

Enhancing the identification of small molecules based on tandem mass spectra and combinatorial fragmentation

Dissertation

zur Erlangung des

Doktorgrades der Naturwissenschaften (Dr. rer. nat.)

der Naturwissenschaftlichen Fakultät III
Agrar- und Ernährungswissenschaften, Geowissenschaften und Informatik
der Martin-Luther-Universität
Halle-Wittenberg,

vorgelegt von

Christoph Karl Heinz Ruttkies

Geb. am 27.12.1985 in Merseburg

Gutachter:

1. Dr. Steffen Neumann
2. Prof. Dr. Dirk Walther

Tag der Verteidigung: 24. November 2022

Acknowledgements

Throughout the writing of this thesis I have received a great deal of support and assistance of so many scientists, colleagues and friends making it hardly possible for me to even try thanking them all on one single page. Nevertheless, I will make an attempt.

I would first like to say a special thank you to my supervisor, Steffen Neumann, for the continuous support and guidance during my thesis, for his expertise, patience and encouragement also through the times when I almost lost sight of the goal. Your trust in me opened up the opportunities to work in several exciting projects in which I could develop and refine skills required for my scientific career and beyond.

Among all the people I met during my work I would like to thank one particular person, Emma Schymanski. You are an inspiring scientist that brought so many invaluable input and ideas into my work. Your scientific eagerness was encouraging for me. I really enjoyed the time in Zurich when I was working with you in person and could meet your family.

Many thanks to Stefan Posch for his support, patience and insights especially in the second half of my work. I enjoyed our discussions with your expertise being extremely beneficial for our project. Bioinformatics is a very diverse research field and I was always impressed and thankful for your willingness and the amount of time you spent to support me in the topics I brought to you.

I would like to thank the IPB including the former head of the SEB department, Dierk Scheel, for giving me the opportunity to be part of this excellent community. I really enjoyed the time as Ph.D. student and will always remember it as a wonderful period of my life.

Many thanks to all members and alumni of my former research group Mass Spectrometry and Bioinformatics at the IPB for their support and discussions. I really enjoyed the pleasant atmosphere in our office and the time we shared in after work activities.

I would like to thank the Eawag and the head of the Department Environmental Chemistry, Juliane Hollender, for the collaboration and the opportunity to work at your facility. I felt very comfortable and enjoyed the time at Dübendorf with the Ph.D. students and staff members I met.

Many thanks to Nadine Strehmel, Martin Krauss and Michael Witting for their collaboration and its outcome being an important part of my thesis.

Thanks to Susann Lindemeyer and Martin Scharm for pushing me in the final phase of my work and the encouragement to finish this thesis.

In addition, I would like to thank my family including my children for being happy distractions to rest my mind outside of my research and giving me the strength to finish this thesis. Now I am sitting here and can finally say, I made it!

Summary

Small molecules play a critical role within living beings and their environment. On the one hand they are important for biological processes and on the other hand they can cause damage to living organisms. Only if we know which molecules are present in biological samples we are able to put them into a functional context. The elucidation of their structures from these samples is necessary for their identification. Besides nuclear magnetic resonance spectroscopy, tandem mass spectrometry (MS/MS) is the analytical tool of choice when small molecules need to be identified. The manual interpretation of resulting datasets requires expert knowledge but is impossible for many MS/MS spectra. This led to the development of computational tools performing an automated interpretation. Examples are rule-based methods, combinatorial fragmentation and statistical approaches that emerged for a large-scale assessment of candidate structures retrieved from small molecule databases. Few existing computational approaches integrate additional information from empirical, statistical and experimental methods. Besides improving the candidate evaluation, the integration of information from different sources also adds more confidence in the identification process. An approach that combines strategies for the identification of small molecules would represent a vital contribution to the research community.

The main goal of my work is to improve the existing combinatorial fragmentation pipeline, MetFrag, and to develop a solution that enables the integration of data from different sources and acquired by different analytical methods. I have extended the scope of application by a novel approach that exploits the idea of the existing MetFrag methodology and provide solutions beyond combinatorial fragmentation. To integrate data from additional analytical methods and to combine different data sources, it is necessary to reconsider and reengineer the existing MetFrag approach. On the one hand I want to improve the performance of identification compared to the existing pipeline and on the other hand to add more confidence to suggested and scored molecular candidates. I will also show how statistical methods combined with combinatorial fragmentation can be used to achieve an improvement in performance and confidence. The enhanced approach will also be compared with other available computational methods to demonstrate the potential of combinatorial fragmentation when combined with additional data sources and statistical models developed in this cumulative thesis.

My solutions and enhancements invented in this work proved to be of importance in the research community and showed to even outperform existing state-of-the-art automated identification methods. The ideas that I combined in a flexible way are a major improvement in the process of the mass spectrometry based identification of small molecules.

Zusammenfassung

Niedermolekulare Verbindungen spielen eine entscheidende Rolle innerhalb und in der Umwelt von Lebewesen. Zum einen regulieren sie biologische Prozesse und zum anderen können sie lebenden Organismen schaden. Nur mit dem Wissen welche Moleküle in biologischen Proben auftreten, sind wir in der Lage, diese in einen funktionalen Zusammenhang zu setzen. Die Aufklärung ihrer Strukturen aus diesen Proben ist für deren Identifikation unerlässlich. Neben der Kernspinresonanzspektroskopie, ist die Massen- (MS) bzw. Tandem-Massenspektrometrie (MS/MS) das Werkzeug der Wahl, um die Struktur kleiner Moleküle aufzuklären. Die manuelle Interpretation der resultierenden Datensätze erfordert Expertenwissen und ist für eine Vielzahl an MS/MS-Spektren unmöglich. Dies führte zur Entwicklung computergestützter Methoden, die eine automatisierte Interpretation ermöglichen. Beispiele sind regelbasierte Verfahren, kombinatorische Fragmentierung und statistische Ansätze, die Kandidaten aus Moleküldatenbanken bewerten. Obwohl diese Methoden bereits effektiv arbeiten, integrieren nur wenige der bestehenden Ansätze zusätzliche Informationen, die man aus empirischen, statistischen und experimentellen Verfahren gewinnen kann. Neben der Verbesserung der Kandidatenbewertung, kann diese Integration zur Erhöhung der Zuversicht im Identifizierungsprozess beitragen. Ein Ansatz, der verschiedene Strategien für die Identifizierung niedermolekularen Verbindungen kombiniert, kann einen entscheidenden Beitrag für die Forschungsgemeinschaft leisten.

In meiner Dissertation mache ich mir die kombinatorische Fragmentierung zu nutze, um die computergestützte Identifizierung von Molekülen auf der Grundlage von MS/MS-Daten zu verbessern. Die Bündelung verschiedener Methoden und Informationsquellen ist das Hauptziel meiner Arbeit. Ich werde zeigen, wie man dabei von der Integration struktureller Informationen aus zusätzlichen experimentellen Methoden profitiert. Zudem werde ich durch die Verwendung von Metadaten aus verschiedenen Datenbanken zeigen, wie man am Beispiel der Anzahl an Patenten und Literaturreferenzen die Bewertung der Kandidatenmoleküle optimieren kann. Meine Erweiterungen beinhalten ebenfalls die Zusammenführung kombinatorischer Fragmentierung und statistischer Methoden, deren Leistungsfähigkeit im Vergleich mit bestehenden Ansätzen verdeutlicht wird.

Mit meinen Beiträgen konnte ich das Anwendungsfeld kombinatorischer Fragmentierung bedeutend erweitern. Die in meiner Arbeit entwickelten Lösungen und Erweiterungen erweisen sich noch heute als ein wichtiger Beitrag zur Forschung und zeigten ebenfalls, dass sie bestehende computergestützte Identifizierungsmethoden übertreffen konnten. Die Ideen, die ich in meiner Arbeit auf eine flexible Art und Weise umgesetzt und kombiniert habe, sind ein wichtiger Fortschritt im Prozess der Identifizierung von niedermolekularen Verbindungen mithilfe von MS-Daten.

Contents

1	Introduction	1
1.1	Small molecules with big impact	2
1.1.1	Metabolites and further examples of small molecules	2
1.1.2	Structural diversity and analytical complexity	4
1.1.3	Isotopes and mass definitions	5
1.2	Mass spectrometry for small molecule research	6
1.2.1	Structure and functionality of a mass spectrometer	7
1.2.2	Mass accuracy and resolution in mass spectrometry	8
1.2.3	Mass spectrum explained with a simulated example	9
1.2.4	Tandem mass spectrometry	10
1.2.5	Identification of small molecules based on tandem mass spectrometry	12
1.2.6	Coupling of chromatography and mass spectrometry	13
1.3	Computational mass spectrometry for the identification of small molecules	14
1.3.1	Databases for small molecules	14
1.3.2	Computational annotation and identification of small molecules based on tandem mass spectrometry	17
1.3.3	Combinatorial <i>in silico</i> fragmentation for the identification of small molecules exemplified by MetFrag	20
1.3.4	Open contest for critical assessment of small molecule identification	22
2	Enhancing the MetFrag pipeline for small molecule annotation	25
2.1	Candidate selection	26
2.2	Fragmentation	27
2.3	Fragment annotation	28
2.4	Candidate scoring	28
2.5	Candidate ranking	29
2.6	Method evaluation	30
3	Discussion	33
3.1	Developed combined scoring principle is major enhancement for MetFrag	33
3.2	Confidence scoring on lipid samples illustrates potential for broader application	34

3.3	Combinatorial structure generation as basis for enlargement of the MetFrag application	35
3.4	Structural elucidation of small molecules remains a topic of interest . . .	36
3.5	Controlled evaluation studies for computational methods are insightful .	37
4	Conclusion	39
5	Peer-reviewed publications	53
5.1	Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an <i>in silico</i> generated compound database and MetFrag	55
5.2	MetFrag relaunched: incorporating strategies beyond <i>in silico</i> fragmentation	67
5.3	Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag	85
5.4	Improving MetFrag with statistical learning of fragment annotations . .	105
5.5	LipidFrag: Improving reliability of <i>in silico</i> fragmentation of lipids and application to the <i>Caenorhabditis elegans</i> lipidome	121
5.6	Tackling CASMI 2012: Solutions from MetFrag and MetFusion	147
5.7	Solving CASMI 2013 with MetFrag, MetFusion and MOLGEN-MS/MS	163
5.8	Critical Assessment of Small Molecule Identification 2016: automated methods	175

1 Introduction

Understanding nature with all its facets is a challenge humanity has been devoting itself for a long time. Growing crops, breeding animals and fighting diseases are examples of achievements that emerged out of the inner pursuit of understanding how life works and improving living conditions (Hayden et al., 1981; Lev-Yadun et al., 2000). These and other research topics, even in their simplest and earliest forms, arose from the current problems and interests of humans and were dependent on available technologies and measurement techniques. The modern human was able to develop tools to extend its borders of perception preset with its equipped senses. For example, with the invention of the light microscope a groundbreaking instrumentation was conceived. Enabled by the enhancement of this technology over the years, Robert Koch was able to detect living microorganisms in 1876 and thus the cause of many diseases (Masters, 2008). However, the invention of the microscope marked just the beginning of the research of the “tiny things” and a more detailed understanding of life. Undetectable for microscopes, tinier entities humans are exposed to are a thousand times smaller than the bacillus *Mycobacterium tuberculosis* discovered by Koch (Brock, 1999). Molecular compounds, including small molecules, influence all biological organisms (Keunen et al., 2016; Massalha et al., 2017). These small molecules might be of biological or industrial origin, such as secondary metabolites and chemicals, respectively. The investigation of the effect they have on biological organisms supports the development of effective natural products and pharmaceuticals (Tian et al., 2019; Schmieder et al., 2014). There are a variety of measurement techniques developed to identify and monitor small molecules in complex samples. The growth of data sets and the speed of acquisition, requires automated methods which can be used for their annotation.

In the next sections, I introduce the basic concepts to describe small molecules and the role they play for biological organisms. Furthermore, I describe the principles of different measurement techniques, namely mass spectrometry, tandem mass spectrometry and chromatography. Following this, an introduction of related computational methods used for computational structure elucidation of small molecules and the challenges in this field is given. The contributions I developed are explained in the in Section 2, “Enhancing the MetFrag pipeline for small molecule annotation”, followed by the Discussion (Section 3) and Conclusion (Section 4) sections. In Section 5 the published manuscripts as part of this cumulative thesis are attached.

1.1 Small molecules with big impact

Small molecules are organic molecules of usually below 900 Da. Despite their relatively small size with up to several dozen atoms, small molecules show a high structural variability caused by the high number of atom types and combinations thereof. Small molecules are regarded as key players that can regulate the activity of different macromolecules in biological systems or have an effect on DNA (genomics), RNA (transcriptomics) and proteins (proteomics) (Schreiber, 2005).

1.1.1 Metabolites and further examples of small molecules

Small molecules that are directly involved in metabolism and transformed within various processes in an organism are called *metabolites*. They are usually formed by enzymatic reactions and located within cells, biofluids and tissues of biological organisms. The set of metabolites that are synthesized by such a biological system is regarded as its *metabolome*. The scientific field investigating these molecules is known as *metabolomics* (German et al., 2005).

While research in the fields of proteomics and also genomics is an ongoing process since several decades, research on small molecules, especially on those connected to the metabolism of biological organisms, is still relatively young (Dettmer et al., 2007). Metabolites are considered as “the end products of cellular regulatory processes, and their levels can be regarded as the ultimate response of biological systems to genetic or environmental changes” (Fiehn, 2002) and can directly be linked to the phenotype of an organism (Nobeli and Thornton, 2006). Thus, there is a growing interest in the investigation of metabolites in scientific research. The number of publications in PubMed Central increased by more than a tenfold in the last decade (2010: 1 105, 2020: 13 212)¹.

Figure 1.1 shows examples of small molecules including several metabolites. Hormones, as one representative group, act in very low concentrations such as the plant-related indole-3-acetic acid (Figure 1.1(a)), also known as auxin. It has a direct influence on plant growth and development (Benjamins and Scheres, 2008). The gonadal steroids, such as testosterone (Figure 1.1(b)), have a pivotal role in the sex development of vertebrates and are directly involved in the reproduction phase of animals. Caffeine (Figure 1.1(c)), predominantly consumed via coffee and related drinks, is famous for the inhibiting effect on receptors of adenosin, reducing its sleep-inducing function. Produced by plants, it may act as a natural occurring pesticide to defend against predators (Nathanson, 1984). Moreover, caffeine is a typical example of a metabolite that is not directly involved in the energy metabolism or growth of plants, but provides significant

¹Publication numbers retrieved on <https://www.ncbi.nlm.nih.gov> with search term “metabolomics”

advantages as they serve survival and defense functions, thus they are called *secondary metabolites*. These metabolites can for example help the plant to protect against competitors and pathogens like bacteria, fungi or insects. There is a high interest in scientific secondary metabolites as they can act as active substances in human medicine, such as antibiotics (Demain and Fang, 2000).

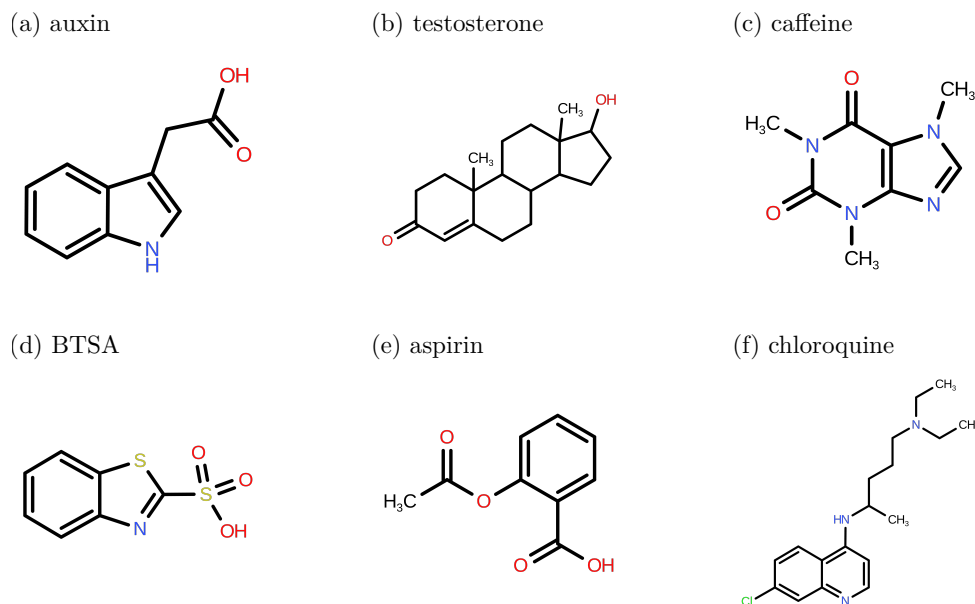


Figure 1.1: Examples of small molecules. The six examples of small molecules shown here are (a) indole-3-acetic acid (auxin), (b) testosterone, (c) caffeine, (d) benzothiazole-2-sulfonic acid (BTSA), (e) acetylsalicylic acid (aspirin) and (f) chloroquine.

Further groups of small molecules are chemicals including pharmaceuticals, synthetic drugs, food additives or contaminants formed by e.g. industrial processes. Many of them are also metabolized by organisms and can be found directly or as biotransformation products in the environment. Even in low concentrations they can have wanted or unwanted effects on metabolism. The chemical benzothiazole-2-sulfonic acid (BTSA) (Figure 1.1(d)) is a derivative of benzotriazole. It is widely used, e.g. as vulcanisation accelerator in rubber and tyre production, thus reaching the environment and being regularly detected as pollutant in industrial wastewater. The European Chemicals Agency² has classified BTSA as a toxic chemical that is harmful to biological organisms (Reemtsma, 2000; Kloepfer et al., 2004). Acetylsalicylic acid (Figure 1.1(e)), also known as aspirin, is a famous pain killing drug. It is indirectly reducing the formation of prostaglandins known to be involved in the modulation of inflammation and pain. Chloroquine (Figure 1.1(f)) was initially developed for the treatment of malaria. Its antiviral properties also resulted in a successful treatment of patients infected by the human immunodeficiency virus (HIV) (Plantone and Koudriavtseva, 2018) and lead to

²<https://echa.europa.eu>

further investigations for an effect on Covid-19 disease during the pandemic. So far no hard evidence on the effective role of chloroquine in the treatment for COVID-19 exists (Gasmi et al., 2021). These six small molecules already show highly diverse activities and indicate their huge influence on biological organisms.

Metabolites can be grouped by their structural characteristics. Carbohydrates, peptides or lipids are typical examples of structural metabolite categories. Representatives of the latter group are involved in important cellular functions like signalling, storage of energy and the formation of building blocks of membranes. They can be further classified into fatty acyls, glycerolipids, glycerophospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids and polyketides, each having several subclasses (Fahy et al., 2005). Although showing a high diversity, many lipid classes are characterized by consisting of fatty acids, esterified to alcohol groups, e.g. glycerides, and to amino groups, e.g. sphingolipids (Sargent et al., 2003).

1.1.2 Structural diversity and analytical complexity

Small molecules, including metabolites, show a broad variability in their chemical structure and properties. Whereas DNA, RNA and proteins are macromolecules composed of well-defined building blocks with four different nucleotides (genome and transcriptome) or the 20 proteinogenic amino acids (proteome), the metabolome shows a variety of extremely diverse chemical compound classes that cannot simply be elucidated by a single experimental method as it is the case for sequencing (DNA). Their diversity ranges from ionic inorganic species to hydrophilic carbohydrates, volatile alcohols and ketones, amino and non-amino organic acids, hydrophobic lipids, to complex natural products (Villas-Bôas et al., 2005). The variability in structure also causes a variability of their physicochemical and geometrical properties, namely polarity, hydrophobicity, weight, confirmation and size. Thus, it is impossible to simultaneously determine the complete metabolome with a single analytical method. To obtain as much information as possible for the analysis of small molecules, combinations of different analytical techniques have been used.

The structural diversity of small molecules is extremely huge. Already the plant metabolome, as a small subset of the entire small molecule universe, is expected to consist of already 200,000 different metabolite structures of which only a part is already known. Enumerations of organic molecules with upto 166 billion generated compounds with less than 17 atoms (Ruddigkeit et al., 2012) still seem to underestimate the expected number of small molecules of the whole chemical space. Thus, scientists enter highly complex spheres when working in this field of research (Fiehn, 2002; Reymond and Awale, 2012). The analysis, discovery and the monitoring of small molecules is important in science, but also for industrial processes and to preserve public health. The latter relates, e.g., to

the monitoring of food or drinking water quality. It is common for samples of complex mixtures to contain thousands of different organic molecules (Vermeulen et al., 2020).

1.1.3 Isotopes and mass definitions

In general, a molecule is built of atoms of one or different elements that are connected via bonds. The connectivity and arrangement of all atoms of a molecule is regarded as its *molecular structure* which uniquely represents the single molecule. A molecule or atom that carries a non-zero net electric charge is called *ion*, specifically an *anion* carries a positive and a *cation* a negativ charge.

The complete elemental composition of a molecule including the numbers of each element is given by its *molecular formula* (e.g. $\text{C}_2\text{H}_5\text{Cl}$). Molecules with the same molecular formula but different (molecular) structures are called *isomers*.

Each element is uniquely identified by its *atomic number*, which represents its number of protons and position in the periodic table. The element carbon (C) for example has an atomic number of six, thus six protons. Two atoms of the same element (same atomic number) and different number of neutrons are called *isotopes*. For some elements only one isotope exists in nature (e.g. Natrium), whereas most elements occur as isotope mixtures with a defined natural distribution. Stable and naturally occuring isotopes of carbon are carbon-12 (^{12}C) (six neutrons) and carbon-13 (^{13}C) (seven neutrons). Their natural distribution is approximately 98.93 % to 1.07 %, making ^{12}C the most abundant isotope of carbon. Due to their almost identical chemical properties, isotopes are usually indistinguishable and not separated in nature. However, due to the different numbers of neutrons, isotopes have different atomic masses. The *atomic mass* is the mass of an atom, which is often expressed in the *atomic mass unit* (u) or in *dalton* (Da), where 1 u is defined as 1/12 of the mass of a ^{12}C atom (Mortimer et al., 2015):

$$1 \text{ u} = 1 \text{ Da} = \frac{m(^{12}\text{C})}{12} = 1.660540 \cdot 10^{-24} \text{ g}$$

The *molecular mass* of a molecule is calculated from the atomic masses of all atoms in that molecule. The IUPAC³ defines further mass values related to the isotopic composition of a molecule. The *exact mass* of an ion or molecule is its calculated mass with specified isotopic composition. The *monoisotopic mass* is the exact mass of an ion or molecule using the mass of the most abundant isotope of each element. The *average mass* is the mass of an ion or molecule weighted for its isotopic composition (Murray et al., 2013).

Table 1.1 shows the calculation of the monoisotopic and average mass exemplified by a molecule with a molecular formula of $\text{C}_2\text{H}_5\text{Cl}$. Chlorine (Cl) occurs as chlorine-

³International Union of Pure and Applied Chemistry

Element	Average mass	Isotopic composition	Atomic mass
H	1.008 Da	$^1\text{H} \rightarrow 0.9999$	1.008 Da
		$^2\text{H} \rightarrow 0.0001$	2.014 Da
C	12.011 Da	$^{12}\text{C} \rightarrow 0.9893$	12.000 Da
		$^{13}\text{C} \rightarrow 0.0107$	13.003 Da
Cl	35.453 Da	$^{35}\text{Cl} \rightarrow 0.7577$	34.969 Da
		$^{37}\text{Cl} \rightarrow 0.2424$	36.966 Da

Average mass			Monoisotopic mass		
	2 x	12.011 Da		2 x	12.000 Da
+	5 x	1.008 Da	+	5 x	1.008 Da
+	1 x	35.453 Da	+	1 x	34.969 Da
=		64.515 Da	=		64.009 Da

Table 1.1: Calculation of average and monoisotopic mass at the example of $\text{C}_2\text{H}_5\text{Cl}$. The first part (top) of the table shows the three present elements, their average masses, their isotopes with the natural abundances and their atomic masses. The second part (bottom) shows the calculation of the average and monoisotopic mass of the molecular formula. Mass values are rounded to three decimal places.

^{35}Cl with 75.77 % and chlorine-37 (^{37}Cl) with 24.24 % in nature. The atomic masses of these isotopes are approximately 34.969 Da and 36.966 Da (average mass: $34.969 \text{ Da} \cdot 0.7577 + 36.966 \text{ Da} \cdot 0.2424 \approx 35.453 \text{ Da}$). The average masses of carbon and hydrogen are approximately 12.011 Da and 1.008 Da which results in an average mass of 64.515 Da for $\text{C}_2\text{H}_5\text{Cl}$. The monoisotopic mass amounts to 64.009 Da using the atomic masses of the most abundant isotopes ($m(^1\text{H}) = 1.008 \text{ Da}$, $m(^{12}\text{C}) = 12.000 \text{ Da}$, $m(^{35}\text{Cl}) = 34.969 \text{ Da}$). The used atomic masses and isotopic compositions can be found in Berglund and Wieser (2011) and Wieser and Berglund (2009).

1.2 Mass spectrometry for small molecule research

Technology advancements enabled and accelerated small molecule research, e.g. in mass spectrometry (MS). In 1922, Francis Aston won the Noble Prize, as he was able to measure masses of charged atoms and to detect their isotopes by MS. Aston was a protégé of the well-known physicist J.J Thomson who discovered the electron in 1897. By the 1940s, when MS instruments were commercially available, industry could use them to control production processes by measuring concentrations of known substances in mixtures. From the 1950s, the contributions of F. McLafferty, K. Biemann and C. Djerassi showed that MS could also be used to elucidate the structures of unknown molecules (Griffiths, 2008). The investigation of commonly occurring fragmentation and rearrangement processes of molecules within the mass spectrometer played an important

role. Further noble prizes were awarded (Tabet and Rebuffat, 2003) for the development of soft ionization techniques in the 1980s which enabled the reduction of unwanted fragmentation. Thus, researchers could also investigate larger macromolecules that remained mostly intact within the MS instrument (Fenn et al., 1989; Karas et al., 1985). This was the beginning of a new era marked with the characterization of proteins using MS. The ability to investigate single molecules in complex mixtures with high accuracy, even in low concentrations has made it the tool of choice in proteomics until today (Aebersold and Mann, 2003).

1.2.1 Structure and functionality of a mass spectrometer

Today, MS is an established technique used for the quantification, structural elucidation and identification of molecules. The general concept is the generation of gas-phase ions (ionized molecules) that are separated by their mass to charge ratio (m/z) (Murray et al., 2013). Figure 1.2 shows the general principle of an MS instrument. It usually performs three steps: ionization, separation and detection. The ionization takes place in the ion source of the instrument for which different ionization methods are available. In metabolomics, electron impact (EI), electrospray ionization (ESI), atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionisation (APPI) are widely used (Williams et al., 2006). ESI and APCI can be considered as soft ionization techniques, as they reduce unintentional fragmentation of ions, which supports the identification process through knowledge of the exact mass of the intact precursor ion. Depending on their properties, the ionization quality differs under specific conditions. Each ionization technique has its preferred scope of application defined by different mass ranges, volatility and polarity of a molecule. Many ionization techniques can be operated in positive or negative mode, resulting in either positively or negatively charged ions. Besides their protonated ($[M+H]^+$) and deprotonated ($[M-H]^-$) form, resulting ions can arise in different cationized or anionized adducts. Typical examples in positive ion mode are $[M+Na]^+$ and $[M+NH_4]^+$ and in negative ion mode $[M+Cl]^-$, where M is the neutral precursor molecule which has been ionized.

From the ion source the ionized molecules are guided by magnetic or electrical fields through the mass spectrometer. Their movement is affected by their m/z ratio, which underlies the main principle of MS-based separation (Ho et al., 2003). Different methodologies exist for the separation performed by the mass analyzer. With the application of an electrical field the quadrupole mass analyzer is able to filter ions with a specific m/z . The quadrupole can scan larger m/z ranges within milliseconds. The time-of-flight (TOF) mass analyzer uses a different principle: it measures the time ions need to traverse through a field-free flight tube until they hit the detector. One of the recent developments are the fourier transform ion cyclotron resonance (FT-ICR) and

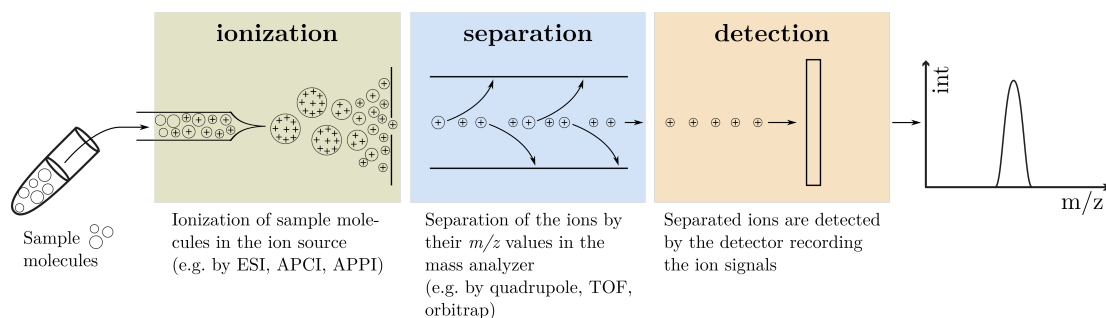


Figure 1.2: Illustration of the MS principle. Molecules of the sample are ionized in the ion source during the ionization step before they enter the mass spectrometer. The charged molecules are guided through the mass analyzer which separates these ions by their m/z values in the separation step. The separated ions with a specific m/z value are detected by the detector. The recording of these ion signals in the detection step results in an m/z peak with a defined intensity.

orbitrap mass analyzers. They have ion storage capabilities where ions are trapped and rotate in a cell using a magnetic field. The acquired frequencies of rotating ions are used to determine their m/z values (Shimadzu, 2019). For small molecule identification, especially in metabolomics, these mass analyzers are frequently used (Sussulini, 2017). Finally, separated ions need to be recorded by detectors that report a signal intensity corresponding to the amount of ionized molecules detected for a specific m/z value.

1.2.2 Mass accuracy and resolution in mass spectrometry

Each MS instrument measures m/z values with a certain error, so the experimentally acquired mass differs from the theoretical mass of the underlying ion (Balogh, 2004). This mass deviation depends on the used instrumentation type, where usually the mass analyzer makes the difference. The mass error of an acquired peak can be expressed as the absolute (in Da) and the relative mass deviation (usually given in ppm, parts per million). In general, the relative mass deviation is used to specify the accuracy of an MS instrument. Instruments with a TOF mass analyzer have a mass deviation of usually 10-20 ppm. Instruments with ultra high mass accuracy such as the FT-ICR mass spectrometer can achieve mass errors of 1 ppm and below. In that case an ion with a theoretical mass of 400 Da would be acquired with a mass deviation of less than 0.0004 Da.

The resolution is another important parameter for the evaluation of a mass spectrometer. As already illustrated in Figure 1.2, the separation of an ion by the mass analyzer and the recording of its signal by the detector results in a peak with a distribution around the real m/z value of the ion. The width of that distribution is determined by the resolving power of the instrument: the smaller the width the higher the resolution

of the mass spectrometer. The resolution is determined using the full-width half-height maximum (FWHM) method where the width (in Da) of a peak's distribution is determined at its half height (i.e. 50 % of its intensity). The m/z value of the apex of the peak's distribution is then divided by the determined width which results in the peak's resolution. A mass spectrometer with a peak measured at m/z 400 with a width (at 50 % intensity) of 0.1 Da has a resolution of 4 000. TOF instruments show mass resolutions of around 10 000 to 30 000 while high resolution instruments, such as the FT-ICR mass spectrometer, have a resolution of more than 1 million.

The identification of small molecules is highly dependent on the mass accuracy and resolution of the acquired mass spectral data. Low accuracy and poor resolution reduce the chances to identify sample molecules, as assignment of their masses becomes ambiguous (Kind and Fiehn, 2007).

1.2.3 Mass spectrum explained with a simulated example

An acquired mass spectrum consists of ion signals recorded with their m/z values and intensities. Intensities are often expressed as normalized values (*relative intensities*) as ratios of the resolved peak and the resolved peak with the highest recorded signal in the spectrum (*base peak*) (Ho et al., 2003). Figure 1.3 shows a simulated spectrum in positive mode of the molecule zeatin. Zeatin belongs to the class of the phytohormons cytokinins, which are known to regulate plant growth and development and play an important role in plant immunity (Schäfer et al., 2015; Großkinsky et al., 2013).

The first peak at around m/z 220.1193 represents the $[M+H]^+$ adduct ion. An additional signal is present at around m/z 221.1218. Both signals correspond to the same ionized molecule of zeatin and are considered as isotopologue ions. They differ only in the isotopic composition of one or more of the constituent atoms (Murray et al., 2013). In the shown example the first peak (at m/z 220.1193) is also considered as the monoisotopic peak: its underlying ion contains only elements with their most abundant isotopes. Typical isotopes with an additional neutron observed in a mass spectrum in metabolomics are carbon-13, nitrogen-15 and hydrogen-2 (deuterium). Carbon-12 occurs in nature with an abundance of around 98.93 %, while its heavier form (^{13}C) occurs by only 1.07 %. As exemplified by the relative intensities of both isotopologue ions in Figure 1.3 this distribution is also reflected in the mass spectrum. The second peak (at m/z 221.1218) has an intensity of about 12 % and represents isotopologue ions of the protonated zeatin containing one isotope with one additional neutron, e.g. either ^{13}C or ^{15}N . For the separation of these isotopologue ions high resolution mass analyzers, such as the FT-ICR, are needed. For lower resolution instruments these isotopologue ions are accumulated in one single peak as it is the case for this simulated spectrum. The contribution to the intensity of these isotopes depends on the abundance-weighted sum

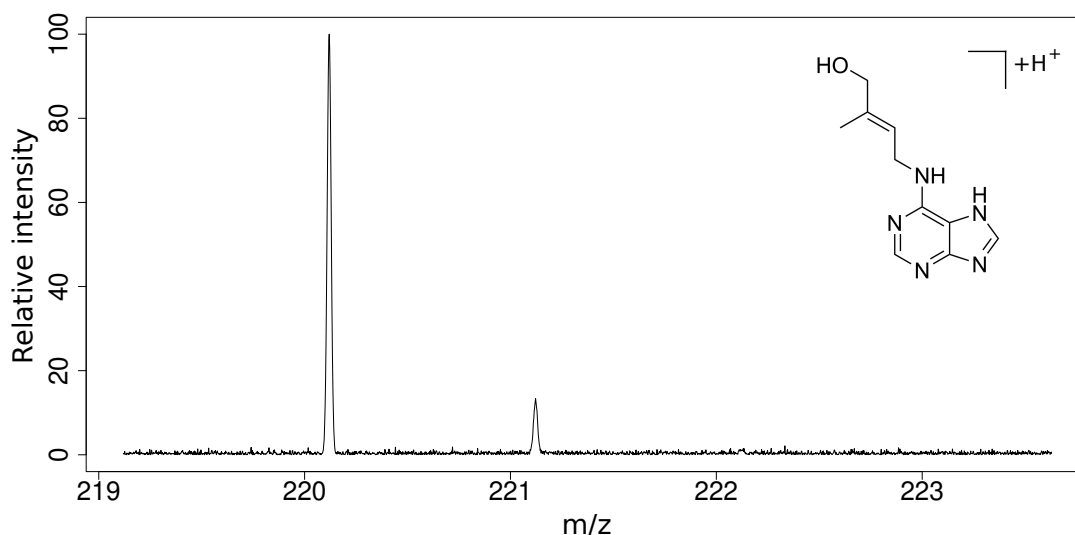


Figure 1.3: Simulated mass spectrum of zeatin. Spectral data was generated using the enviPat Web 2.4 (Loos et al., 2015) and the molecular formula of zeatin ($C_{10}H_{13}N_5O$). (Settings: Adducts - 'positive, M+H'; Resolution - '10 000'; Output - 'Profile'). Noise was simulated by adding a random absolute number drawn from a normal distribution ($\mu = 0$, $\sigma = 0.5$) to each intensity value.

of each element present (Carr and Burlingame, 1996). In a metabolomics experiment, the ^{13}C isotope typically contributes most to the abundance of the second peak in an isotope cluster due to its relatively high probability. Algorithms have been developed to exploit isotope ratios of an MS measurement to determine the molecular formula of the underlying ion (Kind and Fiehn, 2007). The molecular formula can be a first filter criterium to restrict the number of possible molecular candidate structures.

In MS, isotopic labelling experiments are an adequate method to shift the natural abundances of specific elements. Results from such experiments provide additional information that can be used to analyze e.g. metabolic fluxes in biological organisms (Weindl et al., 2015) or assist the identification process of molecular structures contained in the sample (Neumann et al., 2014).

1.2.4 Tandem mass spectrometry

While unintended fragmentation was reduced by softer ionization techniques, intended and controlled fragmentation is helpful to obtain additional information about a single precursor ion of interest. Thus, tandem mass spectrometry (MS/MS) was introduced as a further development of MS to induce an intended fragmentation to a previously selected ion. The resulting fragments support the elucidation of their precursor ion structure. Specialized mass spectrometers are able to perform this intended fragmentation of the precursor ion. Typically, these instruments use two or more connected mass

analyzers (de Hoffmann, 1996). This principle is illustrated in Figure 1.4 which is also called tandem in space.

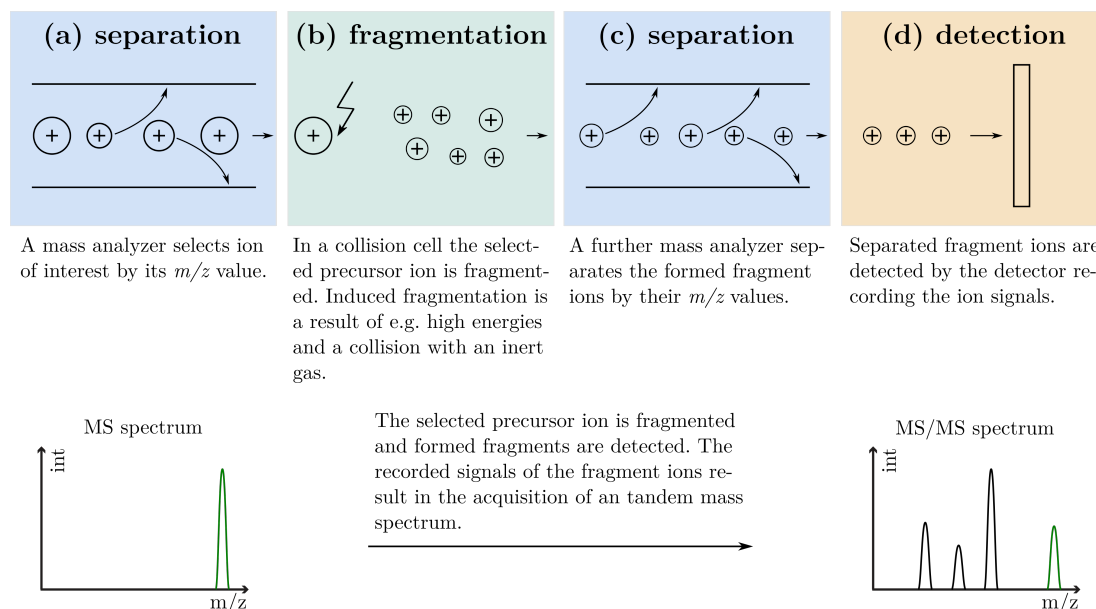


Figure 1.4: Illustration of MS/MS principles. This illustration complements Figure 1.2 to explain the principle of MS/MS with the example of multiple connected mass analyzers (tandem in space). (a) The precursor ion is separated and selected for fragmentation by the first mass analyzer. (b) By applying e.g. higher energies and collision with an inert gas (collision-induced dissociation, CID) in the collision cell the the selected ion is fragmented. (c) The formed fragments are separated by a further mass analyzer and (d) ion signals are recorded by the detector. The result is a tandem mass spectrum (MS/MS spectrum) consisting of ion signals retrieved due to the formed fragments.

The first mass analyzer selects an ion of a molecular precursor based on its m/z value. The ion is guided into a collision cell where its fragmentation is induced. The collision with an inert gas is one example to induce the fragmentation, also known as collision-induced dissociation (CID). The resulting fragments are separated by an additional mass analyzer and recorded by the detector. Due to its successive ion separation steps connected with an induced fragmentation of a selected precursor ion this principle is called MS/MS. Examples of mass spectrometers that are able to perform MS/MS are instrumentations with triple quadrupole (QqQ) (Yost and Enke, 1978) or quadrupole-time-of-flight (QqTOF) (Chernushevich et al., 2001) mass analyzers. Besides connecting multiple mass analyzers where separation and fragmentation takes place in different spaces, specific mass spectrometers have mass analyzers, such as the orbitrap, that perform these steps sequentially in a single cell (tandem in time).

1.2.5 Identification of small molecules based on tandem mass spectrometry

Characteristic m/z fragment peaks of an MS/MS spectrum support the identification of the underlying precursor molecule. Manual annotation of fragment structures requires expert knowledge. Analysts annotate m/z peaks in the spectrum with matching fragment structures by comparing the experimental with theoretical masses. The lookup in spectral databases is also an effective method to find putative fragment structures or even precursor molecules with the match of the entire spectrum. Annotated fragment structures and the information retrieved by the MS measurement (mass and/or molecular formula) support the elucidation of the underlying molecular structure.

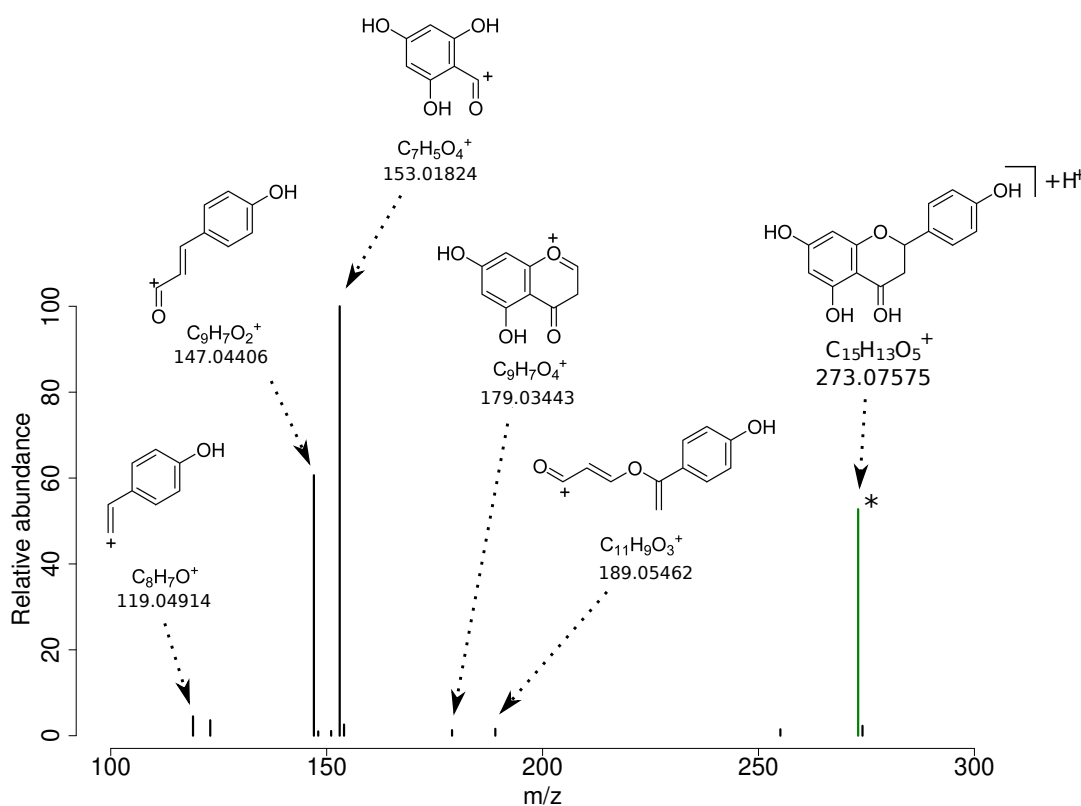


Figure 1.5: Example of a MS/MS spectrum retrieved from MassBank spectral database (ACCESSION ID: PB000123) manually annotated with putative fragment structures, acquired in positive ion mode by an LC-ESI-QTOF instrument. The m/z peak marked with an asterisk represents the protonated monoisotopic precursor ion of naringenin.

Figure 1.5 shows an example of a manually annotated MS/MS spectrum acquired from naringenin in positive ion mode. The m/z peak marked with an asterisk represents the precursor ion of naringenin with an m/z value of approximately 273.076. From annotated fragment peaks, substructures of the precursor can be deduced. Thus, this strategy supports the determination of the entire molecular structure.

For few spectra this approach is feasible. The manual analysis of larger data sets with thousands of spectra obtained from high-throughput experiments such as LC-MS/MS studies is impossible. Thus, computer-assisted methods need to be applied for an automated processing of the generated spectra.

1.2.6 Coupling of chromatography and mass spectrometry

The term chromatography was introduced by Mikhail Tswett who is considered the inventor of this pioneering technique in 1906, which separates components of a sample within a column (Ettre and Sakodynskii, 1993). Almost 50 years later, Archer Martin and Richard Synge received the Noble Prize for the development of the partition chromatography (Martin, 1953). Their separation is based on the partitioning of the sample molecules by the use of a mobile and a stationary phase. The separation of the molecules is possible due to different grades of solubility in the stationary phase. The higher the solubility and its interaction with this phase, the higher is the molecule's retention time in the column. The mobile phase transports sample molecules through the column and can either be a liquid (LC) or a gas (GC). Different chromatographic techniques were developed and have been improved toward faster and more efficient separation. GC and high-performance liquid chromatography (HPLC) have become the dominating separation techniques today (Lundanes et al., 2013).

A molecule's retention time indicates its degree of interaction with the stationary and the mobile phase. Depending on the separation technique, the retention time of a molecule can be used to make conclusions about different physicochemical or geometrical properties including its hydrophobicity, polarity, shape and size (Issaq, 2001). Due to the relation of retention time and structural properties, chromatography is another important source of information for the identification of an investigated molecule.

Chromatography can be coupled with MS instruments to combine two orthogonal separation methods. The investigation of samples retrieved from biological or industrial processes, medical studies or criminal investigations produce mixtures with unknown components that are too complex to be directly analyzed with MS (Karasek and Clement, 2012; Gohlke and McLafferty, 1993). In the 1950s, GC-MS became the first application of coupling chromatography and MS. This enabled the detection of a huge number of different components in complex samples (Gohlke, 1959). Further enhancements were used in the Viking Space Probe mission, where a GC-MS instrument was sent to Mars to search for organic molecular structures (Biemann, 1979). Due to its chromatographic properties GC-MS is an excellent tool for the analysis of volatile polar and nonpolar sample components, however is not directly applicable for nonvolatile and semipolar components (Lee et al., 2013).

To enhance separation and the capability of compound identification derivatization of the sample to be analyzed is used in GC (and GC-MS). Types of derivatization include alkylation (e.g. methylation), silylation (e.g. formation of trimethylsilyl derivatives) and several others. Besides decreasing the boiling point of molecules, there are several other intended effects of derivatization in GC-MS. Derivatized sample molecules may have significantly different properties from each other and their underivatized precursors, allowing separations that are difficult to achieve otherwise (Moldoveanu and David, 2018).

Besides GC-, LC-MS is another method of choice when complex mixtures need to be investigated. Especially for the analysis of metabolic components innumerable studies prove that the coupling of chromatography and MS is indispensable as described in Fiehn (2002).

The permanent acquisition by MS of continuously eluting components from chromatographic partitioning results in a three-dimensional data set. These data sets may consist of several thousands of mass spectra separated by the additional retention time dimension usually given in seconds or minutes.

1.3 Computational mass spectrometry for the identification of small molecules

A variety of computational (pre-)processing steps for acquired raw data are required before computational methods for the identification of small molecules can be applied. Computational scientists especially from the field of metabolomics made significant contributions by developing methods for processing and pre-processing of mass spectral data. An overview about pre-processing steps for retrieving MS/MS peak lists from raw data can be found in (Katajamaa and Orešič, 2007) and (Sumner et al., 2005).

In the following, I will describe computational identification of small molecules with a focus on databases designed for small molecule research and their necessity in the process of structure elucidation. I will introduce different computational approaches that perform identification of small molecules on the basis of MS/MS data with the focus on combinatorial fragmentation at the example of MetFrag (Wolf et al., 2010) as the major basis of my work.

1.3.1 Databases for small molecules

Small molecule databases are a valuable resource of information used in computational MS. There are many databases available for different purposes, ranging from small molecule and pathway databases (Frolkis et al., 2010) via metabolomics experiment data-

bases (Haug et al., 2013) to mass spectral library databases and databases containing structural information about small molecules linked to different resources including meta data.

Database	Description
KEGG	- small molecule related content focused on metabolites - links metabolites via pathways to genes and proteins
PubChem	- largest open resource for small molecular structures - contains information from many resources (including patents & literature) - includes a large amount of bioactivity data
ChemSpider	- large resource for small molecular structures - links to information from different resources
LipidMaps	- resource for lipid molecular structures - provides a hierarchical lipid classification system
HMDB	- contains information about metabolites found in human - metabolites are linked to human related data sets
CompTox	- dashboard that integrates data on chemicals from different platforms for environmental sciences - includes physicochemical properties and data on exposure, toxicity, bioassays and more
<i>Databases with focus on mass spectral data content</i>	
MassBank	- public sharing of reference mass spectra - MS/MS spectra measured from different setups and conditions - shared over servers around the world - MassBank of North America (MoNA) was introduced in 2015
GNPS	- spectral networks created from uploaded spectra - contains annotated reference MS/MS spectra
GMD	- focused on GC-MS spectra of biologically active plant metabolites
LipidBlast	- computer-generated MS/MS database for lipid annotation

Table 1.2: Overview of selected databases related to small molecule research

There are two types of databases for small molecules that are relevant to my work: (1) databases providing structural information (also known as compound databases) and (2) databases containing reference mass spectra of known molecules (also known as mass spectral libraries). Table 1.2 gives an overview of available compound databases and spectral libraries used in this thesis. A more general overview about compound and mass spectral databases is given in Vinaixa et al. (2016).

Besides structural information, compound databases contain information about chemical and physical properties, biological functions and different identifiers usually

linked to other available resources (Kim et al., 2016; Pence and Williams, 2010). The number of molecular structures contained in these databases varies significantly depending on their scope. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000; Kanehisa et al., 2006) and the Human Metabolome Database (HMDB) (Wishart et al., 2007) with around 18 500 and 114 200 molecular structures are mainly focused on metabolites. LipidMaps (Sud et al., 2007) with around 43 700 structures contains only lipids. PubChem (NCBI, 2020) and ChemSpider (RSC, 2020) are the largest databases with around 103 and 81 million compounds. They both integrate hundreds of different data sources linked to the structures. Both databases serve as an important source of molecular candidates for the annotation and identification in computational MS. However, due to the limited biological relevance of the majority of contained chemicals, both compound databases play a limited role in metabolomics. The CompTox Chemicals Dashboard (Williams et al., 2017) provides data about chemicals for environmental scientists and computational toxicologists. Initiated and maintained by the U.S. Environmental Protection Agency’s (EPA), this public dashboard links various data sources and contains over 800 000 compounds (numbers accessed 03/2020).

Mass spectral libraries contain mass spectrometric reference data sets including MS/MS spectra of known precursor molecules. They are important for the development of data-driven approaches for the identification of small molecules (Scheubert et al., 2013). Examples are the Global Natural Products Social Molecular Networking (GNPS) (Wang et al., 2016b), MassBank (Horai et al., 2008, 2010), MassBank of North America (MoNA) and Golm Metabolome Database (GMD) (Kopka et al., 2005). Their content is usually much smaller compared to available compound databases caused by the cost-intensive and time-consuming collection and acquisition of the experimental spectra, as well as the limited number of reference standards available. With 161,100 experimental spectra (accessed 03/2020) MoNA is one of the larger open mass spectral libraries. Most MS/MS spectral libraries contain multiple spectra of one molecule acquired under different conditions and with different instruments reducing the number of unique molecules covered. It is common that compound databases also integrate mass spectral data sets connected to their compounds, such as LipidMaps, HMDB, PubChem, CompTox and ChemSpider.

To overcome the lack of available reference spectra, there is a rising trend of generating computationally predicted libraries of MS/MS spectra. Different approaches have emerged to perform spectra prediction on the basis of an input molecular structure. There are quantum chemistry-based methods that excel in prediction accuracy, which are however very time-consuming especially for larger molecules (Bauer and Grimme, 2016; Cautereels et al., 2016). LipidBlast uses a heuristic and rule-based approach to create MS/MS spectra of lipid molecular structures (Kind et al., 2013). In addition, statistical methods have been developed that use reference spectra to learn fragmentation processes for the prediction of fragments and intensities. Competitive fragmentation modeling

implemented in the tool CFM-ID (Allen et al., 2014) uses a probabilistic generative model to estimate the likelihood of a specific fragmentation processes. The different approaches have several limitations including their application domain. LipidBlast is only applicable to molecules with well-defined patterns of fragmentation, which is true for many lipid molecular structures. They usually show a similar fragmentation pattern within one lipid class. CFM-ID is more flexible as it learns the patterns directly from the training data. However, the application domain is still limited by the used training data (usually only several thousand compounds). The validation of computational approaches used for spectra prediction in terms of accuracy and precision is extremely important as this goes along with the quality of the generated spectral library (Kind et al., 2018).

1.3.2 Computational annotation and identification of small molecules based on tandem mass spectrometry

Given an MS/MS spectrum of an unknown precursor ion, spectral library search is one method to support the identification. The library is typically screened for spectra that are similar to the acquired MS/MS spectrum. Precursor molecules of matching library spectra are putative candidates for the identification. Different measures are used to compare MS/MS spectra and to find matches in larger spectral libraries. An overview about available strategies is given in Kind et al. (2018). If similar spectra are present in the database there is a high probability for the identification of the unknown molecule. However, the coverage of molecules with MS/MS spectra is rather small (Vinaixa et al., 2016) which makes alternative strategies for structure elucidation even more important.

In the following, computational strategies will be described. Several approaches try to model or reconstruct the fragmentation process of candidate molecules. The generated fragments are used to annotate peaks in the query MS/MS spectrum. Some of these approaches are rule-based, such as MassFrontier (Thermo Scientific, 2020) or MS-Finder (Tsugawa et al., 2016). Others use a greedy combinatorial fragmentation procedure to generate many putative fragments that are evaluated by different scoring mechanisms. MetFrag (Wolf et al., 2010), as one of the first tools, MAGMa (Ridder et al., 2014) and MIDAS (Wang et al., 2014) are examples for combinatorial fragmenters. Further methods use quantum chemistry-based (Mayer and Gömöry, 1993, 1994) or statistical methods for fragment prediction (Allen et al., 2014). Besides fragmentation process modelling, different approaches emerged that predict molecular feature (or substructure) vectors, such as molecular fingerprints, for a given query MS/MS spectrum. The predicted fingerprint is used to query compound databases to retrieve putative candidates (Heinonen et al., 2012; Dührkop et al., 2015; Brouard et al., 2016). Recently, several approaches have been published that also make use of neuronal networks and

deep learning methods for the interpretation of MS/MS spectra (Shrivastava et al., 2021; Stravs et al., 2021). In the following, I will highlight the three strategies: combinatorial fragmentation, spectrum and fingerprint prediction.

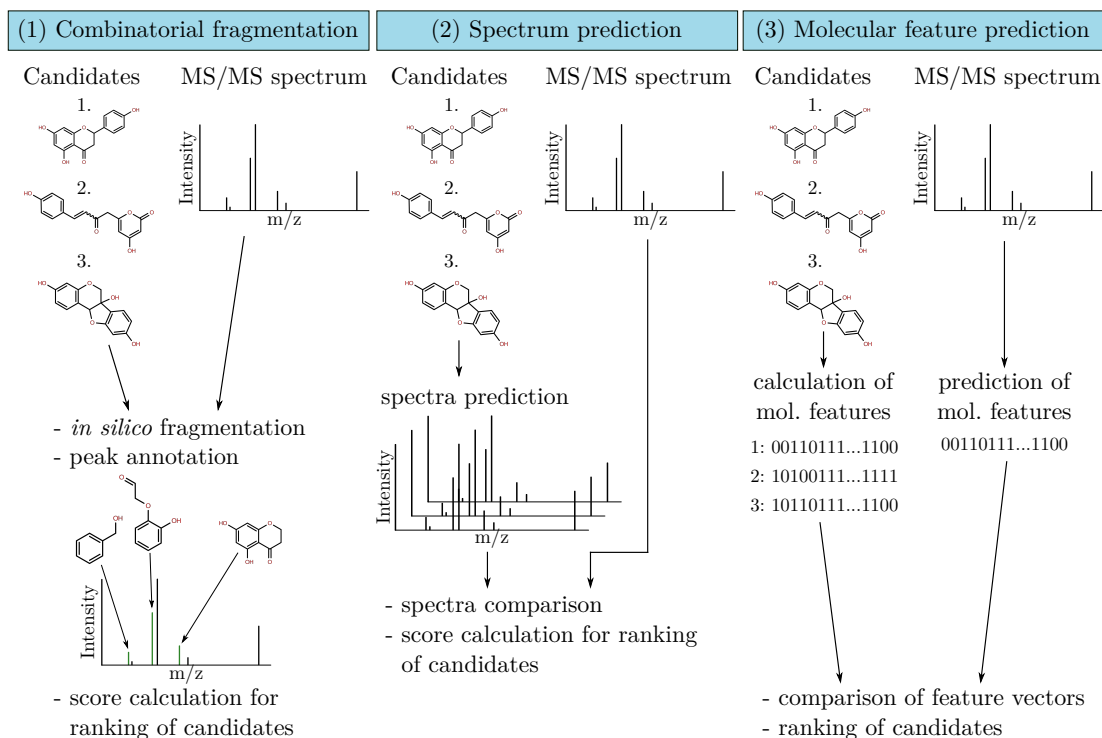


Figure 1.6: Selected methods for computational identification of small molecules based on MS/MS spectra. (1) For a given candidate molecule combinatorial approaches generate putative fragments that are assigned to m/z peaks in a query MS/MS spectrum. Based on the number of annotated peaks candidate scores are calculated and used for candidate ranking. (2) MS/MS spectrum prediction is used to generate spectra for candidate molecules *in silico*. These spectra are compared to the query MS/MS spectrum which is the basis for the ranking of the underlying candidate molecules. (3) Given an MS/MS spectrum, predicted molecular features are matched against a list of putative candidate molecules used for ranking.

In Figure 1.6 the principles of combinatorial fragmentation, spectrum and fingerprint prediction for the identification of small molecules are illustrated. As shown in Figure 1.6 (1), combinatorial fragmentation is directly used to annotate the *in silico* generated fragments of a candidate molecule to the given m/z peaks in the MS/MS spectrum. By evaluating the assigned fragment structures and the annotated peaks, the candidates are scored and sorted to create a ranked candidate list. The bottleneck is the enumeration of all possible fragments which is oftentimes restricted to reduce the number of generated fragments and the runtime of these approaches. Depending on the candidate’s molecular structure, thousands of possible fragments can be generated using combinatorial approaches.

As already mentioned in Section 1.3.1, computational methods exist that predict MS/MS spectra for given candidate molecules. These methods are also used for the identification of small molecules. The principle is illustrated in Figure 1.6 (2). Besides the tool CFM-ID (Allen et al., 2014), quantum chemistry-based methods like quantum molecular dynamics or density functional theory are also in use to predict fragment peaks. These methods are known to be very accurate but usually have high computational costs making them unsuitable for the processing of many molecular candidates (Borges et al., 2021). CFM-ID uses reference MS/MS spectra to train a Markov chain model for the evaluation of transitions for different fragments. This approach is used to reconstruct the whole fragmentation process of a single molecule. Compared to combinatorial fragmenters, the fragment prediction is more specific. Given the statistical prediction model, only fragments with high probabilities are predicted. As also predictions for intensities can be made together with the m/z values, this method is used to simulate MS/MS spectra for molecular candidates. Due to their reasonable computational time, statistical approaches represent a good compromise for fragmentation modelling as for each putative candidate a MS/MS spectrum needs to be predicted. Each predicted spectrum is compared to the experimental query MS/MS spectrum. The candidates are then ranked by the scores calculated on the basis of the spectral similarities.

The third main methodology illustrated in Figure 1.6 (3) makes use of an entirely different approach, as it does not try to reconstruct the fragmentation process of a given molecule. First, it makes use of molecular feature vectors that encode molecular characteristics. Calculated for a molecule, each position represents a defined characteristic that is set if it is found in the molecule. Heinonen *et al.* (Heinonen et al., 2012) introduced a machine learning method based on support-vector machines that predicts molecular fingerprints for MS/MS spectra. The training of the model requires MS/MS reference spectra and molecular fingerprints calculated for the precursor molecules. Given a query MS/MS spectrum, the trained model predicts a fingerprint, which is used to query compound databases for the retrieval of candidates. These candidates are ranked by the similarity of their calculated fingerprints to the one predicted. This approach is further enhanced by using fragmentation trees in CSI:FingerID (Dührkop et al., 2015) or by using an input output kernel regression learning method (Brouard et al., 2016).

The majority of the mentioned methods produce a ranked molecular candidate list for a given query spectrum. The better the method, the higher the chance to identify the correct molecule. Different computational methods are typically evaluated by the comparison of the ranking positions of the correct candidate for a set of reference MS/MS spectra. Dührkop *et al.* (Dührkop et al., 2015) and Schymanski *et al.* (Schymanski et al., 2017b) evaluated different computational tools on a large set of reference MS/MS spectra. They showed that statistical models, in particular fingerprint prediction models, are the most powerful approaches for the identification of small molecules.

1.3.3 Combinatorial *in silico* fragmentation for the identification of small molecules exemplified by MetFrag

The software MetFrag was published in Wolf et al. (2010) and is a typical example that makes use of combinatorial fragmentation. It was one of the first open source approaches used for the annotation of MS/MS spectra to identify small molecules. Figure 1.7 shows the five major steps performed by the MetFrag pipeline: candidate selection, (combinatorial) fragmentation, fragment annotation, candidate scoring and ranking. In step (1), the given information about the precursor molecule is used to retrieve molecular candidates. Compound databases such as KEGG, PubChem or ChemSpider are queried using either the exact mass of the precursor with a mass deviation or, if provided, its molecular formula.

The main goal of the fragmentation step (2) is the enumeration of fragments for each selected candidate. The molecular structure of each candidate is represented as a graph with nodes representing the atoms and edges representing the bonds. The enumeration problem is solved by creating a fragmentation tree. The root of this tree consists of the intact candidate molecule and each of the inner nodes represents a fragment. In the ideal case, the removal of a bond from the molecular graph results in two subgraphs representing two fragments. If the initial bond removed is located in a ring, then at least one additional bond needs to be removed. With an iterative breadth-first algorithm, each bond of the intact molecule is removed to create possible fragments. With this procedure, the first layer of the fragmentation tree is built. The expansion of the fragmentation tree is realized by successively removing bonds from the fragment subgraphs. Due to its combinatorial nature, the fragmentation step is the speed-limiting step of the whole MetFrag pipeline. Thus, the enumeration of fragments is restricted by a specified maximum tree depth of the fragmentation tree, although it can lead to the possible loss of a chemically meaningful fragment structure. In Wolf et al. (2010) the maximum tree depth was set to a value of 2.

During the fragment annotation step (3), the generated fragments are assigned to m/z peaks of the given MS/MS spectrum. By comparing the theoretical mass of the created fragments with the masses of the acquired m/z peaks, fragment peak pairs can be assigned. Due to mass errors in the measurement, a specified mass deviation is considered in this comparison.

After a candidate's fragments have been assigned to the m/z peaks, a score is calculated in step (4). The score calculation involves the intensities and m/z values of the peaks annotated with a generated fragment. Moreover, the weights of bonds being removed to form the assigned fragments are also considered for the score calculation. In Wolf et al. (2010), bond dissociation energies (BDE) are used as bond weights that are defined for each bond type. This energy, usually given in kJ/mol, defines the enthalpy (per mole) required to break a specific bond (Muller, 1994). Equation 1.1 shows the

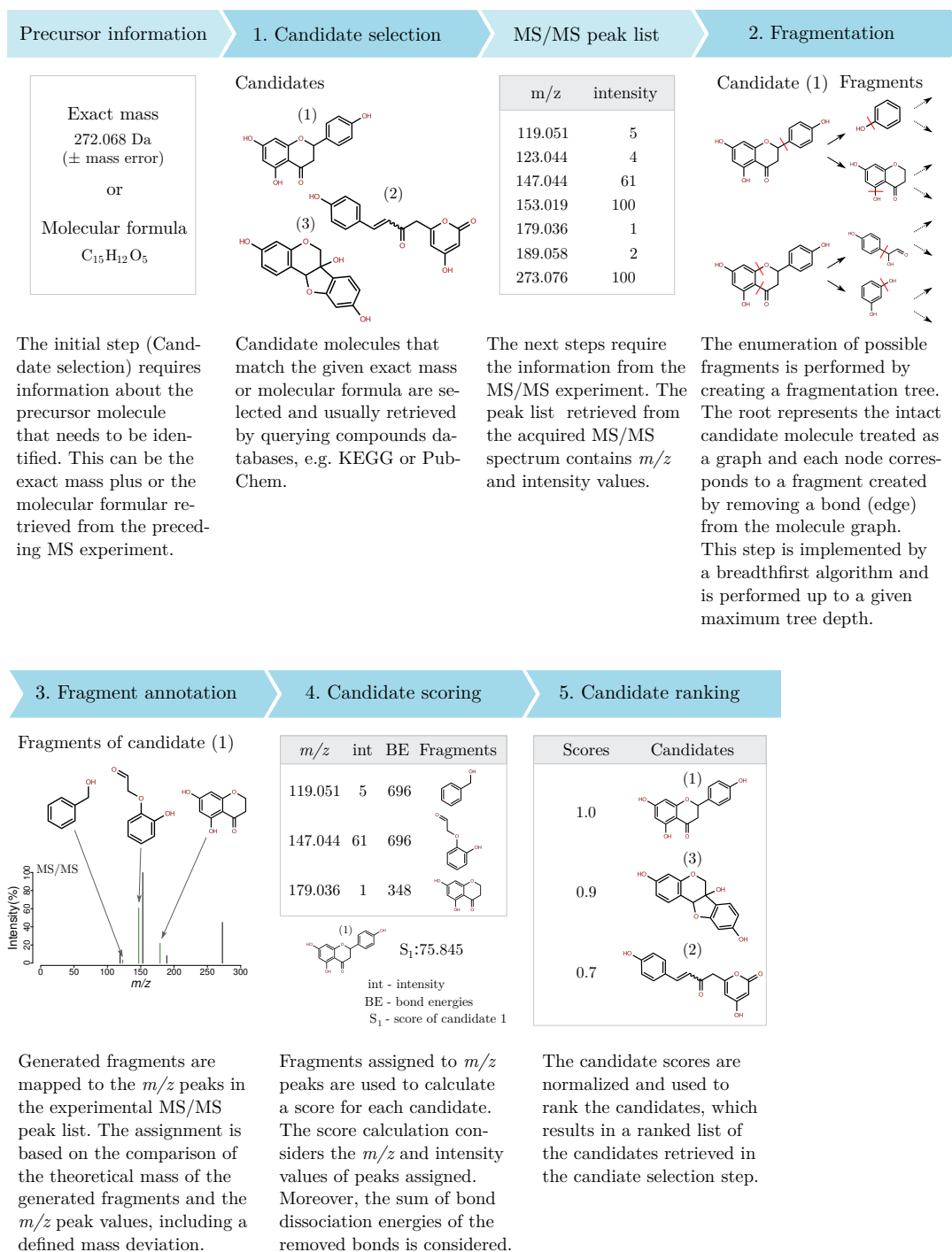


Figure 1.7: Workflow describing the MetFrag pipeline initially published by Wolf et al. (2010). The pipeline consists of five major steps: Candidate selection, Fragmentation, Fragment annotation, Candidate scoring and Candidate ranking. The major inputs are also included: Precursor information and MS/MS peak list.

calculation of a score of candidate i as it was used by Wolf et al. (2010):

$$S_i = \frac{1}{\max_{i=1..n}(w_i)} \cdot w_i - \frac{1}{2 \max_{i=1..n}(e_i)} \cdot e_i \quad (1.1)$$

where

$$w_i = \sum_{f \in F_i} (\text{int}_f)^{0.6} \cdot (\text{mass}_f)^3$$

$$e_i = \frac{1}{|F_i|} \sum_{f \in F_i} \sum_{b \in B_f} \text{BDE}_b$$

The number of retrieved candidates is given by n . The score of candidate i consists of two main terms, w_i and e_i . The term w_i evaluates each peak annotated with a fragment $f \in F_i$, which is the set of fragments of candidate i annotated to a peak in the spectrum. The term w_i , also known as the weighted peak count, sums up the products of intensity (int_f) and m/z (mass_f) of the peaks annotated with a fragment f . The intensity and mass of the annotated peak are weighted by different exponents taken from the literature with $m = 0.6$ for intensity and $n = 3$ for the m/z value. Thus, explained peaks with higher m/z and intensity values have a higher impact on the candidate score S_i as they are presumed to be more characteristic than peaks with lower m/z and intensity values. The second term of the candidate score (e_i) evaluates the annotated fragments as it sums up the BDEs of all bonds $b \in B_f$ that were removed to generate these fragments. Higher BDEs contribute to a decrease of the candidate score S_i . The two main terms are normalized by the factors $1/\max(w)$ and $1/(2 \max(e))$.

The fragmentation (2), fragment annotation (3) and candidate scoring (4) steps are performed separately for each of the selected candidates. The calculated candidate scores are used in the final candidate ranking step (5) to create a ranked candidate list with the goal to rank the correct candidate at the top of this list.

As described in (Wolf et al., 2010), the complete MetFrag pipeline incorporates further steps. These include (i) the integration of neutral loss rules to account for molecular rearrangements during the fragmentation process, (ii) the elimination of redundant fragments during the fragmentation step and (iii) the clustering of the candidate molecules based on their structural similarity.

1.3.4 Open contest for critical assessment of small molecule identification

The open contest ‘‘Critical Assessment of Small Molecule Identification’’ (CASMI) was founded to assess the performance of state-of-the-art approaches for structure elucidation on the basis of mass spectral data. This contest represents an excellent opportunity

for experts in the field of computational mass spectrometry to test and compare their approaches. This comparison reveals the current status of computational tools and their performance in structure elucidation.

The contest initially was held in 2012 (Schymanski and Neumann, 2013) and aimed for the exchange of ideas and to compare different identification approaches of the participants. MS data sets mainly including MS/MS spectra connected with the precursor information were provided. Participants were invited to submit a scored candidate list for each challenge (i.e. MS/MS spectrum). After the deadline, submitted candidate lists were automatically evaluated based on different evaluation criteria (e.g. number of correct highest scored candidates) to rank the submissions of the different participants. The contest was subdivided into different categories, providing data sets for different tasks to be solved by the participants. These tasks ranged from the identification of the correct molecular formula to the correct molecular structure. Over the years 2012 and 2013 it attracted either participants submitting results retrieved by computational methods or using manual identification approaches. For these earlier contests, manual approaches often outperformed the *in silico* methods. The increasing number of challenges over the years 2014, 2016 and 2017 was one reason why the submissions based on automated methods increased from contest to contest. From 16 in 2012 the number of challenges raised to up to 243 challenges in 2017.

The CASMI contest represented a valuable resource to evaluate, compare and enhance state-of-the-art methods used for the identification of small molecules. In addition, it showed how computational methods improved by comparing the submission results over the different years, thus was an excellent way for the generation of knowledge to the community in the field of computational mass spectrometry.

2 Enhancing the MetFrag pipeline for small molecule annotation

Among the different strategies, combinatorial fragmentation has already proven to be an excellent tool for MS/MS-based elucidation of molecular structures (Wolf et al., 2010; Ridder et al., 2014; Wang et al., 2014). However, the potential that lies in this methodology has not been fully exploited so far, although combinatorial fragmentation is well suited to be enhanced and combined with different strategies. Computational approaches are mainly restricted to the use of MS/MS data only. Databases related to small molecules contain valuable information about selected molecular candidates that is hardly used in this automated annotation process. This data could support computational approaches to further corroborate and narrow down the number of candidates to those most relevant for the current investigation. Moreover, structure elucidation that predominantly uses the limited information available from the MS/MS spectrum, might not be sufficient to identify the underlying molecule unequivocally. The lack of evidence for all possible structural features and the sometimes sparse information contained in the spectrum might be reasons (Stein, 2012). Thus, there is a huge potential to include further information such as polarity or the isotopic pattern as obtained by chromatography or isotopic labelling. The main goal of my thesis is to improve the existing MetFrag pipeline and to develop a solution that enables the integration of data from different sources and acquired by different analytical methods. The solutions developed to achieve this goal are described in the following sections and more extensively in the related peer-reviewed publications.

I have extended the scope of application by a novel approach that exploits the idea of the existing MetFrag methodology and provide solutions beyond combinatorial fragmentation. To integrate data from additional analytical methods and to combine different data sources, it is necessary to reconsider and reengineer the existing MetFrag approach. On the one hand I want to improve the performance of identification compared to the existing pipeline and on the other hand to add more confidence to suggested and scored molecular candidates. I will also show how statistical methods combined with combinatorial fragmentation can be used to achieve an improvement in performance and confidence. The enhanced approach will also be compared with other available computational methods to demonstrate the potential of combinatorial fragmentation

when combined with additional data sources and statistical models developed in this cumulative thesis.



Figure 2.1: Peer-reviewed and published manuscripts contributing to my cumulative thesis. They are classified by their contribution to the different topics of the MetFrag pipeline aligned with Figure 1.7. The additional topic “Method evaluation” has been added. For each publication the section number is included where the original manuscript can be found in this thesis.

Figure 2.1 is adapted from Figure 1.7 and illustrates the contribution of each manuscript to the enhanced MetFrag pipeline. The section number attached to each publication listed in Figure 2.1 indicates where the manuscript can be found in my thesis. The classification of the manuscripts are chosen by their main contribution to the MetFrag pipeline entitled with “Candidate selection”, “Fragmentation”, “Candidate scoring”, “Candidate ranking” and “Method evaluation”. Note that a single manuscript may have contributed to several aspects and in this case occurs more than once (Sections 5.2 & 5.3). In the following, the listed manuscripts are put in the context of the objectives.

2.1 Candidate selection

The enhancement of the MetFrag pipeline goes along with the enlargement of the applicability to different available analytical methods. Compound databases provide candidates to be processed in the pipeline, thus are a limiting factor when MetFrag is applied to an analytical method of choice. There are two crucial drawbacks of compound databases. First, they are not complete, hence if the correct candidate is missing there is no chance to annotate the correct compound to the query MS/MS spectrum.

Unless there is a similar compound present in the database, the chance of a successful elucidation of the correct structure is quite low. Second, compound databases are rarely specific for a certain analytical method such as derivatization (e.g. methylation, silylation) or isotopic labelling. This results in the accumulation of many false positive candidates making the candidate list unnecessarily complex.

In Ruttkies et al. (2015) (Section 5.1) I developed a combinatorial workflow to generate *in silico* derivatized compounds by well-defined rules. The heuristic uses existing compound databases, such as KEGG, as input and creates databases with altered structures that were applied to data acquired in GC/APCI-MS/MS experiments. The experimental design required an alteration of the sample to be analyzed through its derivatization prior to data acquisition. At first, the created *in silico* derivatization approach was validated against the GMD as reference database. To demonstrate the workflow, *in silico* generated databases were applied together with MetFrag to MS/MS spectra acquired from GC/APCI-MS/MS profiles of *Arabidopsis thaliana* and *Solanum tuberosum*. The comparison with the GMD revealed a true positive rate of 94 %. *In silico* annotation using MetFrag showed good results with 57 % of the correct candidates ranked at first position when a derivatized KEGG database served as candidate source.

The developed approach was adapted in Ruttkies et al. (2019b) (Section 5.3) to be used for the processing of MS/MS data acquired in hydrogen-deuterium exchange experiments. Here, candidate molecules were deuterated *in silico* by incorporating assumptions about easily and partially exchangeable hydrogens.

2.2 Fragmentation

Due to its combinatorial nature the *in silico* fragmentation is the most time-consuming step in the MetFrag pipeline. Limiting the tree depth of the breadth-first algorithm is one way to reduce the runtime and memory consumption. Further methods implemented in the original MetFrag approach are redundancy checks of generated fragments based on the molecular formula. Fragments in one tree depth cycle are regarded as duplicates and are discarded if they have the same molecular formula even if they have a different connectivity. This weak redundancy check can cause many false fragment removals and a possible loss of reliable fragment annotations.

While relying on the combinatorial breadth-first algorithm in the new MetFrag version developed in Ruttkies et al. (2016) (Section 5.2) as a powerful method for small molecule annotation, algorithmic and data structure improvements have been implemented to reduce runtime and make former redundancy checks unnecessary. Generated fragments are now stored as bit vectors referencing atoms in the original precursor molecule. This allows a fast mass calculation and a reduction of memory consumption. The omission of redundancy checks and allowing for higher tree depths

can lead to more reliable fragment annotations.

The refinements contributed to an improved annotation performance indicated by the median rankings of the correct candidates which decreased from 8 to 4 compared to the original MetFrag approach using ChemSpider database. The 473 MS/MS spectra used in this comparison were processed with an average runtime of 54 s per spectrum. Moreover, the number of correct candidates ranked among the first ten hits increased from 258 to 320 compared to the original MetFrag version as shown in Ruttkies et al. (2016) (Section 5.2).

2.3 Fragment annotation

The annotation of *in silico* generated fragments to m/z peaks in the query MS/MS spectrum is based on the comparison of the theoretical with the experimental mass. Rules have been implemented in the original MetFrag approach to account for rearrangements during neutral loss fragmentation processes as additional mass shifts need to be considered due to the loss of additional hydrogens. To integrate data acquired in isotopic labelling experiments such as hydrogen-deuterium exchange additional rules need to be applied. These need to account for additional mass shifts in the MS/MS spectrum due to the presence of heavier isotopes.

In Ruttkies et al. (2019b) (Section 5.3) the fragment annotation approach is enhanced by the incorporation of additional rules considering data acquired in hydrogen-deuterium exchange experiments. This setup consists of two independent LC-MS/MS runs of one sample, where in the first run the acquisition is performed normally with undeuterated solvents (e.g., MeOH/H₂O) and during the second acquisition at least one of the mobile phases is replaced with a deuterated equivalent (e.g., MeOD/D₂O, ACN/D₂O). For each molecule in the sample two MS/MS spectra are acquired, one normal and one deuterated. Thus, a substantial further development is the processing of several query MS/MS spectra in a single MetFrag run which includes the annotation of fragments to m/z peaks of more than one spectra. Furthermore, additional rules are applied to correct theoretical masses of generated fragments by a variable number of exchanged hydrogens.

2.4 Candidate scoring

Retrieved candidates are assigned to the query MS/MS spectrum and prioritized by calculated scores. The higher the candidate score the better the fit to the spectrum. The original MetFrag approach uses information of assigned MS/MS peaks and annotated fragments such as intensities and BDEs of removed bonds. To integrate information

from additional analytical methods and data sources, the scoring function needs to be extended to allow additional scoring terms. Furthermore, it needs to be flexible and applicable irrespective of whether a high amount of additional information is available for a specific query MS/MS spectrum or not. Additionally, information included in the score calculation needs to be weighted depending on its importance and confidence.

Ruttkies et al. (2016) (Section 5.2) introduced a consensus scoring that includes additional weighted scoring terms. These scoring terms are calculated based on the additional information available. In the study the number of references and patents for a candidate, and the retention time retrieved from LC/MS have been used to exemplify the novel candidate scoring. To further enhance flexibility, extra terms, called “user-defined scores”, were tested by the incorporation of candidate scores calculated by CFM-ID. By using scores calculated on MS/MS information only, MetFrag and CFM-ID had 30 and 43 correct candidates ranked in first position, respectively, using PubChem as a candidate database. Including reference and retention information in MetFrag improved this to 420 and 336 correct candidates ranked first with ChemSpider and PubChem (89 and 71 %), respectively, and even up to 343 (PubChem) when combining with CFM-ID.

In Ruttkies et al. (2019b) (Section 5.3) the novel scoring function has been exploited to include additional information retrieved from hydrogen-deuterium exchange experiments. The three additional scoring terms developed in this study evaluated (1) the match of a candidate to the deuterated MS/MS spectrum, (2) the number of matching normal and deuterated fragment pairs and (3) the expected number of easily exchangeable hydrogens. On a set of 765 MS/MS spectral pairs (normal and deuterated) MetFrag could rank 104 instead of 72 of the correct candidates at first position when using the additional scoring terms. The number of correct candidates ranked among the top ten positions could even be increased from 345 to 481.

In Ruttkies et al. (2019a) (Section 5.4) the consensus scoring was complemented by using statistical approaches where annotations of m/z fragment peaks to fragment-structures were learned in a training step. Based on a Bayesian model, two additional scoring terms have been integrated and were evaluated on a test data set from CASMI 2016 contest consisting of 87 MS/MS spectra. The number of correct candidates ranked in first position increased from 5 to 21 and among the first ten from 39 to 55 both showing higher values than retrieved by CSI:IOKR the winner of this contest.

2.5 Candidate ranking

The scoring of molecular candidates results in a ranked candidate list for a query MS/MS spectrum where in the ideal case top ranked candidates give hints for the structure of the underlying precursor molecule. In high-throughput analysis the generation of

several thousand lists of ranked candidates is possible for which manual investigation is impractical. Typically, the quality of the candidate lists differs due to various reasons and thus the confidence of their meaningfulness. (1) Structural characteristics of the candidate molecules influence the number of generated fragments. So, there are candidates that produce more or less fragments in the *in silico* fragmentation step. (2) The quality of MS/MS spectra and their number of characteristic peaks can be different. Both factors (1) and (2) influence the number of fragments that can be annotated to the MS/MS spectrum which results in different ranges of calculated scores. Thus, distributions of candidate scores vary between different MS/MS spectra and candidate lists making it difficult to decide which ranges for which type of candidates are reliable. There is a huge scientific interest in approaches that assign confidence values to candidate lists, as they are known from MS/MS based peptide identification in proteomics Käll et al. (2008).

In Witting et al. (2017) (Section 5.5) a method is suggested to improve the ranking of candidates, which is also used to evaluate the reliability and confidence of candidate lists produced by MetFrag. In this case study, the approach is tested on MS/MS spectra from lipid molecular structures as they are categorized via a well-defined classification system. Moreover, investigations in this study revealed a characteristic distribution of calculated MetFrag scores for the investigated lipid main and sub classes. These score distributions have been used to develop classifiers for seven different lipid sub and main classes to differentiate ranked candidate lists of good and bad quality. With this approach, named LipidFrag, the number of false positive assignments could be reduced from 91 % to 57 % for positive ion mode and from 93 % to 27 % for negative mode on a reference data set of 960 MS/MS spectra originating from lipid molecular structures. Furthermore, comparison with LipidBlast, one of the most utilized tool for lipid spectra prediction, showed comparable results for both approaches where LipidFrag could annotate 819 and LipidBlast 716 of the MS/MS spectra.

2.6 Method evaluation

In order to evaluate their performance, developments presented in this section have been evaluated on experimental data in the related publications. Performances were also compared with state-of-the-art computational approaches. Besides these examinations, a more general comparison of available identification approaches is even more meaningful especially when performed on completely independent data and under standardized conditions. Ideally, each approach participating is optimized by its experts to guarantee a fair comparison and evaluation of the tools.

For this reason MetFrag participated in the open CASMI contests in the years 2012, 2013 and 2016 resulting in three peer-reviewed publications. The first two contests

took place in the early stages of my work. Thus, results submitted were dominated by the methods of the original MetFrag approach. Ruttkies et al. (2013) (Section 5.6) describes the results of MetFrag applied on the MS/MS data set of 15 spectra published in Category 2 of CASMI 2012, where the correct candidate of the first six challenges were known to be a natural product and the remaining from an environmental source. At this initial stage of my work, MetFrag reached a median rank of 280 for the natural product and 32 for the environmental spectra and a relative ranking position (RRP: calculated regarding the number of candidates where a value of 1 marks the best and a value of 0 the worst possible result) of 0.874 and 0.939. Moreover, MetFrag was teamed with an additional scoring term to improve ranks of candidates from biological origin in the natural product challenges 1-6. This improved the median rank to 145 and the RRP to 0.921.

In Schymanski et al. (2014a) (Section 5.7) MetFrag was combined with MetFusion (Gerlich and Neumann, 2013) and its usage of information from MassBank spectral library. The tools were applied on the 16 challenge MS/MS spectra published in Category 2 of CASMI 2013. Results were obtained by joining candidate lists retrieved from the three small molecule databases ChemSpider, PubChem and KEGG. Thus, the correct candidate was found in each candidate list for all 16 challenge MS/MS spectra, respectively. The aim of improving top rankings of the correct candidates compared with the previous CASMI contest was reached by ranking the correct candidate in first position in seven of the 16 challenges.

The CASMI 2016 contest was special as it provided 208 challenge MS/MS spectra, an amount perfectly suited for the evaluation of computational approaches. Schymanski et al. (2017a) (Section 5.8) was a joined publication by all participants to provide a general comparison of all participating computational approaches. MetFrag was used to produce results submitted in Category 2 (*in silico* fragmentation only) and Category 3 where also additional information beyond *in silico* fragmentation was allowed. In Category 2 the best performing methods used statistical approaches such as machine learning, which at that time was not yet part of MetFrag. In the post-contest evaluation of Category 3 MetFrag could benefit from the use of additional information such as retention time and the number of references to outperform all other participating approaches.

3 Discussion

3.1 Developed combined scoring principle is major enhancement for MetFrag

The scoring mechanism developed in my work, enables an easy integration of different information and data sources to assign better measures of quality to molecular candidates for a given query MS/MS spectrum. As an exemplary study, Ruttkies et al. (2015) (Section 5.1) integrates an additional scoring term to privilege potential metabolites (Peironcely et al., 2011) among the candidate structures from PubChem which includes both biological and non-biological compounds. This leads to improvements in the analysis for metabolomics experiments, which were confirmed with the results achieved in CASMI 2012 published in Ruttkies et al. (2013) (Section 5.6). While Ruttkies et al. (2015) (Section 5.1) represented a special case with the usage of a metabolite-likeness score, in Ruttkies et al. (2016) (Section 5.2) I was able to incorporate several scoring terms in a general approach used for various datasets and compound databases. With these scoring terms additional information such as retention time, literature and patent citations were integrated. I could successfully test the developed approach in CASMI 2016, which was published in Schymanski et al. (2017a) (Section 5.8). The general scoring approach I developed in Ruttkies et al. (2016) (5.2) was enhanced and utilized in Ruttkies et al. (2019b) (Section 5.3) on data obtained by isotopic labelling using hydrogen-deuterium exchange. This is the first study exploiting such data in a computational method such as MetFrag for the identification of small molecules in a high-throughput manner.

This flexible scoring mechanism can be regarded as major outcome of my work as it was the basis for several further enhancements. Even for future applications, MetFrag’s scoring approach can be easily extended by the use of “user-defined scores” allowing any kind of information to be included in case it can be represented numerically.

However, the question of which scoring terms contributed most to the improvement in the identification of small molecules could not be solved entirely. A partial analysis is provided in Ruttkies et al. (2016) (5.2), Schymanski et al. (2017a) (Section 5.8), Ruttkies et al. (2019b) (Section 5.3) and Ruttkies et al. (2019a) (Section 5.4). A study combining all scoring terms has not been performed yet which was due to the complexity and the

number of terms, as well as the different experimental contexts in the presented studies. To also integrate the information from isotopic labelling, MS/MS spectra retrieved by hydrogen-deuterium exchange experiments would be required for an assessment of the full set of scoring terms. The dataset analyzed in Ruttkies et al. (2019b) (Section 5.3) for which hydrogen-deuterium exchange information is already available could be used to combine the developed statistical scoring terms, additional experimental information, such as retention time and spectral libraries, and meta information.

The usage of meta information such as the number of citations and patents of a molecular candidate is highly dependent on the experimental context and question. Usually, meta information has no causative link to the experimental data, but in some contexts such as environmental screening, can provide valuable information on the relevance of certain candidates. For this reason, it should always be combined with methods relying on experimental data such as the query MS/MS spectrum and its usage should always be considered as a supporting (and not the only) method in the process of small molecule identification.

The integration of information obtained from additional experiments or analytical methods, such retention time or isotopic labelling can be included in the structure elucidation process in most cases. That being said, isotopic labelling experiments, such as hydrogen-deuterium exchange, have considerable additional experimental requirements, which are not always practical and feasible. However, if available and of good quality, this data can not only improve the ranking of the correct candidate but also provide more confidence.

An advantage of the statistical scoring implemented in Ruttkies et al. (2019a) (Section 5.4) is that it can be integrated independently of the availability of any additional data. As it only relies on the statistical model trained a priori and the mandatory query MS/MS spectrum, these scoring terms can be added in most applications. The integration of additional training spectra from different origin, such as electron impact, could even enlarge the application domain of the this approach.

3.2 Confidence scoring on lipid samples illustrates potential for broader application

In Witting et al. (2017) (Section 5.5) I took advantage of the hierarchical classification system of lipid molecular structures that categorizes lipids in different main and sub classes. Initially, we wanted to demonstrate the applicability of combinatorial fragmentation for the elucidation of molecular structures in the field of lipidomics. Interestingly, the fragmenter score implemented in Ruttkies et al. (2016) (Section 5.2), calculated on MS/MS spectra from representatives of different lipid classes, showed a characteristic

distribution for the investigated lipid main and sub classes. It can be expected that lipids from the same main or sub class show a similar fragmentation pattern. Thus, these distributions award a certain chemical meaning to the fragmenter score as it reflected these expected similarities. In addition, these distributions were the basis for the training of lipid class-dependent bayesian models. Given a query MS/MS spectrum, these models could assign probabilities to fragmenter scores, which were used to filter false positive assignments of candidates to improve the ranking of the correct candidate. Moreover, in high-throughput analysis with many MS/MS spectra, these probabilities were used to filter out unreliable lists of candidates that could be caused by insufficient spectral quality.

Due to the limited range of lipid classes this study was applied to, no reliable statements can be made whether the approach can be expanded to more diverse classes of compounds. However, this study can be treated as a showcase for a broader application in future. The well-defined classification system of lipids played a significant role in this study. A similar and more general classification system would be required for small molecules prior to further application. Chemical ontologies could be used for a broader classification such as Chemical Entities of Biological Interest (ChEBI), which was mainly built for metabolites (Hastings et al., 2016). ClassyFire (Feunang et al., 2016) or SODIAC (Bobach et al., 2012) are examples for tools developed for the automated classification of molecules. However, these approaches are inherently limited by integrated rules and the training set Sha et al. (2019). Moreover, it needs to be determined whether molecules grouped together by these approaches share similar fragmentation patterns. Another possibility might be to group molecules by their fragmentation patterns themselves. Clusters of MS/MS spectra as created by GNPS Wang et al. (2016a) could also show specific score distributions that could be used for statistical modelling.

3.3 Combinatorial structure generation as basis for enlargement of the MetFrag application

With the workflow I developed in Ruttkies et al. (2015) (Section 5.1) the integration of MS/MS data from GC/APCI-MS/MS experiments was possible. This workflow contained a rule-based and combinatorial method to create structures of molecular candidates that are only scantily covered by existing compound databases. My workflow for the generation of molecular structures was also applied in Allen et al. (2016) to enhance CFM-ID for the analysis of GC electron impact spectra. It could be reused in Ruttkies et al. (2019b) (Section 5.3) to alter structures of candidates to simulate the exchange of hydrogens with deuterium in hydrogen-deuterium exchange experiments.

The successful application of my strategies for this type of experiments encourages the integration of further labelling methods such as ^{13}C in future studies. However, the combinatorial nature of the algorithm induces an overgeneration of molecules resulting in many false positive candidate structures. A more sophisticated method that makes use of more chemical knowledge could help to reduce this overgeneration. Even statistical models could be considered, although training data might be difficult to find. The advantage of the rule-based approach is that if rules used for the combinatorial generation of candidate structures are chosen well, the chance of undergeneration is very low. Compared to overgeneration, the absence of a correct structure (false negative) would clearly reduce or even eliminate the chance for the correct annotation of the query MS/MS spectrum. The high amount of false positive structures as they occur when overgenerating could be discarded by appropriate scoring as applied by MetFrag in the presented applications.

3.4 Structural elucidation of small molecules remains a topic of interest

The community has made huge progress the last decade in developing computational methods for structural elucidation of small molecules based on MS/MS spectra. The progression of the evaluation performed in the CASMI contests over the years illustrates the improvements achieved until today. Broad evaluation measures like the median rank or the relative ranking position as they have been used in the early CASMI contests have been replaced by measures like the number of correctly ranked candidates. These metrics are more relevant especially for analytical scientists. Also the growth of available training data and its increasing quality has contributed to this positive development.

Despite this positive trend computational methods can still only be used as supporting methods as in most cases reference standards need to be used for confirmation of a putative identification (Schymanski et al., 2014b). Although this support already reduces the burden and costs for analysts a lot, the final goal of assigning a single compound to the spectrum with high confidence seems still far away. Most computational methods perform a “soft” assignment of putative candidates by using calculated scores. A hard filter even with high confidence might cause a loss of the correct structure. So, the last decade the community followed the goal to bring as many correct hits among the top positions as possible because scarcely anybody investigates the entire candidate list. As we have learned from Google search results, most people never visit the second page and even expect the best hit to be in first position.

The usage of spectral libraries and small molecule databases also remains an important factor that limits computational approaches. Although, the usage of spectral

reference data for structure elucidation is powerful (independent whether used directly or in machine learning approaches), the coverage compared to the size of the entire chemical space is rather small. Although this problem becomes smaller when relying on structure databases there is still a relevant chance of missing the correct candidate due to an unknown compound. This is why *de novo* methods that are independent of the content available in small molecule databases are of interest Stravs et al. (2021).

3.5 Controlled evaluation studies for computational methods are insightful

An important driver and benchmark for the developments of computational methods in the last decade were the CASMI contests. These contests can be considered for being an excellent assessment of the current state of progress made during my work. The independent comparison of results submitted by research groups all over the world additionally showed the current state among the entire community. Moreover, the discussions and the exchange of expertise during and after a contest was an important driving force for further improvements. Thus, CASMI provided important data for the testing and comparison of the developed approaches.

CASMI 2016 played a particular role for the comparison of computational approaches among all CASMI contests. This can not only be attributed to the relatively high amount of available high quality MS/MS spectra. Additionally, for the first time training data was provided to be used for optimization of parameters required by statistical methods. Moreover, retention time data was given to be included in the identification process. The preparation and provision of fixed candidate lists for each challenge provided common ground for a valid comparison of participating computational approaches. Although a lot has been done to achieve this fair and valid comparison, the use of different training data sets by the participating statistical methods could not be avoided in the end. To overcome this problem, organizers encouraged a post contest resubmission of results by participants who used statistical methods trained on a joint training data set. The analysis obtained by this resubmission was an important outcome of Schymanski et al. (2017a) (Section 5.8) as the method comparison is of paramount importance. Even until today spectral data provided in CASMI 2016 is still used for the development and the evaluation of current computational methods (Li et al., 2020; Fan et al., 2020).

In CASMI 2016, the methods that have been developed in Ruttkies et al. (2016) (Section 5.2) were successfully tested in Category 3 (additional results) where it could outperform all other participants. In an extra study, I could enhance MetFrag by the integration of a statistical model published in Ruttkies et al. (2019a) (Section 5.4), combining combinatorial fragmentation and statistical methods and could outperform

post contest submitted results of other statistical approaches for Category 2. The combination showed the best performance for the negative mode test data among all participants and thus can overcome the unpleasant but usual situation when only few training data is available. Thus, I consider the combination of statistical models and combinatorial fragmentation as an important outcome of my work.

4 Conclusion

During my doctoral studies, I developed a novel and flexible approach that exploits combinatorial fragmentation and combined additional analytical data, statistical methodologies and different information sources to improve annotation of molecular structures on the basis of MS/MS data. This resulted in a completely refactored and extended version of the MetFrag software (Wolf et al., 2010) that now is able to exploit the full potential of combinatorial fragmentation. The achievements of my studies have been published in several peer-reviewed publications that represent the essence of my work.

My developed strategies outperformed not only combinatorial fragmentation implemented in existing software tools but also competing state-of-the-art approaches. The work I published together with Schymanski *et al.* (Ruttkies et al., 2016) formed the basis for further developments and applications presented in this thesis and for the MetFrag approach. With 546 citations¹ this work has a large impact in the community working on or using MS/MS-based computational methods for the identification of small molecules. Initially published in 2010, MetFrag has now celebrated its tenth anniversary and this milestone marks a time of fundamental enhancements and an appreciation of its role as a key playing tool in the field of metabolomics and beyond. However, this is not only indicated by the high citation rate of Ruttkies et al. (2016), but also by the many and increasing application papers using MetFrag beyond close collaborators, and continuing developments beyond the work of my studies. During this decade, the performance of computational small molecule identification has been optimized in a way that evaluation metrics like the median rank or the relative ranking position as predominantly reported for CASMI 2012, has receded into background. Measures like the number of correctly top ranked candidates, previously too small to be usable, have been proven to be more relevant as performance improved. These achievements can be regarded as a milestone in computational small molecule identification. Furthermore, the size of spectral datasets used for evaluation has increased drastically. This shows how practical tools have become so that they can be used for real world data. In this regard, I was able to enlarge MetFrag’s application domain during my studies from metabolomics to data sets originating from environmental science experiments. The success in combining different information and data sources in the process of small molecule identification made MetFrag a solution used by US EPA’s Chemistry Dash-

¹<https://scholar.google.com> (accessed on 04/2020)

board (McEachran et al., 2018) and Bruker’s MetaboScape software (Bruker, 2020). My work has successfully illustrated how to integrate additional experimental data from different analytical methods such as isotopic labelling and statistical methods with computer assisted combinatorial fragmentation. Developed approaches are in common use by the community and are incentives to further integrate different data sources and analytical methods to accelerate computer assisted identification of small molecules.

References

- Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207.
- Allen, F., Greiner, R., and Wishart, D. (2014). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*.
- Allen, F., Pon, A., Greiner, R., and Wishart, D. (2016). Computational prediction of electron ionization mass spectra to assist in gc/ms compound identification. *Analytical chemistry*, 88(15):7689–7697.
- Balogh, M. (2004). Debating resolution and mass accuracy in mass spectrometry. *SPECTROSCOPY-SPRINGFIELD THEN EUGENE THEN DULUTH-*, 19:34–34.
- Bauer, C. A. and Grimme, S. (2016). How to compute electron ionization mass spectra from first principles. *The Journal of Physical Chemistry A*, 120(21):3755–3766.
- Benjamins, R. and Scheres, B. (2008). Auxin: the looping star in plant development. *Annu. Rev. Plant Biol.*, 59:443–465.
- Berglund, M. and Wieser, M. E. (2011). Isotopic compositions of the elements 2009 (iupac technical report). *Pure and applied chemistry*, 83(2):397–410.
- Biemann, K. (1979). The implications and limitations of the findings of the viking organic analysis experiment. *Journal of Molecular Evolution*, 14(1-3):65–70.
- Bobach, C., Böhme, T., Laube, U., Püschel, A., and Weber, L. (2012). Automated compound classification using a chemical ontology. *Journal of cheminformatics*, 4(1):1–12.
- Borges, R. M., Colby, S. M., Das, S., Edison, A. S., Fiehn, O., Kind, T., Lee, J., Merrill, A. T., Merz Jr, K. M., Metz, T. O., et al. (2021). Quantum chemistry calculations for metabolomics: Focus review. *Chemical reviews*.
- Brock, T. D. (1999). *Robert Koch: a life in medicine and bacteriology*. Zondervan.
- Brouard, C., Shen, H., Dührkop, K., d’Alché Buc, F., Böcker, S., and Rousu, J. (2016). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36.

-
- Bruker (2020). MetaboScape 4.0.
- Carr, S. and Burlingame, A. (1996). The meaning and usage of the terms monoisotopic mass, average mass, mass resolution, and mass accuracy for measurements of biomolecules. appendix xi. *Mass Spectrometry in the Biological Sciences, Burlingame, AL and Carr, SA, Eds., Humana Press, Totowa, NJ*, pages 546–553.
- Cautereels, J., Claeys, M., Geldof, D., and Blockhuys, F. (2016). Quantum chemical mass spectrometry: ab initio prediction of electron ionization mass spectra and identification of new fragmentation pathways. *Journal of Mass Spectrometry*, 51(8):602–614.
- Chernushevich, I. V., Loboda, A. V., and Thomson, B. A. (2001). An introduction to quadrupole–time-of-flight mass spectrometry. *Journal of mass spectrometry*, 36(8):849–865.
- de Hoffmann, E. (1996). Tandem mass spectrometry: A primer. *Journal of Mass Spectrometry*, 31(2):129–137.
- Demain, A. L. and Fang, A. (2000). The natural functions of secondary metabolites. *History of modern biotechnology I*, pages 1–39.
- Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78.
- Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585.
- Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*.
- Ettre, L. and Sakodynskii, K. (1993). Ms tswett and the discovery of chromatography i: Early work (1899–1903). *Chromatographia*, 35(3-4):223–231.
- Fahy, E., Subramaniam, S., Brown, H. A., Glass, C. K., Merrill Jr, A. H., Murphy, R. C., Raetz, C. R., Russell, D. W., Seyama, Y., Shaw, W., et al. (2005). A comprehensive classification system for lipids. *European journal of lipid science and technology*, 107(5):337–364.
- Fan, Z., Alley, A., Ghaffari, K., and Ransom, H. W. (2020). Metfid: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics*, 16(10):1–11.
- Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., and Whitehouse, C. M. (1989). Electrospray ionization for mass spectrometry of large biomolecules. *Science*, 246(4926):64–71.

- Feunang, Y. D., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., et al. (2016). Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics*, 8(1):61.
- Fiehn, O. (2002). Metabolomics — the link between genotypes and phenotypes. *Plant Molecular Biology*, v.48, 155-177 (2002), 48.
- Frolkis, A., Knox, C., Lim, E., Jewison, T., Law, V., Hau, D. D., Liu, P., Gautam, B., Ly, S., Guo, A. C., et al. (2010). Smpdb: the small molecule pathway database. *Nucleic acids research*, 38(suppl_1):D480–D487.
- Gasmi, A., Peana, M., Noor, S., Lysiuk, R., Menzel, A., Benahmed, A. G., and Bjørklund, G. (2021). Chloroquine and hydroxychloroquine in the treatment of covid-19: the never-ending story. *Applied microbiology and biotechnology*, pages 1–11.
- Gerlich, M. and Neumann, S. (2013). Metfusion: integration of compound identification strategies. *Journal of Mass Spectrometry*, 48(3):291–298.
- German, J. B., Hammock, B. D., and Watkins, S. M. (2005). Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics*, 1(1):3–9.
- Gohlke, R. S. (1959). Time-of-flight mass spectrometry and gas-liquid partition chromatography. *Analytical Chemistry*, 31(4):535–541.
- Gohlke, R. S. and McLafferty, F. W. (1993). Early gas chromatography/mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 4(5):367–371.
- Griffiths, J. (2008). A brief history of mass spectrometry. *Anal. Chem*, 80(15):5678–5683.
- Großkinsky, D., Edelsbrunner, K., Pfeifhofer, H., Van der Graaff, E., and Roitsch, T. (2013). Cis-and trans-zeatin differentially modulate plant immunity. *Plant signaling & behavior*, 8(7):e24798.
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). Chebi in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(D1):D1214–D1219.
- Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., Mahendrakar, T., Williams, M., Neumann, S., Rocca-Serra, P., et al. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic acids research*, 41(D1):D781–D786.

-
- Hayden, B., Bowdler, S., Butzer, K. W., Cohen, M. N., Druss, M., Dunnell, R. C., Goodyear, A. C., Hardesty, D. L., Hassan, F. A., Kamminga, J., et al. (1981). Research and development in the stone age: technological transitions among hunter-gatherers [and comments and reply]. *Current Anthropology*, 22(5):519–548.
- Heinonen, M., Shen, H., Zamboni, N., and Rousu, J. (2012). Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*, 28(18):2333–2341.
- Ho, C. S., Lam, C., Chan, M., Cheung, R., Law, L., Lit, L., Ng, K., Suen, M., and Tai, H. (2003). Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical Biochemist Reviews*, 24(1):3.
- Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714.
- Horai, H., Arita, M., and Nishioka, T. (2008). Comparison of esi-ms spectra in massbank database. In *2008 International Conference on BioMedical Engineering and Informatics*, volume 2, pages 853–857. IEEE.
- Issaq, H. J. (2001). *A century of separation science*. CRC Press.
- Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of proteome research*, 7(01):29–34.
- Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.
- Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. *Nucleic acids research*, 34(suppl_1):D354–D357.
- Karas, M., Bachmann, D., and Hillenkamp, F. (1985). Influence of the wavelength in high-irradiance ultraviolet laser desorption mass spectrometry of organic molecules. *Analytical chemistry*, 57(14):2935–2939.
- Karasek, F. W. and Clement, R. E. (2012). *Basic gas chromatography-mass spectrometry: principles and techniques*. Elsevier.
- Katajamaa, M. and Orešič, M. (2007). Data processing for mass spectrometry-based metabolomics. *Journal of chromatography A*, 1158(1-2):318–328.
- Keunen, E., Schellingen, K., Vangronsveld, J., and Cuypers, A. (2016). Ethylene and metal stress: small molecule, big impact. *Frontiers in plant science*, 7:23.

- Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. (2016). Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213.
- Kind, T. and Fiehn, O. (2007). Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*, 8(1):105.
- Kind, T., Liu, K.-H., Lee, D. Y., DeFelice, B., Meissen, J. K., and Fiehn, O. (2013). Lipidblast in silico tandem mass spectrometry database for lipid identification. *Nature methods*, 10(8):755.
- Kind, T., Tsugawa, H., Cajka, T., Ma, Y., Lai, Z., Mehta, S. S., Wohlgemuth, G., Barupal, D. K., Showalter, M. R., Arita, M., et al. (2018). Identification of small molecules using accurate mass ms/ms search. *Mass spectrometry reviews*, 37(4):513–532.
- Kloepfer, A., Gnirss, R., Jekel, M., and Reemtsma, T. (2004). Occurrence of benzothiazoles in municipal wastewater and their fate in biological treatment. *Water Science and Technology*, 50(5):203–208.
- Kopka, J., Schauer, N., Krueger, S., Birkemeyer, C., Usadel, B., Bergmüller, E., Dörmann, P., Weckwerth, W., Gibon, Y., Stitt, M., et al. (2005). Gmd@ csb. db: the golm metabolome database. *Bioinformatics*, 21(8):1635–1638.
- Lee, D.-K., Yoon, M. H., Kang, Y. P., Yu, J., Park, J. H., Lee, J., and Kwon, S. W. (2013). Comparison of primary and secondary metabolites for suitability to discriminate the origins of schisandra chinensis by gc/ms and lc/ms. *Food chemistry*, 141(4):3931–3937.
- Lev-Yadun, S., Gopher, A., and Abbo, S. (2000). The cradle of agriculture. *Science*, 288(5471):1602–1603.
- Li, Y., Kuhn, M., Gavin, A.-C., and Bork, P. (2020). Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features. *Bioinformatics*, 36(4):1213–1218.
- Loos, M., Gerber, C., Corona, F., Hollender, J., and Singer, H. (2015). Accelerated isotope fine structure calculation using pruned transition trees. *Analytical chemistry*, 87(11):5738–5744.
- Lundanes, E., Reubsæet, L., and Greibrokk, T. (2013). *Chromatography: basic principles, sample preparations and related methods*. John Wiley & Sons.
- Martin, A. J. P. (1953). *The development of partition chromatography*. Norstedt.

- Massalha, H., Korenblum, E., Tholl, D., and Aharoni, A. (2017). Small molecules below-ground: the role of specialized metabolites in the rhizosphere. *The plant journal*, 90(4):788–807.
- Masters, B. R. (2008). *History of the Optical Microscope in Cell Biology and Medicine*. American Cancer Society.
- Mayer, I. and Gömöry, Á. (1993). Use of energy partitioning for predicting primary mass spectrometric fragmentation steps: A preliminary account. *International Journal of Quantum Chemistry*, 48(S27):599–605.
- Mayer, I. and Gömöry, Á. (1994). Semiempirical quantum chemical method for predicting mass spectrometric fragmentations. *Journal of Molecular Structure: THEOCHEM*, 311:331–341.
- McEachran, A. D., Mansouri, K., Grulke, C., Schymanski, E. L., Ruttkies, C., and Williams, A. J. (2018). “ms-ready” structures for non-targeted high-resolution mass spectrometry screening studies. *Journal of cheminformatics*, 10(1):45.
- Moldoveanu, S. C. and David, V. (2018). Derivatization methods in gc and gc/ms. *Gas Chromatography-Derivatization, Sample Preparation, Application*.
- Mortimer, C., Müller, U., and Beck, J. (2015). *Chemie: Das Basiswissen der Chemie*. Thieme.
- Muller, P. (1994). Glossary of terms used in physical organic chemistry (iupac recommendations 1994). *Pure and Applied Chemistry*, 66(5):1077–1184.
- Murray, K. K., Boyd, R. K., Eberlin, M. N., Langley, G. J., Li, L., and Naito, Y. (2013). Definitions of terms relating to mass spectrometry (iupac recommendations 2013). *Pure and Applied Chemistry*, 85(7):1515–1609.
- Nathanson, J. A. (1984). Caffeine and related methylxanthines: possible naturally occurring pesticides. *Science*, 226(4671):184–187.
- NCBI, N. C. f. B. I. (2020). Pubchem database. <https://pubchem.ncbi.nlm.nih.gov>.
- Neumann, N. K., Lehner, S. M., Kluger, B., Bueschl, C., Sedelmaier, K., Lemmens, M., Krska, R., and Schuhmacher, R. (2014). Automated lc-hrms (/ms) approach for the annotation of fragment ions derived from stable isotope labeling-assisted untargeted metabolomics. *Analytical chemistry*, 86(15):7320–7327.
- Nobeli, I. and Thornton, J. M. (2006). A bioinformatician’s view of the metabolome. *Bioessays*, 28(5):534–545.

- Peironcelly, J. E., Reijmers, T., Coulier, L., Bender, A., and Hankemeier, T. (2011). Understanding and Classifying Metabolite Space and Metabolite-Likeness. *PLoS ONE*, 6(12):e28966.
- Pence, H. E. and Williams, A. (2010). Chempidder: an online chemical information resource.
- Plantone, D. and Koudriavtseva, T. (2018). Current and future use of chloroquine and hydroxychloroquine in infectious, immune, neoplastic, and neurological diseases: a mini-review. *Clinical drug investigation*, 38(8):653–671.
- Reemtsma, T. (2000). Determination of 2-substituted benzothiazoles of industrial use from water by liquid chromatography/electrospray ionization tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, 14(17):1612–1618.
- Reymond, J.-L. and Awale, M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS chemical neuroscience*, 3:649–57.
- Ridder, L., van der Hooft, J. J. J., and Verhoeven, S. (2014). Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrometry*, 3(Special Issue 2):S0033–S0033.
- RSC, R. S. o. C. (2020). ChemSpider. <http://www.chemspider.com/>.
- Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875.
- Ruttkies, C., Gerlich, M., and Neumann, S. (2013). Tackling casmi 2012: Solutions from metfrag and metfusion. *Metabolites*, 3(3):623–636.
- Ruttkies, C., Neumann, S., and Posch, S. (2019a). Improving metfrag with statistical learning of fragment annotations. *BMC bioinformatics*, 20(1):376.
- Ruttkies, C., Schymanski, E. L., Strehmel, N., Hollender, J., Neumann, S., Williams, A. J., and Krauss, M. (2019b). Supporting non-target identification by adding hydrogen deuterium exchange ms/ms capabilities to metfrag. *Analytical and bioanalytical chemistry*, 411(19):4683–4700.
- Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., and Neumann, S. (2016). Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, 8(1):3.
- Ruttkies, C., Strehmel, N., Scheel, D., and Neumann, S. (2015). Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem

- mass spectrometry data using an in silico generated compound database and metfrag. *Rapid Communications in Mass Spectrometry*, 29(16):1521–1529.
- Sargent, J. R., Tocher, D. R., and Bell, J. G. (2003). The lipids. In *Fish nutrition*, pages 181–257. Elsevier.
- Schäfer, M., Brütting, C., Meza-Canales, I. D., Großkinsky, D. K., Vankova, R., Baldwin, I. T., and Meldau, S. (2015). The role of cis-zeatin-type cytokinins in plant growth regulation and mediating responses to environmental interactions. *Journal of experimental botany*, 66(16):4873–4884.
- Scheubert, K., Hufsky, F., and Böcker, S. (2013). Computational mass spectrometry for small molecules. *Journal of cheminformatics*, 5(1):12.
- Schmieder, R., Hoffmann, J., Becker, M., Bhargava, A., Müller, T., Kahmann, N., Ellinghaus, P., Adams, R., Rosenthal, A., Thierauch, K.-H., et al. (2014). Regorafenib (bay 73-4506): antitumor and antimetastatic activities in preclinical models of colorectal cancer. *International journal of cancer*, 135(6):1487–1496.
- Schreiber, S. L. (2005). Small molecules: the missing link in the central dogma. *Nature chemical biology*, 1(2):64–66.
- Schymanski, E. L., Gerlich, M., Ruttkies, C., and Neumann, S. (2014a). Solving casmi 2013 with metfrag, metfusion and molgen-ms/ms. *Mass spectrometry*, 3(Special_Issue_2):S0036–S0036.
- Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., and Hollender, J. (2014b). Identifying small molecules via high resolution mass spectrometry: communicating confidence.
- Schymanski, E. L. and Neumann, S. (2013). Casmi: And the winner is... *Metabolites*, 3(2):412–439.
- Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., et al. (2017a). Critical assessment of small molecule identification 2016: automated methods. *Journal of cheminformatics*, 9(1):22.
- Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F., Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T., Fiehn, O., Ghesquière, B., and Neumann, S. (2017b). Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9(1):22.

- Sha, B., Schymanski, E. L., Ruttkies, C., Cousins, I. T., and Wang, Z. (2019). Exploring open cheminformatics approaches for categorizing per-and polyfluoroalkyl substances (pfass). *Environmental Science: Processes & Impacts*, 21(11):1835–1851.
- Shimadzu (2019). Fundamental guide to liquid chromatography-mass spectrometry (lcms). <https://www.shimadzu.com/an/lcms/support/fundamental/index.html>. Accessed: 2020-01-29.
- Shrivastava, A. D., Swainston, N., Samanta, S., Roberts, I., Wright Muelas, M., and Kell, D. B. (2021). Massgenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793.
- Stein, S. (2012). Mass spectral reference libraries: an ever-expanding resource for chemical identification.
- Stravs, M. A., Dührkop, K., Böcker, S., and Zamboni, N. (2021). Msnovelist: De novo structure generation from mass spectra. *bioRxiv*.
- Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., Merrill Jr, A. H., Murphy, R. C., Raetz, C. R., Russell, D. W., et al. (2007). Lmsd: Lipid maps structure database. *Nucleic acids research*, 35(suppl_1):D527–D532.
- Sumner, L. W., Urbanczyk-Wochniak, E., and Broeckling, C. D. (2005). Metabolomics data analysis, visualization, and integration. In *Plant bioinformatics*, pages 409–436. Springer.
- Sussulini, A. (2017). *Metabolomics: from fundamentals to clinical applications*, volume 965. Springer.
- Tabet, J.-C. and Rebuffat, S. (2003). Nobel prize 2002 for chemistry: mass spectrometry and nuclear magnetic resonance. *Medecine Sciences: M/S*, 19(8-9):865–872.
- Thermo Scientific (2020). MassFrontier. <https://www.thermofisher.com/de/de/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/mass-frontier-spectral-interpretation-software.html>. Accessed: 2020-04-13.
- Tian, M., Liu, T., Wu, X., Hong, Y., Liu, X., Lin, B., and Zhou, Y. (2019). Chemical composition, antioxidant, antimicrobial and anticancer activities of the essential oil from the rhizomes of zingiber striolatum diels. *Natural Product Research*, pages 1–5.
- Tsugawa, H., Kind, T., Nakabayashi, R., Yukihiro, D., Tanaka, W., Cajka, T., Saito, K., Fiehn, O., and Arita, M. (2016). Hydrogen rearrangement rules: computational ms/ms fragmentation and structure elucidation using ms-finder software. *Analytical chemistry*, 88(16):7946–7958.

-
- Vermeulen, R., Schymanski, E. L., Barabási, A.-L., and Miller, G. W. (2020). The exposome and health: Where chemistry meets biology. *Science*, 367(6476):392–396.
- Villas-Bôas, S. G., Mas, S., Åkesson, M., Smedsgaard, J., and Nielsen, J. (2005). Mass spectrometry in metabolome analysis. *Mass spectrometry reviews*, 24(5):613–646.
- Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., and Yanes, O. (2016). Mass spectral databases for lc/ms-and gc/ms-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry*, 78:23–35.
- Wang, M., Carver, J., Phelan, V., Sanchez, L., Garg, N., Peng, Y., Nguyen, D., Watrous, J., Kapon, C., Luzzatto Knaan, T., Porto, C., Bouslimani, A., Melnik, A., Meehan, M., Liu, W.-T., Crüsemann, M., Boudreau, P., Esquenazi, E., Sandoval-Calderón, M., and Bandeira, N. (2016a). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature Biotechnology*, 34:828–837.
- Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapon, C. A., Luzzatto-Knaan, T., et al. (2016b). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837.
- Wang, Y., Kora, G., Bowen, B. P., and Pan, C. (2014). Midas: A database-searching algorithm for metabolite identification in metabolomics. *Analytical Chemistry*, 86(19):9496–9503. PMID: 25157598.
- Weindl, D., Wegner, A., and Hiller, K. (2015). Metabolome-wide analysis of stable isotope labeling—is it worth the effort? *Frontiers in physiology*, 6:344.
- Wieser, M. E. and Berglund, M. (2009). Atomic weights of the elements 2007 (iupac technical report). *Pure and Applied Chemistry*, 81(11):2131–2156.
- Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., et al. (2017). The comptox chemistry dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics*, 9(1):1–27.
- Williams, J. P., Patel, V. J., Holland, R., and Scrivens, J. H. (2006). The use of recently described ionisation techniques for the rapid analysis of some common drugs and samples of biological origin. *Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry*, 20(9):1447–1456.
-

- Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., et al. (2007). Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl_1):D521–D526.
- Witting, M., Ruttkies, C., Neumann, S., and Schmitt-Kopplin, P. (2017). Lipid-frag: Improving reliability of in silico fragmentation of lipids and application to the caenorhabditis elegans lipidome. *PloS one*, 12(3).
- Wolf, S., Schmidt, S., Müller-Hannemann, M., and Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11(1):148.
- Yost, R. and Enke, C. (1978). Selected ion fragmentation with a tandem quadrupole mass spectrometer. *Journal of the American Chemical Society*, 100(7):2274–2275.

5 Peer-reviewed publications

In the following, the peer-reviewed manuscripts published in the preparation process of my thesis are listed. Main authors are underlined and my name is marked in boldface. The copyright declaration for the publications is given together with a short description of the contribution of coauthors, where appropriate.

5.1 Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an *in silico* generated compound database and MetFrag

Christoph Ruttkies, Nadine Strehmel, Dierk Scheel, Steffen Neumann. Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an *in silico* generated compound database and MetFrag. *Rapid Commun. Mass Spectrom.*, 29: 1521-1529, 2015. 15 Citations¹
<https://onlinelibrary.wiley.com/doi/abs/10.1002/rcm.7244>

Contributions

I designed the study together with Nadine Strehmel, who did the experimental lab work, data preprocessing and extraction of the MS/MS peak lists. I developed the computational pipeline, analyzed the data and created the figures describing the workflow and presenting the results. Nadine Strehmel, Steffen Neumann and I prepared the manuscript. The work was supervised and coordinated by Dierk Scheel and Steffen Neumann.

Copyright

I hereby declare that the copyright of this publication is by John Wiley and Sons and the Copyright Clearance Center. The full-text article can be found under the above mentioned URL.

¹<https://scholar.google.com> (accessed on 01/2021)

5.2 MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation

Christoph Ruttkies, [Emma L. Schymanski](#), Sebastian Wolf, Juliane Hollender, Steffen Neumann. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. Journal of Cheminformatics, 8, 3, 2016. 546 Citations²
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-016-0115-9>

Contributions

I designed the study, integrated all features and prepared the manuscript together with Emma L. Schymanski, who also prepared the datasets. Sebastian Wolf developed the original MetFrag software under supervision of Steffen Neumann. I completely refactored the MetFrag pipeline and developed the scoring terms. The work was supervised and coordinated by Juliane Hollender and Steffen Neumann.

²<https://scholar.google.com> (accessed on 01/2021)

SOFTWARE

Open Access



MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation

Christoph Ruttkies^{1††}, Emma L. Schymanski^{2†}, Sebastian Wolf^{1,4}, Juliane Hollender^{2,3} and Steffen Neumann¹**Abstract**

Background: The *in silico* fragmenter MetFrag, launched in 2010, was one of the first approaches combining compound database searching and fragmentation prediction for small molecule identification from tandem mass spectrometry data. Since then many new approaches have evolved, as has MetFrag itself. This article details the latest developments to MetFrag and its use in small molecule identification since the original publication.

Results: MetFrag has gone through algorithmic and scoring refinements. New features include the retrieval of reference, data source and patent information via ChemSpider and PubChem web services, as well as InChIKey filtering to reduce candidate redundancy due to stereoisomerism. Candidates can be filtered or scored differently based on criteria like occurrence of certain elements and/or substructures prior to fragmentation, or presence in so-called "suspect lists". Retention time information can now be calculated either within MetFrag with a sufficient amount of user-provided retention times, or incorporated separately as "user-defined scores" to be included in candidate ranking. The changes to MetFrag were evaluated on the original dataset as well as a dataset of 473 merged high resolution tandem mass spectra (HR-MS/MS) and compared with another open source *in silico* fragmenter, CFM-ID. Using HR-MS/MS information only, MetFrag2.2 and CFM-ID had 30 and 43 Top 1 ranks, respectively, using PubChem as a database. Including reference and retention information in MetFrag2.2 improved this to 420 and 336 Top 1 ranks with ChemSpider and PubChem (89 and 71 %), respectively, and even up to 343 Top 1 ranks (PubChem) when combining with CFM-ID. The optimal parameters and weights were verified using three additional datasets of 824 merged HR-MS/MS spectra in total. Further examples are given to demonstrate flexibility of the enhanced features.

Conclusions: In many cases additional information is available from the experimental context to add to small molecule identification, which is especially useful where the mass spectrum alone is not sufficient for candidate selection from a large number of candidates. The results achieved with MetFrag2.2 clearly show the benefit of considering this additional information. The new functions greatly enhance the chance of identification success and have been incorporated into a command line interface in a flexible way designed to be integrated into high throughput workflows. Feedback on the command line version of MetFrag2.2 available at <http://c-ruttkies.github.io/MetFrag/> is welcome.

Keywords: Compound identification, *In silico* fragmentation, High resolution mass spectrometry, Metabolomics, Structure elucidation

Background

The identification of unknown small molecules from mass spectral data is one of the most commonly-mentioned bottlenecks in several scientific fields, including

metabolomic, forensic, environmental, pharmaceutical and medical sciences. Recent developments to high resolution, accurate mass spectrometry coupled with chromatographic separation has revolutionized high-throughput analysis and opened up whole new ranges of substances that can be detected at ever decreasing detection limits. However, where "peak inventories" are reported, the vast majority of the substances or peaks detected in samples typically remain unidentified [1–3]. Although targeted analysis, where a reference standard is available, remains

*Correspondence: cruttkie@ipb-halle.de

[†]Christoph Ruttkies, Emma L. Schymanski contributed equally to this work

¹Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany
Full list of author information is available at the end of the article



© 2016 Ruttkies et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

the best way to confirm the identification of a compound, it is no longer possible to have access to reference standards for the 100s–1000s of substances of interest in complex samples. While mass spectral libraries are growing for high accuracy tandem and MS^n spectra, the coverage is still relatively small compared with the number of compounds that could potentially be present in typical samples [4, 5]. Thus, for substances without reference standards or not present in the spectral libraries, the challenge of identification still remains. This has spurred activities in computational mass spectrometry, aimed at proposing tentative identifications for the cases where the mass spectrum is not (yet) in a mass spectral library.

The *in silico* fragmenter MetFrag, launched in 2010, was one of the first approaches to address this niche for accurate tandem mass spectra in a fast, combinatorial manner [6]. The MetFrag workflow starts by retrieving candidate structures from the compound databases PubChem [7], ChemSpider [8] or KEGG [9, 10], or accepting the upload of a structure data file (SDF) containing candidates. Candidates are then fragmented using a bond dissociation approach and these fragments are compared with the product ions in the measured mass spectrum to determine which candidates best explain the measured data. The candidate scoring is a function of the mass to charge ratio (m/z), intensity and bond dissociation energy (BDE) of the matched peaks, while a limited number of neutral loss rules (5 in total) account for rearrangements [6]. Searching PubChem, the original MetFrag (hereafter termed “MetFrag2010” for readability) achieved a median rank of 8 (with an average of 338 candidates per compound) when restricted to a Feb. 2006 version of PubChem, and 31.5 querying PubChem in 2009 (average of 2508 candidates per compound) on a 102 compound dataset from Hill et al. [11]. As PubChem is now double the size of the 2009 version, the candidate ranking becomes more challenging over time due to the increase in numbers of candidates. Thus, innovations are required to improve performance and efficiency.

Other methods for *in silico* fragmentation are also available. The commercial software Mass Frontier [12] uses rule-based fragmentation prediction based on standard reactions, a comprehensive library of over 100,000 fragmentation rules, or both. The approaches of MetFrag and Mass Frontier are complementary and have been used in combination to support structure elucidation [13, 14], but Mass Frontier does not perform candidate retrieval or scoring by itself. With increasing amounts of data available, machine learning approaches have been used to train models of the fragmentation process. Heinonen et al. [15] introduced FingerID, which uses a support vector machine to learn the mapping between the mass spectra and molecular fingerprints of

the candidates. Allen et al. [16] use a stochastic, generative Markov model for the fragmentation. Implemented in CFM-ID (competitive fragment modelling), this can be used to assign fragments to spectra to rank the candidates, but also to predict spectra from structures alone. The MAGMa algorithm [17] includes information from MS^n fragmentation data, but also uses the number of references as an additional scoring term. The latest fragmenter, CSI:FingerID combines fragmentation trees and molecular fingerprinting to achieve up to 39 % Top 1 ranks, outperforming all other fragmenters [18]. The MetFusion [19] approach takes advantage of the availability of spectral data for some compounds and performs a combined query of both MetFrag and MassBank [20], such that the scores of candidates with high chemical similarity to high-scoring reference spectra are increased.

Lessons from recent critical assessment of small molecule identification contests (CASMI) [21, 22], which included many of the above-mentioned algorithms, show that the use of smaller, specific databases greatly improves the chance of obtaining the correct answer ranked highly and that the winners gathered information from many different sources, rather than relying on the *in silico* fragmentation alone. Furthermore, performing candidate selection by molecular formula can risk losing the correct candidate if the formula prediction is not certain, such that an exact mass search can be more appropriate in cases where more than one formula is possible. Despite the progress achieved for *in silico* fragmentation approaches, there are still some fundamental limitations to mass spectrometry that mean that candidate ranking cannot be solved by fragment prediction alone. For example, mass spectra that are dominated by one or only a few fragments (e.g. a water loss) that can be explained by most of the candidates simply do not contain enough information to distinguish candidates. Further examples and limitations are discussed extensively in [4].

The aim of MetFrag2.2 was to incorporate many additional features into the original MetFrag *in silico* fragmenter, considering all the information presented above. Features to explicitly include or exclude combinations of elements and substructures by either filtering or scoring were added. Suspect screening approaches, growing in popularity in environmental analysis [1], were also incorporated to allow users to screen large databases (i.e. PubChem and ChemSpider) while being able to check for candidates present in smaller, more specific databases (e.g. KEGG [9], HMDB [23], STOFF-IDENT [24], MassBank [20] or NORMAN suspects [25]), enabling users to “flag” potential structures of interest. The number of references, data sources and/or patents for a substance are now accessible via PubChem and/or ChemSpider web services, and a PubChem reference score has already

been included in the MAGMa web interface [26]. A high number of literature references or patent listings may indicate that the substance is of high use and thus more likely to be found in the environment. Similarly, a higher number of scientific articles for a metabolite could indicate that this has been observed in biological samples before. Reference information has been shown to increase identification “success” in many cases, for example [17, 27, 28], by providing additional information completely independent of the analytical evidence. However, as this information can introduce a bias towards known compounds, this information should be incorporated with caution, depending on the experimental context.

Retention time information is often used for candidate selection in LC/MS. Unlike the retention index (RI) in GC, where the Kovats RI [29] is quite widely applied, there is not yet an established RI per se for LC/MS despite a high interest. Instead, where a reverse phase column is used for the LC method, the octanol–water partitioning coefficient ($\log P$) and retention times (RT) of substances can be correlated due to the column properties [30]. The $\log P$ of the measured standards can be predicted with various software approaches and correlated with the retention times (see e.g. [31] for an overview on different methods). This has already been used in candidate selection (e.g. [13, 32–34]), with various $\log P$ predictions. The orthogonal information proved useful despite the large errors associated with the predictions (e.g. over 1 log unit or up to several minutes retention time window depending on the LC run length). These are due to uncertainties in $\log P$ prediction that are common among different prediction implementations when considering a broad range of substances with different (and many) functional groups and ionization behaviour. As the Chemical Development Kit (CDK [35, 36]) offers $\log P$ calculations, this can be incorporated within MetFrag2.2. Alternative approaches with $\log D$, accounting for ionization, or those requiring more extensive calculations (e.g. [37–39]) can be included via a user-defined score, described further below.

This article details the developments and improvements that have been made to MetFrag since the original publication, including a detailed evaluation on several datasets and specific examples to demonstrate the use of MetFrag2.2 in small molecule identification.

Implementation

MetFrag architecture

MetFrag2.2 is written in Java and uses the CDK [35] to read, write and process chemical structures. To start, candidates are selected from a compound database based on the neutral monoisotopic precursor mass and a given relative mass deviation (e.g. 229.1089 ± 5 ppm),

the neutral molecular formula of the precursor or a set of database-dependent compound accession numbers. Currently, the online databases KEGG [9, 10], PubChem [7] or ChemSpider [8] can be used with MetFrag2.2, as well as offline databases in the form of a structure data file (SDF) or, new to MetFrag2.2, a CSV file that contains structures in the form of InChIs [40] together with their identifiers and other properties. Furthermore, MetFrag2.2 is able to query local compound database systems in MySQL or PostgreSQL, as performed in [41].

MetFrag2010 considered the ion species $[M + H]^+$, $[M]^+$, $[M]^-$ and $[M - H]^-$ during candidate retrieval and fragment generation. While the web interface contained an adduct mass adjustment feature, the presence of adducts was not considered in the fragments. MetFrag2.2 can also handle adducts also appearing in the product ions associated with $[M + Na]^+$, $[M + K]^+$, $[M + NH_4]^+$ for positive ionization and $[M + Cl]^-$, $[M + HCOO]^-$ and $[M + CH_3COO]^-$ for negative ionization. As the candidate retrieval is performed on neutral molecules, the precursor adduct type must still be known beforehand; for high-throughput workflows this information is intended to come from the workflow output.

Additive relative and absolute mass deviation values are used to perform the MS/MS peak matching and can be adjusted according to the instrument type used for MS/MS spectra acquisition. The number of fragmentation steps performed by MetFrag2.2 can be limited by setting the tree depth (default is 2).

The overall score of a given candidate is calculated as shown in Eq. 1.

$$S_{C_{\text{Final}}} = \omega_{\text{Frag}} \cdot S_{C_{\text{Frag}}} + \omega_{\text{RT}} \cdot S_{C_{\text{RT}}} + \omega_{\text{Refs}} \cdot S_{C_{\text{Refs}}} + \omega_{\text{Incl}} \cdot S_{C_{\text{Incl}}} + \omega_{\text{Excl}} \cdot S_{C_{\text{Excl}}} + \omega_{\text{Suspects}} \cdot S_{C_{\text{Suspects}}} + \dots + \omega_n \cdot S_{C_n} \quad (1)$$

The final candidate score $S_{C_{\text{Final}}}$ is the weighted sum of all single scoring terms used, where the weights given by ω_i specify the contribution of each term. All S_C scoring terms used to calculate $S_{C_{\text{Final}}}$ are normalized to the maximum value within the candidate result list for a given MS/MS input. The calculation of individual scoring terms are detailed in the subsections below; all terms besides $S_{C_{\text{Frag}}}$ are new to MetFrag2.2.

A variety of output options are available. Output SDFs contain all compounds with a structure connection table and all additional information stored in property fields. For the CSV and XLS format, the structures are encoded by SMILES [42] and InChI codes, while an extended XLS option is available that includes images of the compounds and/or fragments. In all cases the compounds are sorted by the calculated score by default.

In silico fragmentation refinements

The *in silico* fragmentation part of MetFrag2.2 has undergone extensive algorithmic and scoring refinements. The fragmentation algorithm still uses a top-down approach, starting with an entire molecular graph and removing each bond successively. However, the generated fragments are now stored more efficiently by using only the indexes of removed bonds and atoms, similar to the MAGMa approach [43]. This not only increases processing speed and decreases memory usage, but still allows the fast calculation of the masses and molecular formulas of each fragment. This makes it possible to process MS/MS spectra with higher tree depths to generate reliable fragments for molecules with complex ring structures with lower CPU and memory requirements. As a result, fragment filters such as the molecular formula duplicate filter used in MetFrag2010 to decrease the number of generated structures were no longer required, their removal reduces the risk of missing a potentially correct fragment. The calculation of the fragmentation score, $S_{C_{\text{Frag}}}$, modified from the score given in [6], is shown in Eq. 2 for a given candidate C:

$$S_{C_{\text{Frag}}} = \sum_{p \in P} \frac{\text{RelMass}_p^\alpha \cdot \text{RelInt}_p^\beta}{\left(\sum_{b \in B_f} \text{BDE}_b \right)^\gamma} \quad (2)$$

For each peak p matching a generated fragment, the relative mass RelMass_p and intensity RelInt_p as well as the sum of all cleaved bonds b of the fragment f assigned to p are considered. Where more than one fragment could be assigned to p , the fragment with the lowest denominator value is considered. In contrast to Eq. 2, the MetFrag2010 scoring used the difference between $1/\max(w_c)$ and $1/\max(e) \cdot e_c$, which could lead to negative scores if the BDE penalty was large. The weights α , β and γ were optimized on a smaller subset of spectra from Gerlich and Neumann [19] that was not used further in this work including merged MassBank IPB (PB) and RIKEN (PR) MS/MS spectra and were set to $\alpha = 1.84$, $\beta = 0.59$ and $\gamma = 0.47$. Once $S_{C_{\text{Frag}}}$ has been calculated for all candidates within a candidate list, it is normalised so that the highest score is one.

Compound filters, element and substructure options

The *unconnected compound filter* was already implemented in MetFrag2010 to remove salts and other unconnected substances that could not possibly have the correct neutral mass from the candidate list. InChIKey filtering has now been added to reduce candidate redundancy due to stereoisomerism, as stereoisomers inflate candidate numbers but cannot (usually) be distinguished with MS/MS. The InChIKey filtering is performed using the first block, which encodes the molecular skeleton (or

connectivity), but not the stereochemistry. While this is generally reasonable, some tautomers may have differing InChIKey first blocks (see e.g. [40]), such that not all tautomers will be filtered out. The highest-scoring stereoisomers overall with a matching first block are retained.

Element restrictions have been added to enhance the specificity of the exact mass search. Three options are available to restrict the elements considered: (a) include *only* the given elements, (b) the given elements have to be present, but other elements can also be present (as long as they are not explicitly excluded) and (c) exclude certain elements. Options (b) and (c) can be used in combination. These filters can be used for example to incorporate isotope information (e.g. Cl, S) that has been detected in the full scan (MS1) data.

Substructure restrictions allow the inclusion and exclusion of certain molecular substructures, encoded in SMARTS [44]. Each substructure is searched independently, thus overlapping substructures can also be considered. This option is particularly useful for cases where detailed information about a parent substance is known (e.g. transformation product, metabolite elucidation), or complementary substructure information is available from elsewhere (e.g. MS2Analyzer [45] or other MS classifiers [13]). Candidates containing certain substructures can either be included and/or excluded prior to fragmentation, or scored differently. To calculate a score, the number of matches in the inclusion or exclusion list containing n substructures are added per candidate as given in Eq. 3 (where $M_i = 1$, if substructure i matches candidate C from the given candidate list L or 0 otherwise):

$$N_{C_{\text{Match}}} = \sum M_1 + M_2 + \dots + M_n; \quad M_i \in \{0, 1\} \quad (3)$$

The inclusion ($S_{C_{\text{Incl}}}$) and/or exclusion ($S_{C_{\text{Excl}}}$) score(s) per candidate are then calculated as shown in Eq. 4:

$$\begin{aligned} S_{C_{\text{Incl}}} &= \frac{N_{C_{\text{Match}}}}{\max_{C' \in L} (N_{C'_{\text{Match}}})}; \\ S_{C_{\text{Excl}}} &= \frac{n - N_{C_{\text{Match}}}}{\max_{C' \in L} (n - N_{C'_{\text{Match}}})} \end{aligned} \quad (4)$$

where $\max_{C' \in L} (N_{C'_{\text{Match}}})$ is the maximal value of $N_{C_{\text{Match}}}$ within the candidate list and the scores $S_{C_{\text{Incl}}}$ or $S_{C_{\text{Excl}}}$ are set to 0 when $\max_{C' \in L} (N_{C'_{\text{Match}}}) = 0$ or $\max_{C' \in L} (n - N_{C'_{\text{Match}}}) = 0$, respectively.

Additional substance information

Reference and patent information

While the reference and patent information is represented by the placeholder term $\omega_{\text{Refs}} \cdot S_{C_{\text{Refs}}}$ in Eq. 1, the score can either be composed of several terms or added as a combined term, as described below.

If the query databases is PubChem, the number of patents (PubChemNumberPatents, PNP) and PubMed references (PubChemPubMedCount, PPC) are retrieved for each candidate via the PubChem PUG REST API [46]. These values result in the scoring terms $S_{C_{PNP}}$ and $S_{C_{PPC}}$ which can be weighted individually, or a combined term with either or both parameters. For the latter, first, a cumulative reference term is calculated as shown in Eq. 5, before the PubChem combined reference score ($S_{C_{PCR}}$) is calculated for candidate C in candidate list L as shown in Eq. 6 for PubChem:

$$N_{C_{PCR}} = a_1 \cdot PNP_C + a_2 \cdot PPC_C, \quad a_1, a_2 \in \{0, 1\} \quad (5)$$

$$S_{C_{PCR}} = \frac{N_{C_{PCR}}}{\max_{C' \in L} N_{C'_{PCR}}} \quad (6)$$

For ChemSpider, five values with reference information can be retrieved using the ChemSpider web services [47]), including the number of data sources (ChemSpiderDataSourceCount, CDC), references (ChemSpiderReferenceCount, CRC), PubMed references (ChemSpiderPubMedCount, CPC), Royal Society for Chemistry (RSC) references (ChemSpiderRSCCount, CRSC) and external references (ChemSpiderExternalReferenceCount, CERC). Any combination of these reference sources can be used and weighted individually, yielding the score terms $S_{C_{CDC}}$, $S_{C_{CRC}}$, $S_{C_{CPC}}$, $S_{C_{CRSC}}$ and $S_{C_{CERC}}$. Alternatively, the ChemSpider Combined Reference Scoring term ($S_{C_{CCR}}$) can be calculated, as shown below in Eqs. 7 and 8:

$$N_{C_{CCR}} = b_1 \cdot CRC_C + b_2 \cdot CERC_C + b_3 \cdot CRSC_C + b_4 \cdot CPC_C + b_5 \cdot CDC_C \quad (7)$$

$$b_1, b_2, b_3, b_4, b_5 \in \{0, 1\}$$

$$S_{C_{CCR}} = \frac{N_{C_{CCR}}}{\max_{C' \in L} N_{C'_{CCR}}} \quad (8)$$

The corresponding command line terms are given in the additional information (see Additional files 1, 2, 3).

Suspect lists

Additional lists of substances (so-called “suspect lists”) can be used to screen for the presence of retrieved candidates in alternative databases. The suspect lists are input as a text file containing InChIKeys (one key per line) for fast screening. The first block of the InChIKey is used to determine matches. Example files are available from [25]. This “suspect screening” can be used as an inclusion filter (include only those substances that are in the suspect list) or as an additional scoring term for the ranking of the candidates, yielding the term $\omega_{\text{Suspects}} \cdot S_{C_{\text{Suspects}}}$ given in Eq. 1.

Retention time score via log P

The retention time (RT) scores offered within MetFrag2.2 are based on the correlation of $\log P$ and user-provided RT information. The RTs must be associated with sufficient analytical standards measured under the same conditions as the unknown spectrum (a minimum of ten data points are recommended, depending on the distribution over the chromatographic run). By default, the $\log P$ is calculated using the XlogP algorithm in the CDK library [36, 48, 49]. Alternatively, if PubChem is used as a candidate source, the XLOGP3 value retrieved from PubChem can also be used [50]. The user-provided RTs and their associated $\log P$ values comprise a training dataset to generate a linear model between RT and the $\log P$, shown in Eq. 9, where a and b are determined using least squares regression:

$$\log P_{\text{Unknown}} = a \cdot RT_{\text{Unknown}} + b \quad (9)$$

This equation is then used to estimate $\log P_{\text{Unknown}}$, given the measured RT associated with the unknown spectrum, and compared with $\log P_C$ calculated for each candidate. It is imperative that the $\log P$ calculated for each candidate arises from the same source as the $\log P$ used to build the model in Eq. 9. Lower $\log P$ deviations result in a higher score for a candidate; the score is calculated using density functions assuming a normal distribution with $\sigma = 1.5$ (chosen arbitrarily), as shown in Eq. 10:

$$S_{C_{RT}} = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\log P_{\text{Unknown}} - \log P_C)^2}{2\sigma^2}} \quad (10)$$

Alternative $\log P$ values that are not available within MetFrag2.2 can also be used to establish a model and calculate a different $S_{C_{RT}}$ in a two-step approach. First, MetFrag2.2 can be run either with or without one of the built-in models, so that candidates and all other scores can be obtained. The InChIs or SMILES in the output CSV, or structures in the output SDF can then be used by the user to calculate their own $\log P$ values. These should be included in the output CSV or SDF using the “User-LogP” tag (or a self-defined alternative) and used as input for MetFrag2.2 with the Local Database option and a RT training file containing retention times and the user $\log P$ s with the column header matching the tag in the results file. The values a and b in Eq. 9 are then determined and used to calculate $S_{C_{RT}}$ for the final scoring. Alternative RT models that do not use $\log P$ should be included as a “user-defined score”, as described below.

User-defined scoring functions

The final term in Eq. 1, $\omega_n \cdot S_{C_n}$, represents the “user-defined scoring function”, which allows users to incorporate any additional information into the final candidate scoring. The MetFrag2.2 output (InChIs, SMILES, SDF)

can be used to calculate additional “scores” for the candidates using external methods and these scores can be reimported with the candidates and all other MetFrag2.2 scores in the pipe-separated (`|`) format for final scoring. The scores and weights are matched from the column headers in the input file and the parameter names added to the score list. The commands are given in an additional table (see Additional files 1, 2, 3), with an example (“terbutylazine and isomers”) below.

Results and discussion

The changes to MetFrag2.2 were evaluated on several datasets, described in the following. Further examples are given to demonstrate the use of different new features. Unless mentioned otherwise, candidate structures were retrieved from the compound databases PubChem and ChemSpider in June, 2015. If not stated explicitly, the datasets were processed with a relative and absolute fragment mass deviation of 5 ppm and 0.001 Da, respectively. The resulting ranks, if not specified explicitly, correspond to pessimistic ranks, where the worst rank is reported in the case where the correct candidate has the same score as other candidates. Stereoisomers were filtered to keep only the best scored candidate based on the comparison of the first part of the candidates’ InChIKeys. The expected top ranks calculated as in Allen et al. [16], which handles ties of equally scored candidates in a uniformly random manner, are also given when comparing the two *in silico* fragmenters. This demonstrates the effect of equally scored candidates on ranking results.

The datasets from Eawag and UFZ used in this publication arose from the measurement of reference standard collections at Eawag and UFZ, which comprise small molecules of environmental relevance such as pharmaceuticals and pesticides with a wide range of physico-chemical properties and functional groups, and also include several transformation products which typically have lower reference counts. All spectra are publicly available in MassBank.

In Silico fragmentation performance

Comparison with MetFrag2010

The merged spectra from 102 compounds published in Hill et al. [11], also used in [6, 19], formed the first evaluation set. The candidate sets from Gerlich and Neumann [19] were used as input for MetFrag2.2 and processed with consistent settings: relative mass deviation of 10 ppm and absolute mass deviation of 0 Da, i.e. no absolute error, for a direct comparison with MetFrag2010. With MetFrag2.2, the median rank improved from 18.5 to 14.5, while the number of correct ranked candidates in the top 1, 3 and 5 improved from 8 to 9, 20 to 24 and 28 to 34, respectively.

Baseline performance on Orbitrap XL Dataset

A set of 473 LTQ Orbitrap XL spectra resulting from 359 reference standards formed the second dataset. The spectra were measured at several collision energies with both collision-induced ionization (CID) 35 and higher-energy CID (HCD) 15, 30, 45, 60, 75 and 90 normalized units (see [51] for more details) coupled with liquid chromatography (LC) with a 25 min program on an Xbridge C18 column. The raw files were processed with RMassBank [51, 52], yielding the “EA” records in MassBank. These spectra were merged using the `mzClust_hclust` function in `xcms` [53] (parameters `eppm = 5 × 10-6` and `eabs = 0.001 Da`) to create peaks with the mean *m/z* value and highest (relative) intensity and retained where they contained at least one fragment peak other than the precursor. In total 473 spectra (319 $[M + H]^+$ and 154 $[M - H]^-$) were evaluated with MetFrag2010 using ChemSpider, as well as MetFrag2.2 using either PubChem or ChemSpider. The correct molecular formula was used to retrieve candidates. The results, given in Table 1, show the clear improvement between MetFrag2010 (73 Top 1 ranks with ChemSpider) and MetFrag2.2 (105 top 1 ranks with ChemSpider). This is also indicated by the higher relative ranking positions (RRP) [19] retrieved by MetFrag2.2 where a value of 1 marks the best possible result and 0 the worst possible result. Note that the version used here is 1-RRP as defined in Kerber et al. [54] and Schymanski et al. [55]. The results show that the algorithmic refinements improved the baseline *in silico* fragmentation performance, although it is difficult to tell which of the changes had the greatest influence.

Comparison with CFM-ID using Orbitrap XL Dataset

The same dataset of 473 merged spectra and the corresponding PubChem candidate sets were used as input for CFM-ID [16] version 2.0 (“Jaccard”, RDKit 2015.03.1, `lpsolve 5.5.2.0`, `Boost 1.55.0`), to form a baseline comparison with an alternative *in silico* fragmenter. The results, given in Table 1, show that CFM-ID generally performed better, indicated by the higher number of correct first ranked candidates (43 vs. 30), top 5 (170 vs. 145), top 10 (232 vs. 226) and a lower median and mean rank of 11 versus 12 and 127 versus 141. The expected ranks, including equal ranked candidates, also implied a better performance of CFM-ID (top 1: 43 vs. 57, top 5: 163 vs. 193, top 10: 245 vs. 261). This was not entirely unexpected as CFM-ID uses a more sophisticated fragmentation approach, but also requires a much longer computation time. For run time analysis, 84 of the 473 queries, selected at random, were processed (single-threaded) with MetFrag2.2 and CFM-ID in parallel on a computer cluster with a maximum of 28 (virtual) computer nodes with 12 CPU cores each. The total run times (system +

Table 1 Comparison of *in silico* fragmentation results for 473 Ewag Orbitrap spectra (formula search)

	MetFrag2010	MetFrag2.2		CFM-ID	MetFrag2.2 + CFM-ID
	ChemSpider	ChemSpider	PubChem	PubChem	PubChem
Pessimistic ranks					
Median rank	8	4	12	11	8
Mean rank	74	38	141	127	85
Mean RRP	0.859	0.894	0.880	0.881	0.901
Top 1 ranks	73 (15 %)	105 (22 %)	30 (6 %)	43 (9 %)	62 (13 %)
Top 5 ranks	202	267	145	170	202
Top 10 ranks	258	320	226	232	276
Expected top ranks					
Top 1 ranks	90 (19 %)	124 (26 %)	43 (9 %)	57 (12 %)	70 (15 %)
Top 5 ranks	218	280	163	193	213
Top 10 ranks	274	329	245	261	288

MetFrag2010 and MetFrag2.2 were compared with the same ChemSpider candidate sets; MetFrag2.2 and CFM-ID with the same PubChem candidate sets. Far right: Best top 1 pessimistic ranks obtained by combining MetFrag2.2 and CFM-ID 2.0 with the weights $\omega_{\text{Frag}} = 0.67$ and $\omega_{\text{CFM-ID}} = 0.33$. The expected ranks, which partially account for equally scored candidates as calculated in [16], are shown in the lower part of the table

user runtime, retrieved by linux bash command *time*) were 75 min for MetFrag2.2 and 12,570 min (209.5 h) for CFM-ID. These values represent the runtime on a single CPU core for all 84 queries in series. The average run time per query amounts to 54 s for MetFrag2.2 and 8979 s (150 min) for CFM-ID.

As CFM-ID and MetFrag2.2 use independent *in silico* fragmentation approaches, one can hypothesize that the combination of the approaches should improve the results further. To demonstrate this, the CFM-ID results were incorporated into MetFrag2.2 by introducing an additional scoring term $\omega_{\text{CFM-ID}} \cdot S_{\text{CFM-ID}}$, where $S_{\text{CFM-ID}}$ defines the normalized CFM-ID probability of candidate C . Different contributions of each fragmenter relative to another was determined by randomly drawing 100 combinations of ω_{Frag} and $\omega_{\text{CFM-ID}}$ such that $(\omega_{\text{Frag}} + \omega_{\text{CFM-ID}} = 1)$. The best results, shown in Table 1, were obtained with $\omega_{\text{Frag}} = 0.67$ and $\omega_{\text{CFM-ID}} = 0.33$, where the change in number 1 ranks with weight is shown in Additional file 4. With this best combination, the number of Top 1 ranks improved from 30 to 61, while the median rank improved to 8. This shows that the combination of independent fragmentation methods can indeed yield valuable improvements to the results, shown again in the next paragraph after including the additional information. Further validation was beyond the scope of the current article, as further improvements could be made by retraining CFM-ID on Orbitrap data, but would be of interest in the future.

Adding retention time and reference information

Parameter selection on Orbitrap XL Dataset

The next stage was to assess the effect of references and retention time information on the MetFrag results.

Firstly, each score term (i.e. fragmenter, retention time and/or reference information) was either included or excluded by setting the weight ($\omega_{\text{Frag}}, \omega_{\text{RT}}, \omega_{\text{Refs}}$) to 1 or 0, to assess the impact of the various combinations on the number of correctly-ranked number 1 substances. The results are shown in Table 2. The best result was obtained when all three “score terms” (fragmenter, RT and references) were included in candidate ranking. For PubChem, both RT/log P models (CDK XlogP and XLOGP3 from PubChem directly) were assessed and thus two sets of results are reported. The reference information was included using the combined reference scores introduced in Eqs. 6 and 8, where all combinations of the reference values described above (1–2 for PubChem, 1–5 for ChemSpider, i.e. 3 and 31 combinations in total, respectively), were used to form a cumulative total reference term, shown in Eq. 5 for PubChem and Eq. 7 for ChemSpider. The best results were achieved with PubChem when using both patents and PubMed references ($S_{\text{C}_{\text{PNP+PPC}}}$; $a_1 = 1, a_2 = 1$), while for ChemSpider using the ReferenceCount, ExternalReferenceCount and the DataSourceCount ($S_{\text{C}_{\text{CRC+CERC+CDC}}}$) proved best, i.e. $b_1 = 1, b_2 = 1, b_3 = 0, b_4 = 0, b_5 = 1$. Table 2 contains the number of Top 1 ranks for each combination of $\omega_{\text{Frag}}, \omega_{\text{RT}}, \omega_{\text{Refs}} \in \{0, 1\}$. The results show clearly that, while references alone result in over 311 top 1 ranks (65 % for PubChem), the addition of both fragmentation and retention time information improves the results further, to 69 % of candidates ranked first (PubChem) and even 87 % of candidates ranked first (ChemSpider). For PubChem the distribution of the number of CombinedReferences (including patents and PubMed references) for the 359 queries of the (unique) correct candidates is shown in Additional file 5.

Table 2 PubChem and ChemSpider results (number of pessimistic top 1 ranks) for 473 Eawag Orbitrap spectra

Weight term	Score term	Weights						
ω_{Frag}	$S_{\text{C-Frag}}$	1	1	1	0	1	0	0
ω_{RT}	$S_{\text{C-RT}}$	1	1	0	1	0	1	0
ω_{Refs}	$S_{\text{C-Refs}}$	1	0	1	1	0	0	1
Database	RT source	Top 1 ranks						
PubChem	XLOGP3	325 (69 %)	53	322	315	30	10	311
PubChem	CDK XlogP	326 (69 %)	43	322	316	30	8	311
ChemSpider	CDK XlogP	411 (87 %)	113	411	376	105	41	376

The weights indicate where the score term was included (1) or excluded (0) from the candidate ranking. For PubChem $\omega_{\text{Refs}} \cdot S_{\text{C-Refs}} = \omega_{\text{Refs}} \cdot (S_{\text{C-PNP+PPC}})$; for ChemSpider $S_{\text{C-Refs}} = S_{\text{C-CRC+CERC+CDC}}$ only. See text for explanations

Following this, the combination of each scoring term was assessed by randomly drawing 1000 different weight combinations such that ($\omega_{\text{Frag}} + \omega_{\text{RT}} + \omega_{\text{Refs}} = 1$) to determine the optimal relative contributions of each term for the best results. This was performed for all combinations of reference sources (3 for PubChem, 31 for ChemSpider). The best result was obtained again when using both patents and PubMed references for PubChem ($S_{\text{C-PNP+PPC}}$; $a_1 = 1$, $a_2 = 1$), but using only the ReferenceCount ($S_{\text{C-CRC}}$; $b_1 = 1$, $b_2 = 0$, $b_3 = 0$, $b_4 = 0$, $b_5 = 0$) for ChemSpider. The results are summarized in Table 3 (including the weight terms) and shown in Figs. 1 and 2 for PubChem and ChemSpider respectively. These triangle plots show the top 1 candidates for all ω_i combinations, colour-coded (black—0 % of the correct candidates ranked first, yellow—10 0 % of the correct candidates ranked first) with the ω_i per category increasing in the direction of the arrow. Each corner is $\omega_i = 1$. The 25th and 75th percentiles are shown to give an idea of the distribution of the ranks. The equivalent plots for the number of top 5 and top 10 ranks are given in Additional files 6, 7, 8 and 9. Although the results from (ω_{Frag} , ω_{RT} , $\omega_{\text{Refs}} \in \{0, 1\}$) above indicated that the term $S_{\text{C-CRC+CERC+CDC}}$ yielded the best result for ChemSpider with 411 top 1 ranks, $S_{\text{C-CRC}}$ yielded 410 top 1 ranks for the same calculations, indicating that there is little difference between the two combinations. Using the randomly-drawn weights, the top 1 ranks improved to 420 (ChemSpider) and 336 (PubChem). This proves without a doubt that the addition of reference and retention time information drastically improves the performance, going from 22 to 89 % top 1 ranks (ChemSpider) and 6.3 to 71 % (PubChem).

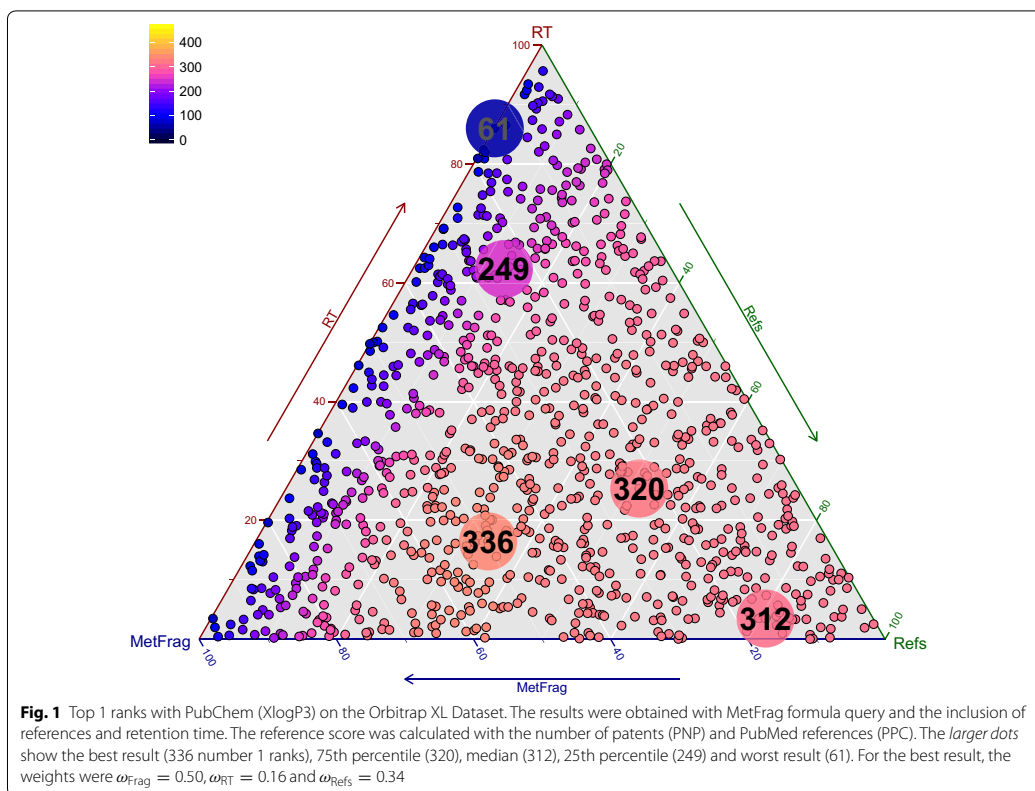
As above, it was interesting to investigate whether the addition of a complementary fragmentation technique, i.e. CFM-ID, would improve the results even further. MetFrag2.2 and CFM-ID were combined with retention time and reference information using 1000 randomly

Table 3 PubChem and ChemSpider results for 473 Eawag orbitrap spectra with formula retrieval, including *in silico* fragmentation, RT and reference information as shown, with the given ω_i for the highest number of Top 1 ranks

	MetFrag2.2			MetFrag2.2 + CFM-ID
Database	ChemSpider	PubChem	PubChem	PubChem
RT/log P Model	CDK XlogP	CDK XlogP	XLOGP3	CDK XlogP
$\omega_{\text{Frag}} (S_{\text{C-Frag}})$	0.49	0.57	0.50	0.33
$\omega_{\text{RT}} (S_{\text{C-RT}})$	0.19	0.02	0.16	0.03
$\omega_{\text{Refs}} (S_{\text{C-Refs}})$	0.32	0.41	0.34	0.35
$\omega_{\text{CFMID}} (S_{\text{C-CFMID}})$	–	–	–	0.29
Median rank	1	1	1	1
Mean rank	6.5	35	41	18
Mean RRP	0.990	0.977	0.977	0.978
Top 1 ranks	420 (89 %)	336 (71 %)	336 (71 %)	343 (73 %)
Top 5 ranks	447	396	398	411
Top 10 ranks	454	422	414	429

For PubChem $\omega_{\text{Refs}} \cdot S_{\text{C-Refs}} = \omega_{\text{Refs}} \cdot (S_{\text{C-PNP+PPC}})$; for ChemSpider $S_{\text{C-Refs}} = S_{\text{C-CRC}}$ only. See text for explanations. Far right: combining CFM-ID results to incorporate complementary fragmentation information

drawn combinations of ω_{Frag} , $\omega_{\text{CFM-ID}}$, ω_{RT} and $\omega_{\text{PNP+PPC}}$ such that ($\omega_{\text{Frag}} + \omega_{\text{CFM-ID}} + \omega_{\text{RT}} + \omega_{\text{PNP+PPC}} = 1$). The results, shown in Table 3, indicate that the PubChem results can be improved further, to 343 top 1 ranks (73 %). This is a drastic improvement from the performance of both original fragmenters alone, with CFM-ID alone yielding between 10 and 12 % top 1 hits (expected rank) in their original publication [16] with an older PubChem, the combination of both fragmenters alone yielding 15 % (expected rank) here. These combined results are also drastically better than the latest *in silico* fragmentation results just published for CSI:FingerID. Dührkop et al. [18] investigated each individual fragmenter currently available and compared the results with



the CSI:FingerID. Despite using different data and settings to those here, their results on the Agilent dataset indicated that MetFrag2010 and CFM-ID achieved 9 and 12 % top 1 (expected) ranks, which are reasonably comparable with the results presented above. FingerID [15] achieved 19.6 %, while CSI:FingerID achieved 39 % top 1 results, which is a dramatic improvement over the other fragmenters. Since the external information boosted the top 1 ranks to 73 % for MetFrag2.2 plus CFM-ID, one could speculate that the combination of CSI:FingerID, MetFrag2.2 and CFM-ID would result in an even greater performance.

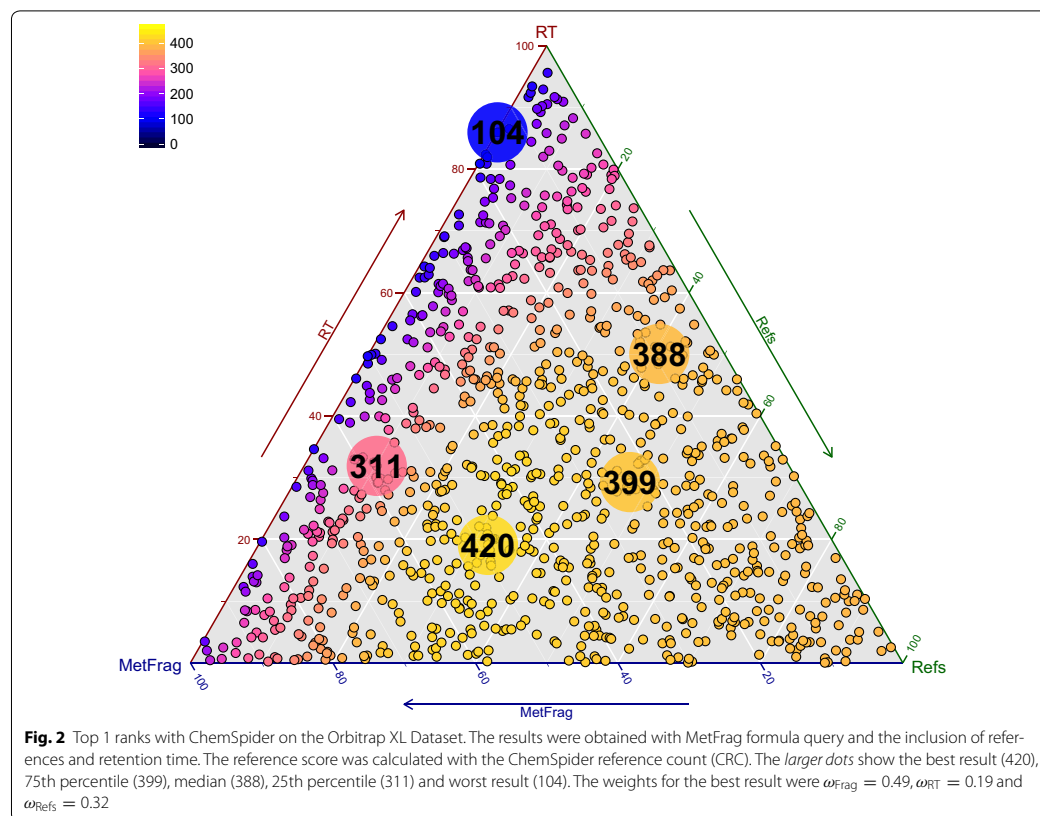
Cross-evaluation on additional datasets

As the RT and reference scores are very subjective to experimental context, MetFrag2.2 now contains so many tuneable parameters that it will be beneficial to users when a few default cases are suggested. Thus, once the optimal reference source combinations were determined as described above, alternative datasets were used to re-determine the optimal weights ω_{Frag} , ω_{RT} and ω_{Refs} to

investigate the variation over different datasets. Three sufficiently large datasets available on MassBank contained good quality MS/MS and RT data, all processed with RMassBank [51].

UF dataset: A subset of the 2758 UFZ Orbitrap XL records were acquired on an Kinetex Core-Shell C18 column from Phenomenex with a 40 min chromatographic program (all others were direct infusion experiments). These MS/MS spectra, arising from $[M + H]^+$ and $[M - H]^-$ precursors, were recorded at four collision energies: CID 35 and 55 as well as HCD 50 and 80. These spectra were merged and processed as described above for the Orbitrap XL dataset, resulting in 225 merged spectra ("UF" dataset) from 195 substances (184 $[M + H]^+$ and 41 $[M - H]^-$).

EQex and EQxPlus datasets: Two additional Eawag datasets were also available. The "EQex" dataset, measured on a Q Exactive Orbitrap, contained MS/MS spectra associated with $[M + H]^+$ and $[M - H]^-$ precursors recorded at six different collision energies (HCD 15, 30, 45, 60, 75 and 90). The "EQxPlus" dataset, measured



on a Q Exactive Plus Orbitrap, contained MS/MS spectra associated with $[M + H]^+$ and $[M - H]^-$ precursors recorded at nine different collision energies (HCD 15, 30, 45, 60, 75, 90, 120, 150, 180).

Both datasets were acquired using the same LC setup as the other Eawag dataset. The MS/MS from these two datasets were merged as above to yield 294 merged spectra from 204 compounds (195 $[M + H]^+$ and 94 $[M - H]^-$) for the “EQEx” dataset and 314 merged spectra from 232 compounds (219 $[M + H]^+$ and 91 $[M - H]^-$) for the “EQExPlus” dataset. There was a very small overlap between the different Eawag datasets (5, 2 and 2 substance overlap between EA and EQEx, EA and EQExPlus and EQEx and EQExPlus, respectively).

The overlap between the UFZ and Eawag datasets was larger, with 97, 16 and 21 substances in common between the UFZ and EA, EQEx and EQExPlus datasets, respectively. The overlap was determined using the first block of the InChIKey. As the spectral and retention time data for the substances in the individual datasets were processed

independently with different collision energies and ionization modes, none of the overlapping substances were removed from the datasets. The retention times extracted from the MassBank records per substance were used to establish the RT–log P model (see Eq. 9) for each dataset independently based on a tenfold cross-validation.

The influence of the different parameters was assessed for each dataset by setting ω_{Frag} , ω_{RT} and ω_{Refs} to either 0 or 1 again; these results are presented in Table 4. As above, the performance improved from between 2 and 9 % of the candidates ranked first using fragmentation alone, through to 64–82 % ranked first when all ω_x were weighted equally, although the results varied quite dramatically between the datasets. The 473 spectrum dataset used above thus fell within this range.

Similarly, the optimization of ω_{Frag} , ω_{RT} and ω_{Refs} was performed again for each dataset independently using the 1000 randomly-drawn weights. The results are presented in Table 5 and show that the percentage of top 1 ranks varies widely between the datasets, from 63 to 82 %; the

Table 4 Results (Top 1, 5 and 10 ranks) using PubChem formula queries on three additional datasets

Weight term	Score Term	Weights						
ω_{Frag}	S_{CFrag}	1	1	1	0	1	0	0
ω_{RTs}	S_{CRT}	1	1	0	1	0	1	0
ω_{Refs}	S_{CRefs}	1	0	1	1	0	0	1
Dataset	Metric	Ranks						
UF (n = 225)	Top 1 ranks	164 (73 %)	9	163	159	3	2	157
UF (n = 225)	Top 5 ranks	186 (83 %)	48	189	189	36	13	199
UF (n = 225)	Top 10 ranks	191 (53 %)	77	196	192	61	25	204
EQex (n = 289)	Top 1 ranks	235 (81 %)	33	232	230	26	11	223
EQex (n = 289)	Top 5 ranks	263 (91 %)	87	260	258	88	38	276
EQex (n = 289)	Top 10 ranks	270 (93 %)	132	269	263	139	55	280
EQexPlus (n = 310)	Top 1 ranks	190 (61 %)	32	183	182	21	8	181
EQexPlus (n = 310)	Top 5 ranks	238 (77 %)	84	246	238	83	28	243
EQexPlus (n = 310)	Top 10 ranks	254 (82 %)	115	258	247	121	37	256

The weights indicate where ranking parameters were included (1) or excluded (0) from the candidate ranking. Retention time score calculation was performed using the XLOGP3 values of PubChem. $\omega_{\text{Refs}} \cdot S_{\text{CRefs}} = \omega_{\text{Refs}} \cdot S_{\text{CRRP+PPC}}$. See text for explanations

original dataset falls in the middle with 71 %. The results in Table 5 also show that the suggested relative weights to one another remain consistent enough to enable default parameter suggestion, with $\omega_{\text{Frag}} \approx 0.5$, $\omega_{\text{RT}} \approx 0.2$ and $\omega_{\text{Refs}} \approx 0.3$. All results for the number of top 1 ranks for the three additional datasets are shown in Additional files 10, 11 and 12.

Specific examples

As the additional features are more difficult to evaluate using large datasets, individual examples are presented below to demonstrate the flexibility of MetFrag2.2 command line (CL), with the corresponding commands give in a different font. Lists of the available parameters are given in Additional files 1, 2 and 3. These examples serve to show how MetFrag2.2 can also be adjusted by the user to explore individual cases in greater detail than during e.g. a high-throughput screening.

Gathering evidence for unknown 199.0428

During the NORMAN Collaborative Non-target Screening Trial [1], a tentatively identified non-target substance of m/z $[M - H]^-$ 199.0431 was reported by one participant as mesitylenesulfonic acid (ChemSpider ID (CSID) 69438, formula $C_9H_{12}O_3S$, neutral monoisotopic mass 200.0507) or isomer. The same unknown was detected in the same sample measured at a second institute, where the standard of mesitylenesulfonic acid was available. Although the retention time was plausible (5.96 min), comparing the MS/MS spectra clearly disproved the proposed identification, with many fragments from the

Table 5 Best Top 1 rank results on three additional datasets using PubChem formula queries including *in silico* fragmentation, RT and reference information as shown, with the given ω_i

Dataset	MetFrag2.2		
	UFZ (n = 225)	EQex (n = 289)	EQexPlus (n = 310)
$\omega_{\text{Frag}} (S_{\text{CFrag}})$	0.40	0.38	0.61
$\omega_{\text{RT}} (S_{\text{CRT}})$	0.23	0.27	0.11
$\omega_{\text{Refs}} (S_{\text{CRefs}})$	0.37	0.35	0.28
Median rank	1	1	1
Mean rank	58.0	14.6	46.2
Mean RRP	0.972	0.981	0.976
Top 1 ranks	165 (73 %)	236 (82 %)	196 (63 %)
Top 5 ranks	188	261	233
Top 10 ranks	191	268	247

Retention time score calculation was performed using the XLOGP3 values of PubChem. $\omega_{\text{Refs}} \cdot S_{\text{CRefs}} = \omega_{\text{Refs}} \cdot S_{\text{CRRP+PPC}}$. See text for explanations

unknown absent in the standard spectrum. Thus, MetFrag2.2 was used to investigate other possibilities.

Firstly, the following parameter combination was used, taking the unknown MS/MS peak list from the second participant: ChemSpider exact mass search, fragment error 0.001 Da + 5 ppm, tree depth 2, unconnected compound and InChIKey filter, filter included elements = C, S (as isotope signals were detected in the full scan), experimental RT = 6.20 min, an RT training set of 355 InChIs and RTs measured on the same system and score weights of 1 (fragmenter and RT score)

and 0.25 each for four ChemSpider reference sources. This yielded 134 candidates with four different formulas ($C_9H_{12}O_3S$, $C_8H_{16}SSi_2$, $C_7H_{13}BO_2SSi$, $C_7H_{10}N_3O_2S$), all fulfilling the element filter (C, S). $S_{C_{Final}}$ ranged from 0.70 to 2.12, where several candidates had high numbers of references and similar number of peaks explained. Three candidates are shown in Table 6, along with a summary of the information retrieved. The clear top match, ethyl *p*-toluenesulfonate (CSID 6386, shown to the left) was unlikely to be correct, as the MS/MS contained no evidence of an ethyl loss and also had a clear fragment peak at m/z 79.9556, corresponding with an SO_3H group (thus eliminating alkyl sulfonates from consideration).

MetFrag2.2 was run again with the SMARTS substructure inclusion filter, which resulted in 31 candidates but with the same top matching structure. However, adding the SMARTS $S(=O)(=O)OC$ to the exclusion list eliminates the alkyl sulfonate species and resulted in 18 candidates, where the top candidate was now the originally proposed (and rejected) identification mesitylenesulfonic acid, shown in the middle of Table 6. The next matches were substitution isomers. Referring to the MS/MS again, another large peak was present at m/z 183.0115, which is often observed in surfactant spectra corresponding with a *p*-ethyl benzenesulfonic acid moiety. Running MetFrag2.2 again with a substructure inclusion of $CCc1ccc(cc1)S(=O)(=O)O$ yielded only two candidates, 4-isopropylbenzenesulfonic acid ($S_{C_{Final}} = 2.5$, CSID 6388), shown to the right in Table 6 and 4-propylbenzenesulfonic acid ($S_{C_{Final}} = 2.0$, CSID 5506213).

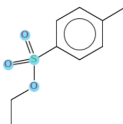
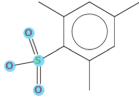
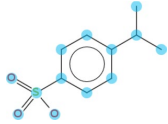
To check the relevance of the proposed candidates in an environmental sample, a “suspect screening” was performed. The STOFF-IDENT database [24] contains over

8000 substances including those in high volume production and use in Europe registered under the European REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals) Legislation. The STOFF-IDENT contents were downloaded (February 2015) and the SMILES were converted to InChIKeys using OpenBabel and given as input to MetFrag as a suspect list. Of the 134 original candidates, only one, 4-isopropylbenzenesulfonic acid, was tagged as being present in the STOFF-IDENT database. This gives additional evidence that indeed 4-isopropylbenzenesulfonic acid is the substance behind the unknown spectrum, however it has not been possible to confirm this identification at this stage due to the lack of a sufficiently pure reference standard.

Terbutylazine and isobars

The example of terbutylazine (CSID 20848, see Table 7) shows how MetFrag2.2 can help in gathering the evidence supporting the identification of isobaric substances. Terbutylazine and secbutylazine (CSID 22172) often co-elute in generic non-target chromatographic methods and have very similar fragmentation patterns, but can usually be distinguished from the other common triazine isobars propazine (CSID 4768) and triethazine (CSID 15157) via MS/MS information. However, during the NORMAN non-target screening collaborative trial [1], all four substances were reported as potential matches for the same mass, showing clearly the danger of suspect screening based only on exact mass. For this example, the merged $[M + H]^+$ MS/MS spectrum of terbutylazine from the EA dataset above (EA02840X) was used as a peak list to run MetFrag2.2, as the correct answer is clear with a reference

Table 6 Top MetFrag2.2 candidates for unknown at m/z 199.0428 with different settings

CSID	6386	69438	6388
			
Original results (134 candidates)			
Rank (n = 134)	1	6	90
#Peaks explained	5	5	5
CDK log P/ $S_{C_{Ref}}$	1.44/0.167	1.50/0.161	2.02/0.107
$\sum S_{C_{Refs}}$	94 + 15 + 7 + 70 = 186	179 + 1 + 0 + 40 = 220	32 + 0 + 0 + 21 = 53
Substructure interpretation			
Included	$S(=O)(=O)O$	$S(=O)(=O)O$	$CCc1ccc(cc1)S(=O)(=O)O$
Excluded	–	$S(=O)(=O)OC$	–
Comment	No ethyl loss in MS/MS	Disproven via standard	Present in suspect list

Structures overlaid with the included substructure were generated with AMBIT [57]. See text for details

spectrum. Table 7 shows the data for the four substances mentioned above plus the top match based on fragmentation data alone, *N*-butyl-6-chloro-*N'*-ethyl-1,3,5-triazine-2,4-diamine (CSID 4954587, given the synonym “*n*Butylazine” hereafter to save space). ChemSpider was used to perform an exact mass search, resulting in a total of 112 structures (data from only five are shown). Five scores were used, all with weight 1: FragmenterScore, ChemSpiderReferenceCount, RetentionTimeScore, SuspectListsScore and SmartSubstructureInclusionScore. To show the inclusion of external log *P* calculations, ChemAxon JChem for Excel [56] was used to predict log *P* and log *D* at pH 6.8 (the pH of the chromatographic program used) for a training dataset of the 810 substances in the Ewag database on MassBank. The log *P* and log *D* predictions were then performed externally for all MetFrag candidates on the dominant tautomeric species and added to the MetFrag CSV file for final scoring. The scores, shown in Table 7, showed that different candidates were the best match for different categories, indicated in italics. The candidates are ordered by the number of references. As above, STOFF-IDENT was used as a suspect list and all four of the substances reported by trial

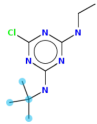
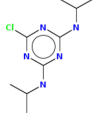
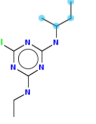

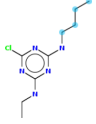
participants were indeed in STOFF-IDENT. However, Table 7 clearly shows that two can be eliminated using $S_{C_{\text{Frag}}}$ and substructure matches (as the MS/MS clearly displays the loss of a C₂H₅ and C₄H₉ group, indicating these are likely attached to a heteroatom, in this case N). Although secbutylazine is scored lower than terbutylazine, the reference count is the main influence here and both substances could be present in an environmental sample—depending on the context.

The large dataset evaluations show that MetFrag2.2 is suitable for high-throughput workflows, with a relatively quick runtime. On the other hand, the detailed examples shows how the various features of MetFrag2.2 can be used to investigate the top candidates in more detail and enhance the interpretation of the results, including the inclusion of external RT/log *P* and/or log *D* information that cannot be calculated within MetFrag2.2 (e.g. due to license restrictions, as in the case of ChemAxon).

Conclusions

In many cases additional information is available and needed from the experimental context to complement small molecule identification, especially where the mass spectrum alone is not sufficient for candidate

Table 7 Summary of MetFrag2.2 results for terbutylazine and four isobars

Name	Terbutylazine	Propazine	Secbutylazine	Triethazine	<i>n</i> Butylazine ^a
CSID	20848	4768	22172	15157	4954587
					
$S_{C_{\text{Frag}}}$	0.958	0.765	0.997	0.653	1.0
#Peaks explained	11/15	10/15	12/15	8/15	12/15
$S_{C_{\text{SRefs}}}$	286	204	56	45	4
ChemAxon log <i>P</i>	1.65	2.75	2.28	1.11	2.31
$S_{C_{\text{RT}}}$ log <i>P</i>	0.159	0.256	0.223	0.103	0.225
ChemAxon log <i>D</i>	1.63	2.75	2.19	0.97	2.23
$S_{C_{\text{RT}}}$ log <i>D</i>	0.249	0.247	0.266	0.192	0.266
Suspect hit	1	1	1	1	0
Substructure hits	2	0	2	1	2
Matches	NC(C)(C)C N[CH ₂][CH ₃]	–	NC(C)CC N[CH ₂][CH ₃]	N[CH ₂][CH ₃]	NCCCC N[CH ₂][CH ₃]
$S_{C_{\text{Final}}}$ (log <i>P</i>)	4.22	3.43	3.69	2.53	2.52
$S_{C_{\text{Final}}}$ (log <i>D</i>)	4.56	3.41	3.85	2.87	2.68
Comment	Correct substance	No longer in use	Can co-elute with 20848		

The predicted log *P* and log *D* from the retention time was 3.17 and 2.18 using a training set of 810 substances calculated externally with ChemAxon and added to MetFrag2.2 via the UserLogP option. Included substructure SMARTS were N[CH₂][CH₃], NCCCC, NC(C)CC, NC(C)(C)C

^aName synonym assigned for space reasons. The values in italics indicates the best result per category. Structures overlaid with the included substructure were generated with AMBIT [57]. See text for details and weights

selection from a large number of candidates. The results for MetFrag2.2 clearly show the benefit of considering this additional information, with a tenfold improvement compared with MetFrag2.2 fragmentation information alone. The flexibility of the new features in addition to the ability to add user-defined scores means that MetFrag2.2 is ideally suited to high-throughput workflows, but can also be used to perform individual elucidation efforts in greater detail. The ability to incorporate CFM-ID as an additional scoring function shows the potential to improve these results further using complementary *in silico* fragmentation approaches. The parameter files including the spectral data, the candidate, result and ranking files of the used EA, UF, EQEx, EQExPlus and HILL datasets are available at <http://msbi.ipb-halle.de/download/CHIN-D-15-00088/> and can be downloaded as ZIP archives. Feedback on the command line version available at <http://c-ruttkies.github.io/MetFrag/> is welcome. The new functions greatly reduce the burden on users to collect and merge ever increasing amounts of information available for substances present in different compound databases, thus enabling them to consider much more evidence during their screening efforts.

Availability and requirements

- Project name: MetFrag2.2;
- Project home page: <http://c-ruttkies.github.io/MetFrag/>;
- Operating system(s): Platform independent;
- Programming language: Java;
- Other requirements: Java ≥ 1.6 , Apache Maven $\geq 3.0.4$ (for developers);
- License: GNU LGPL version 2.1 or later;
- Any restrictions to use by non-academics: none.

Additional files

Additional file 1. MetFrag2.2 Command Line (CL) general parameters.

Additional file 2. MetFrag2.2 CL local database parameters (*MySQL*, *PostgreSQL*)

Additional file 3. MetFrag2.2 CL - Different Scoring terms (MetFragScore-Types) available for online databases used by MetFrag All or a subset of these values can also be used as a total with CombinedReferenceScore (Table in Additional file 1).

Additional file 4. Top 1 ranks of MetFrag2.2. combined with CFM-ID This figure shows the distribution of the number of top 1 ranks with different weights (100 drawn randomly between 0 and 1) for MetFrag2.2 and CFM-ID. Lightest yellow dot marks the maximum, 62 top 1 ranks at $_{MetFrag} = 0.67$ and $_{CFM-ID} = 0.33$. The red dot at the right marks the minimum, 36 top 1 ranks at $_{MetFrag} = 0.997$ and $_{CFM-ID} = 0.003$. The most left dot marks 49 top 1 ranks at $_{MetFrag} = 0.02$ and $_{CFM-ID} = 0.98$.

Additional file 5. Number of patents and PubMed references shown as CombinedReferences retrieved from PubChem for the Orbitrap XL dataset This figure shows the distribution of the number of references and

patents for all candidates (marked by black dots) retrieved from PubChem for the 359 (unique) correct candidates (marked with green line) and the additional (wrong) candidates retrieved for each query. The queries are sorted by the number of CombinedReferences for the correct candidate, respectively. The intensity of the black dots indicate the number of candidates which overlap at that position.

Additional file 6. Top 5 ranks with PubChem (XlogP3) on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (402 in the top 5), 90th percentile (386), median (375), 10th percentile (325) and worst result (145).

Additional file 7. Top 5 ranks with ChemSpider on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (463 in the top 5), 90th percentile (452), median (440), 10th percentile (385) and worst result (195).

Additional file 8. Top 10 ranks with PubChem (XlogP3) on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (422 in the top 10), 90th percentile (406), median (391), 10th percentile (351) and worst result (187).

Additional file 9. Top 10 ranks with ChemSpider on the Orbitrap XL Dataset The results were obtained with MetFrag2.2 formula query and the inclusion of references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (471 in the top 10), 90th percentile (460), median (450), 10th percentile (404) and worst result (223).

Additional file 10. Top 1 ranks with PubChem (XlogP3) on the UFZ dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (165 in the top 1), 90th percentile (159), median (156), 10th percentile (112) and worst result (11).

Additional file 11. Top 1 ranks with PubChem (XlogP3) on the EQex dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (236 in the top 1), 90th percentile (230), median (225), 10th percentile (162) and worst result (29).

Additional file 12. Top 1 ranks with PubChem (XlogP3) on the EQexPlus dataset The results were obtained with MetFrag2.2 formula query and the inclusion of patents, references and retention time. Each small dot shows the number of first ranks with a given combination of weights. The larger dots show the best result (196 in the top 1), 90th percentile (184), median (181), 10th percentile (142) and worst result (28).

Authors' contributions

SW developed the original MetFrag code under supervision of SN. CR rewrote MetFrag and performed all programming. CR and ES developed the additional features for MetFrag together and performed the data evaluation and interpretation. ES generated the test datasets and examples, CR and ES both prepared the manuscript. JH provided analytical advice on the concepts and integration of additional strategies; SN conceptualized MetFrag and provided advice on the informatics optimization. All authors read and approved the final manuscript.

Author details

¹ Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany. ² Ewag: Swiss

Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland. ³ Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland. ⁴ Present Address: R&D NMR Software, Bruker BioSpin GmbH, Silberstreifen, 76287 Rheinstetten, Germany.

Acknowledgements

The authors thank Heinz Singer, Michael Stravs, Birgit Beck and other members of the Environmental Chemistry Department at Eawag; Tobias Schulze, Martin Krauss and others at the Helmholtz Centre for Environmental Research (UFZ) involved in the acquisition of the Eawag and UFZ datasets, as well as Jennifer Schollée for performing the ChemAxon calculations, Martin Krauss for the unknown at *m/z* 199 data and Felicity Allen for her assistance with CFM-ID. CR acknowledges funding from DFG grant NE/1396/5-1, CR and ES acknowledge funding from EU FP7 project SOLUTIONS under Grant Agreement No. 603437.

Competing interests

The authors declare that they have no competing interests. The contributions of SW were independent of his employment at Bruker.

Received: 3 October 2015 Accepted: 8 January 2016

Published online: 29 January 2016

References

- Schymanski EL, Singer HP, Slobodnik J, Ipolyi IM, Oswald P, Krauss M, Schulze T, Haglund P, Letzel T, Grosse S et al (2015) Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal Bioanal Chem* 407(21):6237–6255
- Hug C, Ulrich N, Schulze T, Brack W, Krauss M (2014) Identification of novel micropollutants in wastewater by a combination of suspect and nontarget screening. *Environ Pollut* 184:25–32
- Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, Ripollés Vidal C, Hollender J (2014) Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol* 48(3):1811–1818
- Stein S (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal Chem* 84(17):7274–7282
- Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, Yanes O (2015) Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects. *Trends Anal Chem (TrAC)*. doi:10.1016/j.trac.2015.09.005
- Wolf S, Schmidt S, Müller-Hannemann M, Neumann S (2010) *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform* 11:148
- National Center for Biotechnology Information (2016) PubChem Database. <https://pubchem.ncbi.nlm.nih.gov/search/search.cgi>. Accessed 14 Jan 2016
- Royal Society of Chemistry (2016) ChemSpider. <http://www.chemspider.com/>
- Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30
- Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34(suppl 1):354–357
- Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF (2008) Mass spectral metabonomics beyond elemental formula: chemical database querying by matching experimental with computational fragmentation spectra. *Anal Chem* 80(14):5574–5582
- HighChem Ltd. (2015) Mass Frontier v. 7. HighChem Ltd, Bratislava
- Schymanski EL, Gallampois CMJ, Krauss M, Meringer M, Neumann S, Schulze T, Wolf S, Brack W (2012) Consensus structure elucidation combining GC/EL-MS, structure generation, and calculated properties. *Anal Chem* 84:3287–3295
- Chiaia-Hernandez AC, Schymanski EL, Kumar P, Singer HP, Hollender J (2014) Suspect and nontarget screening approaches to identify organic contaminant records in lake sediments. *Anal Bioanal Chem* 406(28):7323–7335
- Heinonen M, Shen H, Zamboni N, Rousu J (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* 28(18):2333–2341
- Allen F, Greiner R, Wishart D (2015) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 11(1):98–110. doi:10.1007/s11306-014-0676-4
- Ridder L, van der Hooff JJJ, Verhoeven S (2014) Automatic compound annotation from mass spectrometry data using MAGMa. *Mass Spectrom* 3(Special Issue 2):0033. doi:10.5702/massspectrometry.S0033
- Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci* 112(41):12580–12585. doi:10.1073/pnas.1509788112
- Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. *J Mass Spectrom* 48(3):291–298. doi:10.1002/jms.3123
- Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45:703–714
- Kasama T, Kinumi T, Makabe H, Matsuda F, Miura D, Miyashita M, Nakamura T, Tanaka K, Yamamoto A, Nishioka T (2014) Winners of CASMI2013: automated tools and challenge data. *Mass Spectrom* 3(Special Issue 2):S0039. doi:10.5702/massspectrometry.S0039
- Schymanski EL, Neumann S (2013) CASMI: and the winner is . . . *Metabolites* 3(2):412–439
- Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E et al (2013) HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res* 41(Database issue):D801–D807. doi:10.1093/nar/gks1065
- LfU: Bayerisches Landesamt für Umwelt (2016) STOFF-IDENT (login required). <http://bb-x-stoffident.hswt.de/>. Accessed 14 Jan 2016
- NORMAN Association (2016) NORMAN Suspect List Exchange. <http://www.norman-network.com/?q=node/236>. Accessed 14 Jan 2016
- Netherlands eScience Center (2016) MAGMa Web Interface. <http://www.emetabolomics.org/magma>. Accessed 14 Jan 2016
- Little J, Cleven C, Brown S (2011) Identification of known unknown utilizing accurate mass data and chemical abstracts service databases. *J Am Soc Mass Spectrom* 22:348–359. doi:10.1007/s13361-010-0034-3
- Little J, Williams A, Pshenichnov A, Tkachenko V (2012) Identification of known unknowns utilizing accurate mass data and ChemSpider. *J Am Soc Mass Spectrom* 23:179–185. doi:10.1007/s13361-011-0265-y
- Kováts E (1958) Gas-chromatographische Charakterisierung organischer Verbindungen. Teil 1: Retentionsindices aliphatischer Halogenide, Alkohole, Aldehyde und Ketone. *Helv Chim Acta* 41(7):1915–1932. doi:10.1002/hlca.19580410703
- Dunn WJ, Block JH, PR S (1986) Partition coefficient, determination and estimation. Pergamon Press, Oxford
- Mannhold R, Poda G, Ostermann C, Tetko IV (2009) Calculation of molecular lipophilicity: state-of-the-art and comparison of log P methods on more than 96,000 compounds. *J Pharm Sci* 98(3):861–893. doi:10.1002/jps.21494
- Kern S, Fenner K, Singer HP, Schwarzenbach RP, Hollender J (2009) Identification of transformation products of organic contaminants in natural waters by computer-aided prediction and high-resolution mass spectrometry. *Environmental Sci Technol* 43(18):7039–7046
- Bade R, Bijlsma L, Sancho JV, Hernández F (2015) Critical evaluation of a simple retention time predictor based on LogKow as a complementary tool in the identification of emerging contaminants in water. *Talanta* 139:143–149
- Hogenboom A, Van Leerdam J, de Voogt P (2009) Accurate mass screening and identification of emerging contaminants in environmental samples by liquid chromatography-hybrid linear ion trap Orbitrap mass spectrometry. *J Chromatogr A* 1216(3):510–519
- Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The chemistry development kit (CDK): an open-source java library for chemo- and bio-informatics. *J Chem Inf Comput Sci* 43(2):493–500

36. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the chemistry development kit (CDK)—an open-source java library for chemo- and bio-informatics. *Curr Pharm Des* 12(17):2111–2120
37. Ulrich N, Schürmann G, Brack W (2011) Linear solvation energy relationships as classifiers in non-target analysis—a capillary liquid chromatography approach. *J Chromatogr A* 1218(45):8192–8196. doi:10.1016/j.chroma.2011.09.031
38. Miller TH, Musenga A, Cowan DA, Barron LP (2013) Prediction of chromatographic retention time in high-resolution anti-doping screening data using artificial neural networks. *Anal Chem* 85(21):10330–10337. doi:10.1021/ac4024878
39. Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, Jones C (2015) Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. *Metabolomics* 11(3):696–706. doi:10.1007/s11306-014-0727-x
40. Heller SR, McNaught A, Stein S, Tchekhovskoi D, Pletnev IV (2013) InChI—the worldwide chemical structure identifier standard. *J Cheminform* 5(7). doi:10.1186/1758-2946-5-7
41. Ruttkies C, Strehmel N, Scheel D, Neumann S (2015) Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an in silico generated compound database and MetFrag. *Rapid Commun Mass Spectrom* 29(16):1521–1529
42. Daylight Chemical Information Systems, Inc. (2016) SMILES—a simplified chemical language. <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>. Accessed 14 Jan 2016
43. Ridder L, van der Hoof JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid Commun Mass Spectrom* 26(20):2461–2471. doi:10.1002/rcm.6364
44. Daylight Chemical Information Systems, Inc. (2016) SMARTS—a language for describing molecular patterns. <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>. Accessed 14 Jan 2016
45. Ma Y, Kind T, Yang D, Leon C, Fiehn O (2014) MS2Analyzer: a software for small molecule substructure annotations from accurate tandem mass spectra. *Anal Chem* 86(21):10724–10731
46. National Center for Biotechnology Information (2016) PubChem REST Services. https://pubchem.ncbi.nlm.nih.gov/pug_rest/PUG_REST_Tutorial.html. Accessed 14 Jan 2016
47. Royal Society of Chemistry (2016) ChemSpider MassSpec API. <http://www.chemspider.com/MassSpecAPI.aspx>. Accessed 14 Jan 2016
48. Leo AJ (1993) Calculating log P oct from structures. *Chem Rev* 93(4):1281–1306
49. Wang R, LL Fu Y (1997) A new atom-additive method for calculating partition coefficients. *J Chem Inf Comput Sci* 37(3):615–621
50. Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, Li Y, Wang R, Lai L (2007) Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J Chem Inf Model* 47(6):2140–2148
51. Stravs MA, Schymanski EL, Singer HP, Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom* 48(1):89–99
52. Stravs MA, Schymanski EL (2016) RMassBank Package. <http://www.bioconductor.org/packages/devel/bioc/html/RMassBank.html>. Accessed 14 Jan 2016
53. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 78(3):779–787. doi:10.1021/ac051437y
54. Kerber A, Meringer M, Rucker C (2006) CASE via MS: ranking structure candidates by mass spectra. *Croat Chem Acta* 79(3):449–464
55. Schymanski EL, Meringer M, Brack W (2009) Matching structures to mass spectra using fragmentation patterns: are the results as good as they look? *Anal Chem* 81(9):3608–3617. doi:10.1021/ac802715e
56. ChemAxon (2016) JChem for Excel 15.7.2700.2799. <http://www.chemaxon.com>. Accessed 14 Jan 2016
57. AMBIT (2016) AMBIT Web. <https://apps.ideaconsult.net/ambit2/depict>. Accessed 14 Jan 2016

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com

5.3 Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag

Christoph Ruttkies, Emma L. Schymanski, Nadine Strehmel, Juliane Hollender, Steffen Neumann, Antony J. Williams, Martin Krauss. Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag. *Anal Bioanal Chem.* 411: 4683, 2019. 9 Citations³
<https://link.springer.com/article/10.1007/s00216-019-01885-0>

Contributions

I designed the study, performed data analysis and evaluation together with Emma L. Schymanski, who also prepared the datasets. I implemented the software with the additional features and developed the scoring terms. Nadine Strehmel and Martin Krauss did the experimental lab work and provided advice on the concepts and integration of the developed computational methods. Antony J. Williams registered investigated molecular compounds in the CompTox Chemical Dashboard. The work was supervised and coordinated by Steffen Neumann and Juliane Hollender. I prepared the manuscript together with Emma L. Schymanski.

³<https://scholar.google.com> (accessed on 01/2021)



Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag

Christoph Ruttkies¹ · Emma L. Schymanski^{2,3} · Nadine Strehmel¹ · Juliane Hollender^{3,4} · Steffen Neumann^{1,5} · Antony J. Williams⁶ · Martin Krauss⁷

Received: 25 January 2019 / Revised: 8 April 2019 / Accepted: 30 April 2019 / Published online: 17 June 2019
© The Author(s) 2019

Abstract

Liquid chromatography coupled with high-resolution mass spectrometry (LC-HRMS) is increasingly popular for the non-targeted exploration of complex samples, where tandem mass spectrometry (MS/MS) is used to characterize the structure of unknown compounds. However, mass spectra do not always contain sufficient information to unequivocally identify the correct structure. This study investigated how much additional information can be gained using hydrogen deuterium exchange (HDX) experiments. The exchange of “easily exchangeable” hydrogen atoms (connected to heteroatoms), with predominantly $[M+D]^+$ ions in positive mode and $[M-D]^-$ in negative mode was observed. To enable high-throughput processing, new scoring terms were incorporated into the *in silico* fragmenter MetFrag. These were initially developed on small datasets and then tested on 762 compounds of environmental interest. Pairs of spectra (normal and deuterated) were found for 593 of these substances (506 positive mode, 155 negative mode spectra). The new scoring terms resulted in 29 additional correct identifications (78 vs 49) for positive mode and an increase in top 10 rankings from 80 to 106 in negative mode. Compounds with dual functionality (polar head group, long apolar tail) exhibited dramatic retention time (RT) shifts of up to several minutes, compared with an average 0.04 min RT shift. For a smaller dataset of 80 metabolites, top 10 rankings improved from 13 to 24 (positive mode, 57 spectra) and from 14 to 31 (negative mode, 63 spectra) when including HDX information. The results of standard measurements were confirmed using targets and tentatively identified surfactant species in an environmental sample collected from the river Danube near Novi Sad (Serbia). The changes to MetFrag have been integrated into the command line version available at <http://c-ruttkies.github.io/MetFrag> and all resulting spectra and compounds are available in online resources and in the [Electronic Supplementary Material \(ESM\)](#).

Published in the topical collection *Young Investigators in (Bio-)Analytical Chemistry* with guest editors Erin Baker, Kerstin Leopold, Francesco Ricci, and Wei Wang.

Christoph Ruttkies and Emma L. Schymanski contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00216-019-01885-0>) contains supplementary material, which is available to authorized users.

✉ Emma L. Schymanski
emma.schymanski@uni.lu

✉ Martin Krauss
martin.krauss@ufz.de

¹ Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany

² Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367 Belvaux, Luxembourg

³ Eawag: Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

⁴ Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland

⁵ iDiv - German Centre for Integrative Biodiversity Research (iDiv), Halle-Jena-Leipzig Deutscher, Platz 5e, 04103 Leipzig, Germany

⁶ National Centre for Computational Toxicity (NCCT), United States Environmental Protection Agency, Research Triangle Park, NC 27711, USA

⁷ Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany

Keywords Compound identification · In silico fragmentation · Hydrogen deuterium exchange · High-resolution mass spectrometry · Structure elucidation · Metabolomics

Introduction

The identification of unknown chemicals in complex samples via non-target screening with liquid chromatographic (LC) separation followed by high-resolution (HR) mass spectrometric (MS) analysis remains challenging due to the vast chemical space and still relatively limited coverage of spectra in reference libraries [1, 2]. While techniques such as nuclear magnetic resonance (NMR) spectroscopy yield rich structural information and are well-suited for structure elucidation, NMR is often unachievable with the low concentrations available in complex samples. In LC-HRMS, information about structural properties is obtained by fragmenting detected substances to yield MS/MS spectra. The resulting spectra can then be compared to spectral libraries, or interpreted by software using in silico fragmentation approaches. Unlike NMR, however, the MS/MS spectra typical in LC-HRMS/MS are often information-poor. Thus, alternative ways of obtaining additional structural information are needed for non-target identification methods reliant on LC-HRMS. While techniques such as direct labelling experiments can be used in metabolomics experiments to gain additional information [3, 4], this is impractical in the context of most complex real-world samples, such as environmental samples.

Recently, the inclusion of additional metadata within the in silico fragmenter MetFrag was shown to greatly improve the identification success in the environmental context [5]. While 6% of structures were correctly ranked initially using in silico fragmentation alone with PubChem as a database in this study, this increased to 71% when including metadata such as the retention time, reference, and patent information. Similar results were observed for other in silico fragmenters in the 2016 CASMI contest [6, 7]. However, most metadata scoring terms themselves do not explicitly include the use of structural information to limit candidates, beyond the fragmentation score. While metadata terms such as patent and reference counts provide useful information in some contexts, these could potentially bias the results towards well-known substances and are not useful where no external information is available for the sample or candidate, such as for unknown metabolites or transformation products. Including the retention time alone (without reference information) did not improve candidate ranking greatly [5]. Further approaches for identification, especially in metabolomics, are reviewed elsewhere (e.g., [2]). However, additional ways of obtaining structural information are needed for non-target identification methods reliant on LC-HRMS. One such method of obtaining additional information can be achieved by modifying the analytes prior to performing HRMS, e.g., using hydrogen-deuterium exchange

(HDX). This approach is used in proteomics for probing conformation and structural dynamics (with different experimental setups) and has been used occasionally for structure elucidation of small molecules over the last decades (e.g., [8–12]). HDX experiments can be used to provide information about which functional groups may be present in the compound of interest. When the chromatographic system is flooded with deuterated solvents (e.g., D₂O instead of H₂O, MeOD instead of MeOH), the “exchangeable hydrogens” can be replaced (i.e., exchanged) with deuteriums. When combined with routine (undeuterated—hereafter termed “normal”) measurements, the changes in the fragmentation pattern can yield information about the substructures in the molecule. While this experimental setup is quite expensive due to the relatively large amounts of deuterated solvents required, cheaper methods such as post-column deuteration tend to yield very complex deuteration patterns due to changing fractions of undeuterated and deuterated solvents along an LC gradient elution that require rigorous statistical analysis [8, 13]. This approach is therefore less useful for the identification of unknown substances at this stage.

There are essentially three classes of “exchangeable” hydrogens, shown conceptually in Fig. 1, although the borders between the classes are blurred. The “easily exchangeable” hydrogens attached to the heteroatom groups (OH, NH, SH) would generally be completely exchanged in experiments with a deuterium-flooded chromatographic system [14]; typically, the exchange reactions take place in the microsecond to millisecond time range. Those that are sterically hindered or stabilized by hydrogen bonding may take longer to exchange (starting from several millisecond to minutes), but this is still anticipated to occur in most cases within the contact time in the LC system. Partially exchangeable hydrogens, including some conjugated and aromatic hydrogens (e.g., those on pyrrole rings [15] or affected by keto-enol tautomerism [16]), may also exchange in the liquid phase (during LC separation) and/or the gas phase (during ionization and in the MS), with exchange rates depending strongly on experimental conditions [15–17]. However, as shown in Fig. 1, the “unexchangeable” hydrogens, i.e., aliphatic and most aromatic carbons (CH) would not be expected to exchange during an LC-MS run. Thus, a first hypothesis is formed for structure elucidation of small molecules:

All “easily exchangeable” hydrogens should be replaced with deuterium; some conjugated or aromatic hydrogens may be replaced with deuteriums, whereas any aliphatic and most aromatic CH hydrogens would be expected to remain intact.

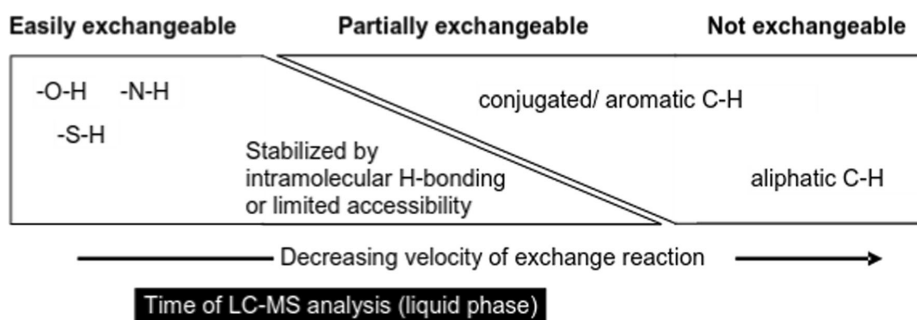


Fig. 1 Conceptual view of the degree of exchangeability of hydrogens relative to the timescale of LC-MS analysis

The influence of deuterium exchange in MS experiments is relevant in both MS1 (full scan) and MS/MS experiments. As deuterium (atomic mass 2.014102 Da) has a different mass to hydrogen (atomic mass 1.007825 Da), the number of deuteriums can be readily determined by the mass difference between the normal and deuterated ion in the full scan (MS1). As the system is flooded with deuterium, the typical ions expected in positive electrospray ionization are no longer $[M+H]^+$, but rather $[M+D]^+$; thus, the presence of two D in the detected ion indicates one exchangeable hydrogen and one D^+ adduct, and so on. In negative ESI, the absence of a mass difference indicates one exchangeable hydrogen, which is abstracted by the ionization process to form an $[M-D]^-$, with an m/z identical to the $[M-H]^-$ ion in the undeuterated eluents (note that without an exchangeable H, ionization in negative mode is difficult). From this information, it is possible to determine the maximum number of easily exchangeable hydrogens available on the molecule. The readiness of partially exchangeable hydrogens to be exchanged within the timeframe of the LC method requires further investigation and this was considered throughout this study. Beyond the full scan, the deuterium mass shift will also be reflected in the MS/MS fragments, and the existence of a deuterated fragment in the MS/MS of the deuterated compound can give valuable information about the molecular structure of the compound.

Thus, the aim of this study was to investigate how hydrogen-deuterium exchange experiments could assist structural elucidation in non-targeted HR-MS experiments using high-throughput, automated in silico fragmentation techniques. The in silico fragmenter MetFrag was modified to include additional scoring terms to account for the HDX starting with the theory discussed above and tested on small datasets. Once the method was established, it was evaluated on a set of several mixtures of environmental chemicals containing 762 unique compounds and analyzed in both positive and negative mode, as well as a smaller dataset of 80 metabolites. HDX experiments were then performed on a water

sample from the river Danube near Novi Sad (Serbia) to assess the feasibility of applying HDX experiments in the context of a complex real-world water sample.

Materials and methods

Experimental data sets

Set 1: Deuterated standards and Orbitrap

To ensure that MetFrag accounted for deuterium exchange substitution correctly during the in silico fragmentation, the initial development was performed on stably labeled deuterated substances (typically used as internal standards) where the location of the deuterium atoms (in the precursor) was known. This also served to diagnose any unexpected phenomena in the fragmentation. A mix of internal standards (1 $\mu\text{g/L}$) was measured on an LTQ Orbitrap XL (Thermo Scientific) with electrospray ionization in positive mode. LC separation was performed in advance on a Kinetex Core-Shell C18 column (3.0 \times 100 mm, 2.6 μm particle size) from Phenomenex with $\text{H}_2\text{O}/\text{MeOH}$ (both with 0.1% formic acid) at a flow rate of 200 $\mu\text{L}/\text{min}$ and a gradient of 90/10 at 0 min, 80/20 at 3.2 min, 5/95 at 17.8 min, 5/95 at 37.8 min, 90/10 at 37.9 min, and 90/10 at 47 min. MS/MS scans were obtained using both higher energy collision-induced dissociation (HCD) at nominal collision energy (NCE) of 100 and collision-induced dissociation (CID) at 35 NCE, an MS/MS isolation width of 1.3 m/z , and resolution of 15,000. Spectra were extracted for DEET-d7, metolachlor-d6, and carbamazepine-d10, summarized in ESM Table S1.

Set 2: HDX and QToF-MS

Individual compounds were dissolved in $\text{MeOH}/\text{H}_2\text{O}$ 80/20 (v/v) at a concentration of 10 mM. Then, ten compounds were combined to one synthetic mixture to give 1 mM and the final concentration of each mixture adjusted to 100 μM using

MeOH/H₂O 50/50 (v/v). Following this, 100 μ L was dried down and the residue redissolved in 100 μ L acetonitrile/deuterium oxide 50/50 (v/v), ultrasonicated for 5 min at room temperature, centrifuged at 16,000 \times g for 2 min, and the supernatant injected onto an UPLC-QTOFMS system (Waters, Eschborn, Germany; Bruker Daltonics, Bremen, Germany) with ESI ionization. For the normal (native, undeuterated) samples, water/formic acid, 99.9/0.1 (v/v), was used as eluent A and acetonitrile/formic acid, 99.9/0.1 (v/v), as eluent B. In contrast, for the deuterium-exchanged samples, deuterium oxide/formic acid, 99.9/0.1 (v/v), was applied as eluent A and acetonitrile/formic acid, 99.9/0.1 (v/v), as eluent B.

Each mixture was measured in both positive and negative ion modes according to [18]. CID mass spectra were acquired using the respective $[M+H]^+$, $[M-H]^-$, or their deuterated equivalent masses, isolated inside the quadrupole using an isolation width of 3 m/z and fragmented inside the collision cell after applying two collision energies (10 eV and 20 eV). All instrument parameters were maintained as previously described in [18]. The resolution was 10,835 (m/z 922) in positive mode and 9632 (m/z 1034) in negative mode, with a mass accuracy of 5 ppm. The MS and MS/MS data were processed with DataAnalysis 4.2 (Bruker Daltonics, Bremen, Germany) prior to use with MetFrag as previously described [19]. Spectra from kinetin, N-(3-indolylacetyl)-L-valine, o-anisic acid, and phlorizin were used in the results presented further below (see ESM Table S2 for more information).

Set 3: Large standard set for HDX and Orbitrap

A total of 22 mixes with 850 substances, already in use at UFZ, were used to measure the large standard set (762 unique substances, i.e., 677, 82, and 3 substances were present once, twice, or three times, respectively, due to the use of the various mixes in the laboratory—see ESM Table S3a). Each mix contained between 10 (mix 15) and 94 (mix 13) substances. Each substance in each mix was assigned a unique identifier, starting at 8000 (a 4-digit number is necessary for RMassBank processing)—such that standards present in more than one mix had two or three identifiers. Each mix was checked for isobars and “near isobars” (substances that would potentially fall within the same MS/MS isolation window of 1.3 m/z); the corresponding identifiers were logged for quality control (see ESM Table S3b). For instance, if the presence of an isobar or near isobar could not be excluded, the substance was eliminated from the test set as the spectral quality could not be guaranteed.

The reference standards were purchased from various suppliers at a minimum purity of 97% and spiked in the mixes at a concentration of 1 μ g/mL. These mixes were then measured on an LC system coupled to a HR-MS/MS (Q Exactive Plus, Thermo). The Ultimate 3000 LC system (Thermo) used a Kinetex C18 EVO column (2.1 \times 50 mm, 2.6 μ M particle size), with a 2.1 \times 5 mm pre-column from Phenomenex and

an injection volume of 5 μ L. The gradient was 95/5 at 0 min, 95/5 at 1 min, 0/100 at 13 min, and 0/100 at 24 min at 300 μ L/min. For normal measurements, solvents A and B were H₂O and MeOH, both with 0.1% formic acid. For the deuterated measurements, the solvents were deuterated water (D₂O, 99.9 atom-% D, Sigma-Aldrich) and deuterated methanol (MeOD, i.e., CH₃OD, 99.5 atom-% D, Sigma-Aldrich), both containing 0.1% (v/v) undeuterated formic acid. Electrospray ionization (ESI) in positive and negative mode was used. MS1 was acquired at a nominal resolving power of 70,000 (referenced to m/z 200); MS/MS were acquired at R = 35,000 using data-dependent acquisition with 5 MS/MS scans following each full scan MS1 and an inclusion list adjusted to each mix. The pesticide mix (mix 13, containing 94 substances) was run three times in positive mode with different inclusion lists to ensure that MS/MS of all compounds were obtained. Higher energy collision dissociation (HCD) was used with stepped 20/35/50 nominal collision energy units (NCE) and an isolation window of 1.3 m/z . All runs were obtained using a range of m/z = 100–1000, except for low mass range runs done on the polar compound mix (mix 19), which was between m/z = 60 and 600. An overview of the mixes and the original acquisition data are given in ESM Table S3a and b, respectively. In addition to this, the polar compound mix (mix 19) was also re-measured on a Synergi Polar RP column (100 \times 3.0 mm, 2.5 μ M particle size, Phenomenex). The dataset for CASMI 2016 [6] was formed from the initial normal measurements of these mixes. A full list of substances and further details (structure, predicted ion masses, etc.) are given in ESM Table S3c.

Environmental water sample

A well-studied sample from the SOLUTIONS project [20] was used to scope the potential to apply HDX to complex environmental samples. The sample was collected from the river Danube near Novi Sad (Serbia) in the plume of an untreated wastewater inlet using on-site large volume solid-phase extraction and enriched 500-fold for analysis as detailed in [21, 22]. The sample was measured under normal and HDX conditions with a data-dependent top 6 experiment (without an inclusion list) and the same collision energies and other conditions as for the large standard set described above, using the Kinetex column. The target analysis results from [22] were used to direct the data evaluation presented in this manuscript, along with a list of suspect surfactants [23–25].

Data processing (set 3)

HDX prediction and registration

The base hypothesis to test was that “easily exchangeable” hydrogens would be exchanged in these experiments; thus,

5.3 Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag

for all 762 substances, a prediction was made to exchange each heteroatom hydrogen with a deuterium (i.e., SH to SD, OH to OD, NH₂ to ND₂). The predicted deuterated formula was then used as a basis to search for deuterated spectra. In terms of the expected mass for each ionization mode, it was assumed that [M+D]⁺ ions would be formed in positive mode and [M-D]⁻ in negative mode (see “Introduction”). An example is given in Fig. 2 and further details are given in the “Implementation” section below. Note that while deuterium is commonly represented as “D,” a convention that we use in the text in this article for readability and consistency, the chemical representation used in the depictions is the isotopic form ²H, which allows for proper interpretation in the cheminformatics toolkits. The predicted deuterated SMILES for all substances are given in ESM Table S3d (note this is the prediction and not all species were observed). These predicted SMILES were used to perform the HDX data extraction (see next section). All observed (and manually verified) HDX features, given in ESM Table S3e-f, were registered in DSSTox, the database behind the CompTox Chemicals Dashboard [26], based on the predicted SMILES and mappings to the original standards. DSSTox was used to generate the remaining structural information presented in ESM Table S3f. The corresponding DSSTox substance identifiers (DTXSIDs) were used to create the HDXNOEX and HDXEXCH lists of undeuterated and deuterated species.

MS data processing

The raw data files from Thermo were converted to mzML using a front-end for MSConvert (from ProteoWizard [28])

written by U. Schmitt (SIS, ETHZ), using vendor centroiding, zero value removal, and zlib compression. The MS/MS of the standards were extracted using RMassBank [29]. The “normal” runs were processed in the typical RMassBank workflow, using the SMILES code for each chemical. As RMassBank could not (initially) handle deuterium when the data was extracted (due to issues with the Chemistry Development Kit that have subsequently been resolved [30]), the HDX data were extracted using the exact mass only, which meant that recalibration and noise removal was not performed on these data. Retention times (RTs) from the normal data were used initially, with a window of 0.4 min. Substances with RTs that were unknown were extracted using the RT at maximum EIC intensity for the precursor mass; for multiple peaks, these were determined manually. All RTs were checked manually and refined where necessary for those substances with missing results. For the normal runs, peak annotation and reanalyzed peaks options were both “true.” The recalibration was performed using loess fitting (see [29]) on assigned fragments and the MS1 data, using dppm. The MS1 and MS/MS were recalibrated together, with an initial window of 15 ppm. The multiplicity filter was set to 1 (as only one spectrum was recorded). All additional settings were the default ones (see file online). The extraction of the MSMS data was checked both visually and using a summary of the data (see Figures and Tables in the ESM). InChIKeys were used to check for duplicate chemical structures, while the spectral hash (SPLASH) [31] was used to detect identical extracted spectra for different substances. Data processing was all performed in the R programming language unless explicitly mentioned elsewhere.

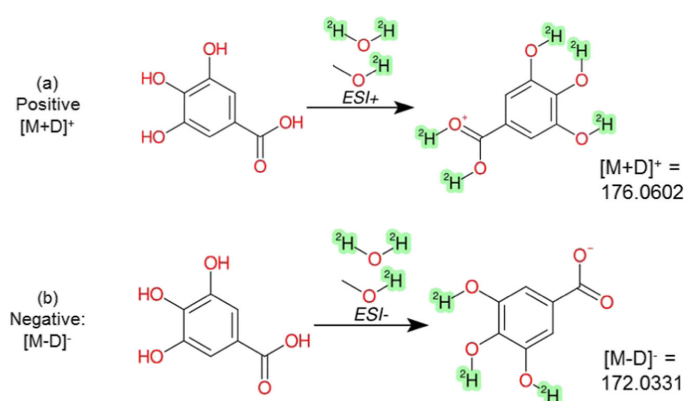


Fig. 2 Example of expected HDX behavior of gallic acid (DTXSID0020650) in the experiment performed here in **a** positive ESI mode to produce [M+D]⁺ and **b** negative ESI mode to produce [M-D]⁻, along with the calculated ion masses that were subsequently observed in the experimental measurements. The quadruply deuterated species of gallic acid is available here (DTXSID60892625). Images created using

CDK Depict [27] with SMARTS highlighting to indicate the deuterium. Note that while we refer to deuterium as “D” throughout the manuscript for simplicity, the depiction with ²H here is consistent with the standard representation of isotopes and enables the SMARTS-based highlighting shown

Implementation of HDX in MetFrag

MetFrag is a Java-based *in silico* fragmenter that uses the Chemistry Development Kit (CDK) [30, 32, 33] to read, write, and process chemical structures. The candidates are generally retrieved from compound databases using the neutral monoisotopic mass (calculated from the precursor) and a given relative mass deviation, the neutral molecular formula of the precursor or a set of database-dependent compound identifiers. Further details on MetFrag are given elsewhere [5, 34].

The starting point for performing MetFrag on HDX data is the acquisition of two independent LC-MS/MS runs of one sample, where the first sample is acquired normally with undeuterated solvents (e.g., MeOH/H₂O) and where at least one of the mobile phases is replaced with a deuterated equivalent during the second acquisition (e.g., MeOD/D₂O, ACN/D₂O). This yields two data sets and corresponding MS/MS spectra pairs (S_H , S_D) have to be collected where the precursor is in its normal form ("H") and in its deuterated form ("D"), where $S_H = \{P_1, \dots, P_N\}$ contains N and $S_D = \{dP_1, \dots, dP_M\}$ M MS/MS peaks (middle part of Fig. 3). Each peak is defined by a m/z (mass to charge ratio) value $m(P_N)$ (for simplicity, we do not take into account intensities here). As reference standards were used in this manuscript, the expected deuterated species were predicted (based on the number of easily exchangeable Hs, as described above). These predicted masses

were then used to extract the HDX MS/MS data, which was verified as described above. The undeuterated candidates were then deuterated *in silico* and matched to the experimental data, then combined using various scoring terms to yield the overall candidate rankings. Details on the generation and combination of these results are given below.

In silico deuteration of candidate structures

To use MetFrag's *in silico* fragment generation for deuterated compounds, the algorithm was adapted to handle deuteriums as well as hydrogens. Furthermore, the MetFrag algorithm was extended to generate an *in silico* deuterated candidate list for a given MS/MS spectrum S_D . First, MetFrag determines the number of experimentally exchanged hydrogens (X), which is calculated by the mass differences of the precursors of S_H and S_D as mentioned earlier. Given the candidate list C derived from a database search (e.g., PubChem [35], ChemSpider [36], or CompTox [26]), based on the precursor information (calculated monoisotopic mass, molecular formula) of the normal spectrum S_H , MetFrag generates an *in silico* deuterated candidate list dC . For a candidate $C_i \in C$, the number of easily exchangeable hydrogens ($eH(C_i)$) are determined by counting the number of hydrogens attached to oxygens, sulfurs, and nitrogens, namely hydroxyl/carboxyl, thiol, and amino groups. A graph-based approach is used to perform a

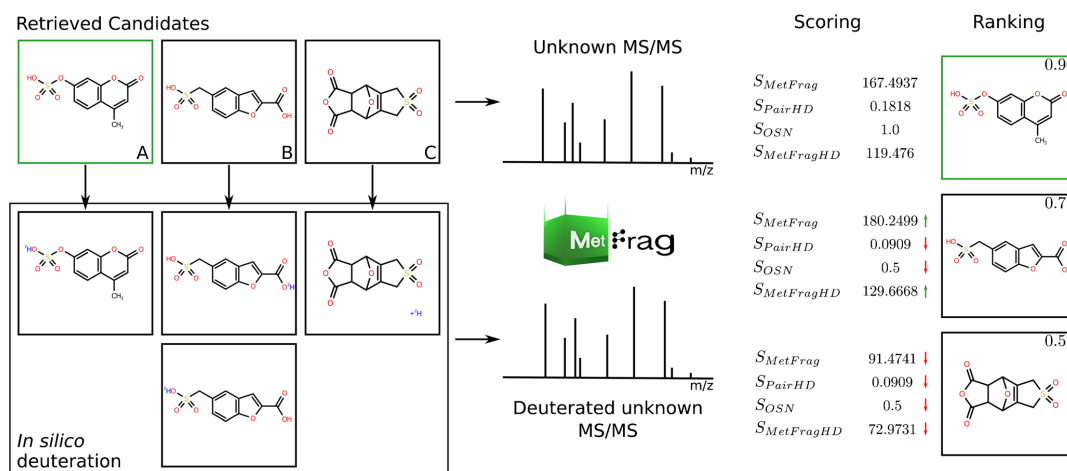


Fig. 3 Workflow for MetFrag to analyze deuterated MS/MS spectra using the example of 4-methylumbelliferyl sulfate (a, green border) of the large standard set. The mass difference of the determined *neutral* precursor masses of the normal (256.0042 Da) and the deuterated (257.0104 Da) spectrum indicated $X = 1$, i.e., one exchanged hydrogen as shown for (a). Two additional selected candidates (b, c) illustrate different *in silico* deuteration cases where the retrieved candidate can result in two deuterated candidates (b) or one candidate of variable deuterium location as no easily exchangeable H is present (c). Processing normal and deuterated candidates with MetFrag-HDX results in four scoring

terms for each candidate, which are combined in a consensus score using weight parameters retrieved during the cross-validation (~ 0.109 , ~ 0.004 , 0.497 , ~ 0.39 ; see **Methods**; note, scores are normalized to range [0, 1]). This resulted in a top 1 ranking of the correct candidate 4-methylumbelliferyl sulfate. Green and red arrows mark scores that are higher or lower compared to those of the correct candidate. Candidate b is the top scoring candidate using $S_{MetFrag}$ alone (without HDX information). This example illustrates both the workflow and the benefit of the additional scoring terms

simple search for the easily exchangeable Hs. During the method establishment, hydrogen/deuterium exchange was predicted assuming that all easily exchangeable hydrogens were 100% replaced with deuterium. This formed the “base case” for in silico deuteration and could be used to reject C_i as potential correct candidate in case ($eH \neq X$). However, there are reasons why $eH(C_i)$ and X can differ, even when C_i is the correct candidate:

- (a) Hydrogens attached to conjugated and/or aromatic carbons could be exchanged due to keto-enol tautomerism or by gas-phase reactions in the ESI source and thus the number of easily exchangeable hydrogens during measurement changes;
- (b) easily exchangeable hydrogens might be stabilized by intramolecular hydrogen-bonding or sterically hindered; and
- (c) the wrong isotopic peak was selected during data-dependent acquisition, leading to the wrong number of experimentally exchanged hydrogens (X).

Thus, different cases need to be handled for the in silico deuteration. Exactly one deuterated candidate is generated by exchanging all easily exchangeable hydrogens in case ($eH = X$). Exactly one candidate is also generated in case ($eH < X$) by exchanging all easily exchangeable hydrogens of C_i and exchanging ($X - eH(C_i)$) variable hydrogens ($vH(C_i)$) of C_i assuming that also aliphatic and/or aromatic hydrogens are replaced without knowing the exact position (as the exact position of the Hs is not necessarily required explicitly during the fragmentation). Where ($eH(C_i) > X$), MetFrag generates every combination of deuterated candidates where X out of $eH(C_i)$ easily exchangeable hydrogens are exchanged by deuterium, which results in (X choose $eH(C_i)$) deuterated candidates for C_i . Figure 3 shows example candidates for all three cases. This approach uses a modified version of the method used for in silico derivatization in [19]. The in silico deuteration method is available as a jar file and included as ESM. The predicted candidates are given in ESM Table S3d.

Scoring terms

To incorporate the information gained by additional deuterated experimental MS/MS spectra, different scores are calculated by MetFrag. Altogether, MetFrag calculates four scoring terms for a candidate C_i that are combined into a final (consensus) score. The regular *FragmenterScore* ($S_{\text{MetFrag}}(C_i)$), already introduced in [5], calculates the match of in silico-generated fragments $\text{Frag}_{i,n}$ of a candidate C_i to the experimental MS/MS peaks P_n of S_H , taking into account the relative intensity of a matched MS/MS peak, the m/z value, and the sum of the bond dissociation energies (BDEs) of the

candidate bonds that were cleaved to generate the matching fragment.

The *HDFragmenterScore* ($S_{\text{MetFragHD}}(C_i)$) uses the same calculation rule as the regular *FragmenterScore* with the same generated fragments but incorporates the information of exchanged hydrogens from the precursor candidate C_i . This information is used to adapt calculated fragment masses to match against m/z peaks dP_m from the deuterated MS/MS spectrum S_D as illustrated in Fig. 4. The mass of a deuterated fragment $d\text{Frag}_{i,n}$ is then calculated as

$$m(d\text{Frag}_{i,n}) = m(\text{Frag}_{i,n}) + eH(\text{Frag}_{i,n}) \cdot (m(D) - m(H)); \quad (1)$$

where $m(\text{Frag}_{i,n})$, $m(H)$, and $m(D)$ are the masses of the normal fragment, a hydrogen, and a deuterium, respectively.

Equation 1 simulates the exchange of a number $eH(\text{Frag}_{i,n})$ of easily exchangeable hydrogens with deuterium for the related fragment. Where $vH(C_i) \neq 0$, MetFrag also tries to find a match based on a variable number of exchanged hydrogens by adapting fragment masses with

$$m(d\text{Frag}_{i,n}) = m(d\text{Frag}_{i,n}) + k \cdot (m(D) - m(H)); \quad (2)$$

where $1 \leq k \leq vH(d\text{Frag}_{i,n})$ to simulate an additional exchange of non-easily exchangeable hydrogens. As for the mass of the normal fragment $\text{Frag}_{i,n}$, the adduct mass value c is added/subtracted also for $d\text{Frag}_{i,n}$, which is usually the mass of a proton in the undeuterated case and thus the mass of D^+ for the deuterated case.

The *HDFragmentPairScore* ($S_{\text{pairHD}}(C_i)$) counts matching fragment pairs ($\text{Frag}_{i,n}$, $d\text{Frag}_{i,n}$) between the normal and deuterated MS/MS spectrum. If a fragment $\text{Frag}_{i,n}$ matches a peak in the normal MS/MS spectrum S_H and the corresponding deuterated fragment $d\text{Frag}_{i,n}$ matches a peak in the deuterated MS/MS spectrum S_D , it will be counted as a valid pair. For the matched MS/MS peaks $P_n \in S_H$ and $dP_m \in S_D$, the number of exchanged hydrogens k can be calculated by

$$|m(P_n) + k \cdot (m(D) - m(H)) - m(dP_m)| \leq c \quad (3)$$

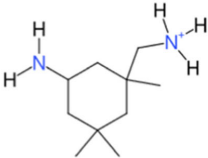
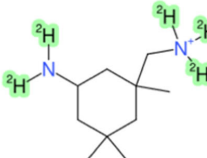
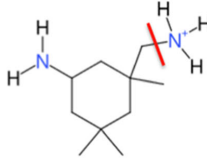
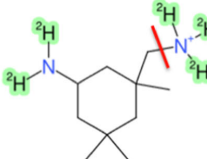
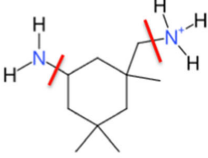
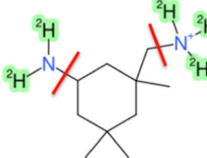
where c is a predefined mass deviation and $k \leq X$. A fragment pair is only counted if the number of deuteriums of $d\text{Frag}_{i,n}$ are equal to k , so

$$eH(d\text{Frag}_{i,n}) + vH(d\text{Frag}_{i,n}) = k; \quad (4)$$

with $0 \leq k$, where a pair is also counted, if $k = 0$ and $eH(d\text{Frag}_{i,n}) + vH(d\text{Frag}_{i,n}) = 0$ meaning $d\text{Frag}_{ij}$ carries no deuterium.

The *HDExchangedHydrogensScore* ($S_{\text{OSN}}(C_i)$) shown in Eq. 5 boosts candidates whose predicted number of easily

Fig. 4 Modified in silico fragmentation workflow, demonstrated on isophorone diamine (DTXSID6027503). In silico-generated fragments from normal mode (left) are modified by exchanging and adding deuteriums at predicted positions (right, green shading) from the precursor molecule. The normal precursor is used to determine possible positions of hydrogen/deuterium exchange (here the amino groups). This information is used during the in silico fragmentation to perform mass calculation of deuterated fragments (left)

	NORMAL	HDX
Precursor		
<i>m/z</i> , formula	171.1856, C ₁₀ H ₂₃ N ₂ ⁺	176.2170, C ₁₀ H ₁₈ D ₅ N ₂ ⁺
Tree Depth 1		
<i>m/z</i> , formula	154.1590, C ₁₀ H ₂₀ N ⁺	156.1716, C ₁₀ H ₁₈ D ₂ N ⁺
Tree Depth 2		
<i>m/z</i> , formula	137.1325, C ₁₀ H ₁₇ ⁺	137.1325, C ₁₀ H ₁₇ ⁺

exchangeable hydrogens $eH(C_i)$ matches the number of experimentally exchanged hydrogens X and discriminates those the more the higher the two values deviate from each other assuming that all and only easily exchangeable hydrogens are exchanged in most of the cases.

$$S_{CI(OSN)} = 1/(|X - eH(C_i)| + 1) \quad (5)$$

The four scoring terms are calculated for all candidates C_i in the candidate list C and are normalized by the maximum value within C . The final score, which is used to rank the candidates C_i , is calculated by the weighted sum (represented by the respective weighting terms ω), as shown in Eq. 6.

$$S_{C_i} = \omega_{MetFrag} \cdot S_{MetFrag}(C_i) + \omega_{MetFragHD} \cdot S_{MetFragHD}(C_i) + \omega_{PairHD} \cdot S_{PairHD}(C_i) + \omega_{OSN} \cdot S_{OSN}(C_i) \quad (6)$$

In case more than one deuterated candidate exists for a given candidate C_i , the maxima of $S_{MetFragHD}(C_i)$ and $S_{PairHD}(C_i)$ over the generated deuterated candidates are used for Eq. 6.

Evaluation and optimization

To test the workflow, the adapted MetFrag algorithm was used to process all spectra pairs from sets 2 and 3. Candidates were retrieved by querying the ChemSpider database (June, 2017) with the formula of the correct precursor molecule. Candidates consisting of non-covalently bound substructures (e.g., salts) and containing non-standard isotopes (like ¹³C) were filtered out and not considered for the final scoring. For the processing of the normal and deuterated MS/MS peak lists, a relative and absolute mass deviation of 5 ppm and 0.001 Da was used for set 3 and 10 ppm and 0.01 Da for set 2 to match in silico-generated fragments to experimental MS/MS peaks. MetFrag calculated the four scoring terms $S_{MetFrag}(C_i)$, $S_{MetFragHD}(C_i)$, $S_{PairHD}(C_i)$, and $S_{OSN}(C_i)$ for each of the candidates. The weights $\omega_{MetFrag}$, $\omega_{MetFragHD}$, ω_{PairHD} , and ω_{OSN} were optimized by a randomized grid search for which 1000 weight combinations were drawn uniformly from the simplex. The optimal weight combination was determined by maximizing the number of correctly top 1 ranked candidates among the MS/MS spectra pairs in the training set. In case several candidates shared the same final score as the correct one, the average rank was reported. Prior to the ranking,

duplicate entries within the candidate list were filtered based on the first part of the candidates' InChIKey. The optimization was performed by a tenfold cross-validation for the large standard set (set 3) with a randomized fold assignment of the spectra pairs. Due to a lower number of spectrum pairs, a leave-one-out cross-validation was used for set 2. To determine the influence of the scoring terms on the ranking results for set 3, the same cross-validation (same fold assignment) was repeated by considering different sets of scoring terms used to calculate the final score S_C . The term combinations considered were $\{S_{\text{MetFrag}}(C_i), S_{\text{MetFragHD}}(C_i), S_{\text{PairHD}}(C_i)\}$, $\{S_{\text{MetFrag}}(C_i), S_{\text{MetFragHD}}(C_i), S_{\text{OSN}}(C_i)\}$, and $\{S_{\text{MetFrag}}(C_i), S_{\text{MetFragHD}}(C_i)\}$.

Results

Set 1: Fragmentation of deuterated standards

To extend MetFrag to deal with deuterium, MS/MS spectra of three deuterated (internal) standards (where the location of deuterium is known and not expected to undergo any form of exchange during the experiment) were extracted using RMassBank and compared with QExactive spectra of the corresponding undeuterated substances available in MassBank. The three standards (DEET and DEET-d7, metolachlor and metolachlor-d6, carbamazepine and carbamazepine-d10) are shown in ESM Table S1, along with database identifiers and the corresponding best-matching MassBank spectrum. Table S4 (see ESM) shows the two main example fragment pairs from DEET and DEET-d7, with formulas as annotated by MetFrag and proposed fragment structures. The corresponding MS/MS spectra are given in ESM Fig. S1.

The spectrum of metolachlor-d6 (see ESM Fig. S2) revealed more interesting fragmentation information than DEET for the MetFrag results, as the deuteration was for only 6 of the total 22 hydrogens. As expected, the undeuterated fragment $C_4H_9O^+$ at m/z 73.0648, lost from the nitrogen, was observed as $C_4H_3D_6O^+$ at m/z 79.1022 for metolachlor-d6 (see ESM Table S1). Corresponding m/z fragments prior to the loss of this group were also seen, e.g., $C_{12}H_{18}N^+$ (m/z 176.1434) in the undeuterated molecule and $C_{12}H_{12}D_6N^+$ (m/z 182.1815) in the deuterated molecule. However, many fragments associated with the aromatic group (originally undeuterated) were also observed incorporating one or more deuteriums. This indicates that the replacement of Hs with Ds can also occur at the aromatic ring in the collision cell, either due to rearrangement reactions involving a movement of Ds in activated gas-phase ions (scrambling) or an exchange with other species present in the cell [37, 38]. Examples observed at high intensities in the MS/MS spectra included $C_7H_7^+$ (m/z 91.0542) to $C_7H_6D^+$ (m/z 92.0603); $C_6H_7N^+$ (m/z 93.0573) to $C_6H_6DN^+$ (m/z 94.0632) and $C_6H_5D_2N^+$ (m/z 95.0698); $C_7H_{10}N^+$ (m/z 108.0807) to $C_7H_9DN^+$ (m/z 109.0872) and

$C_7H_8D_2N^+$ (m/z 110.0933). The most important conclusion from this exercise for MetFrag, apart from the successful method development, that this mobile deuterium in the collision cell should be considered dynamically, similar to hydrogen [5], i.e., fragments can be explained with up to one or two additional hydrogens or deuteriums.

Set 2: QToF HDX experiments

The spectra from this test set, although a minor contribution in comparison to the larger standard set described below, were invaluable in establishing and testing the scoring strategy implemented in MetFrag before the complete large standard set was available. However, the results do illustrate the impact of lower mass accuracy in HDX as obtained by the used QToF instrument. The results retrieved for selected compounds are given in ESM Table S2 along with the structures and the weights of the scoring function and the resulting ranks. The candidates were retrieved with a ChemSpider query as described above. The top row per compound contains the results considering only MetFrag without the deuterated scoring terms, while the lower two rows show results with different weightings (given in ESM Table S2) of all terms. The table shows clearly for each example that the candidate ranking and thus the results are improved when considering the information from the deuterated experiments. Drastic improvements are obtained for the examples N-(3-indolylacetyl)-L-valine and phlorizin where the rankings improved from 97 to 25 and from 14 to 3.5, respectively. While the original results for this test set actually eliminated candidates that exchanged fewer H atoms, subsequent testing revealed that this could potentially result in the elimination of correct candidates. As a result, the methods were adjusted to the final strategy presented in this publication, where all candidates are scored and the scores are used to provide relative rankings, rather than performing a hard elimination of any candidates not exactly matching the theory. All further validation was performed on the large standard set, described below, as this was a much more comprehensive dataset and the greater substance numbers were required for a more comprehensive evaluation of the method.

Set 3: Evaluation on large standard set

Experimental results on large standard set

As described in the methods, several mixtures were measured to obtain the experimental data for the HDX method development and validation. Several re-measurements were undertaken to confirm observations and obtain the highest quality MS/MS spectra possible. In total, pairs of spectra (i.e., valid MS/MS spectra in both normal and HDX measurements) were found for 592 of the 762 unique substances measured. As described in the methods, these were quality controlled with

automated curation, control checks, and automated plotting of extracted spectra and spectral pairs. All spectra were verified manually by at least two of the authorship team, including cross-checks in the vendor software. The results generally matched very well with the theory explained above, and were overall better than anticipated given the large structural diversity and myriad of functional groups and properties in this large standard set. An overview of all observed retention times plus respective columns and measurement is given in ESM Table S3e. The chemical information associated with all of these observed species, including number of deuteriums exchanged and deuterated structures (where applicable), is given in ESM Table S3f. These observed structures are available for readers to download (https://comptox.epa.gov/dashboard/chemical_lists/hdxexch). The full substance listing is also available at https://comptox.epa.gov/dashboard/chemical_lists/hdxnoex (reference standards only, not including the deuterated species).

Example chromatograms (one normal, one HDX, ESI positive mode) for the pesticide mix are given in the ESM (Fig. S3). This shows that overall, the chromatograms look similar in many places, although peaks are clearly shifted slightly (sometimes lower, sometimes higher retention times—for instance, 5.51 to 5.80 min and 13.46 to 13.36 min in normal and HDX conditions, respectively). In the isocratic region (after approx. 15 min), peaks appear at much higher intensity in the HDX chromatogram than in the normal chromatogram for the Kinetex column—a phenomenon that was reproducible in both the standard mixes and environmental samples (discussed further below). The normal vs HDX retention times

over all mixes for the final compiled dataset are plotted in Fig. 5 for the Kinetex column. The retention times are generally on the 1:1 line (with some small deviations at very early retention times) until approximately 13 min, where the elution regime changes from gradient to isocratic with 100% MeOH/MeOD, respectively. Several compounds are still on the 1:1 lineup to 16 min, while others deviate markedly from this trend, eluting up to 25 min in normal mode but by 16 min in HDX. The latter structures were all surfactants with a polar head group and a long, apolar tail. Two of the most extreme examples are dodecylbenzenesulfonic acid (DTXSID8050443) and perfluorotetradecanoic acid (DTXSID3059921), as shown in Fig. 5. Despite these few extreme examples, the average retention time shift over all standards was 0.04 min. A figure showing the retention time vs change in retention time between the columns is included in the ESM (Fig. S4), including additional example structures for standout data points. While the change in physicochemical properties from the normal to the deuterated eluents hardly affects the compound retention during the relatively steep gradient elution, these differences have a much larger effect on surfactants during the isocratic elution. For the Synergi column, the average retention time shift was 0.35 min, but note this was only for 45 substances measured with a long chromatographic gradient to enable better separation.

The majority of MS/MS spectra, 505 pairs, were found in positive ion mode, while 155 pairs of spectra were found in negative ion mode (68 substances had pairs in both modes). A summary of the MS/MS information is given in ESM Table S3g. While fewer substances ionize in negative mode,

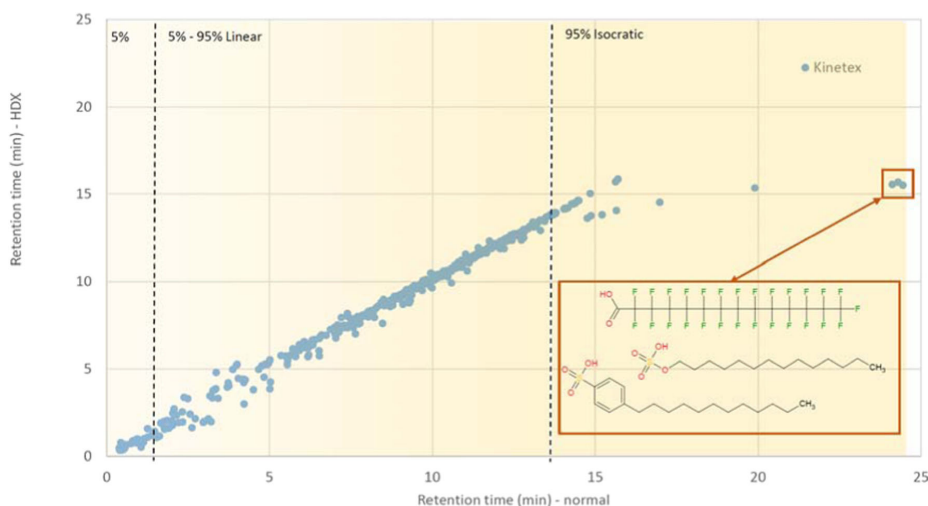


Fig. 5 Retention time (in minutes) of all (unique) substances detected in normal (x axis) and HDX (y axis) conditions for the substances measured on the Kinetex column (both ESI positive and negative modes). The gradient and percentage of methanol (normal) are marked with yellow

highlighting and dashed lines. Examples for the extreme retention time shifts observed are given in the box and in ESM Fig. S4; for explanations, see text

5.3 Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag

there was also a significant loss of intensity in the negative mode HDX spectra (reproducible across several measurements) that contributed to the significantly lower proportion of negative mode pairs. While intensity losses were also observed in positive mode, the generally higher intensity values in positive ESI resulted in many more spectral pairs in positive mode. The average maximum intensities across the MS/MS acquired from the major three chromatographic runs (first measurements and bulk re-measurements on the Kinetex column plus the Synergi runs) were 2.21×10^8 for positive normal, 1.03×10^8 for positive HDX (both over 499 observations), 1.75×10^7 for negative normal, and 9.57×10^6 for negative HDX (over 153 observations). The highest maximum intensities observed in the MS/MS (in the same order) were 4.7×10^9 , 2.1×10^9 , 2.4×10^8 , and 1.3×10^8 , while the lowest maximum intensity was 1.7×10^5 , 5.6×10^4 , 3.8×10^4 , and 2×10^4 . Based on experience, a maximum intensity above 1×10^5 in the MS/MS is required (for this instrument) for a sufficiently informative spectrum; thus, part of the manual checks performed was to judge whether the extracted MS/MS were of sufficient intensity, and thus quality, for the purposes of this study. A further overall factor to consider was the number of fragments observed. The average number of fragments (same order as previously) was 30, 50, 11, and 28 fragments (see ESM Table S3g for a full listing). Note that while more fragments were observed for HDX (50 vs 30, 28 vs 11), this is both due to the potential for more fragments on account of the exchange behavior but also because a less rigorous cleanup was performed (see “Methods” section and Fig. 6

below). Furthermore, the presence of more fragments reduces the intensity of single fragments and this could partially explain the intensity losses observed in the HDX spectra. The maximum number of fragments observed was 267, 383, 104, and 112, respectively, with minimum 1 for all categories except negative HDX (5). Visual checks were performed to eliminate the presence of spectra that may just be noise or where the pairs appeared to completely mismatch, or where only peaks resulting from the precursor (or higher) were present, as these are not accounted for during MetFrag processing. Following all manual checks, 499 spectral pairs remained for positive mode and 148 for negative mode (see ESM Table S3g). This dataset formed the basis for the MetFrag Score validation (see next section).

In the end, matching pairs were observed as one or both of $[M+H]^+/[M+D]^+$ and $[M-H]^-/[M-H/D]^-$ for 592 of the original 762 substances (78%) and 579 (76%) of these were used further for method development following manual checks. For 170 substances, no valid pairs were observed for a number of reasons, which are clarified in the following examples. It is possible that some “pairs” have been falsely eliminated in the quest for optimal data quality. For instance, in positive mode, retention times were determined for 656 of 850 (non-unique) $[M+H]^+$ species over the two major runs of all mixes, whereas only 631 RTs could be determined for the equivalent $[M+D]^+$ species—in the vast majority of cases due to lack of intensity, poor peak shape or evidence of interfering co-elution. Overall, very little evidence of partial or incomplete exchange was observed. For negative mode, retention times could be

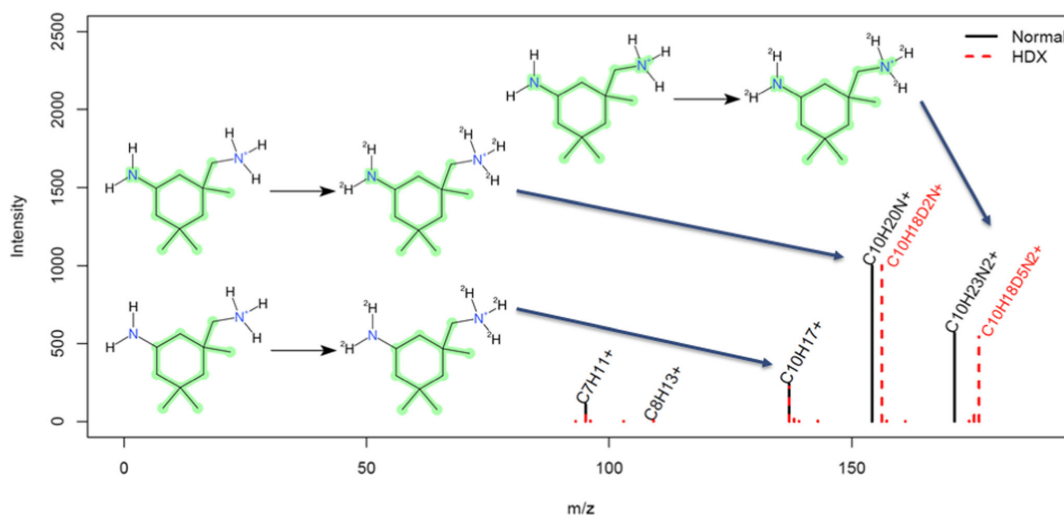


Fig. 6 Observed normal (black) and HDX (red dashed) MS/MS fragments for isophorone diamine (DTXSID6027503) showing the $[M+D]^+$ ion (shifted by 5 mass units, as expected when 4D are exchanged plus an additional D is gained in ionization), then a NH_3/ND_3 loss to yield a

fragment pair with a 2 mass unit shift, then a subsequent NH_2/ND_2 loss to yield the identical $C_{10}H_{17}^+$ fragment with no more deuterium present. Images created using CDK Depict; the highlighting indicates the remaining “backbone” of the structure, as represented in MetFrag

determined for 206 [M-H]⁻ species and 195 [M-H/D]⁻ species according to the theory described in the methods; no substances exhibiting partial exchange were noted, but as stated above, the intensity losses in negative mode made it difficult to find valid pairs in some cases. A few substances were not extracted due to incorrect structural information in the original compound lists used to perform the data extraction (i.e., SMILES and name mismatch, which only became obvious during quality control)—while the tables presented in ESM Table S3 have been extensively curated and present the correct structural information to the best of our knowledge, the spectra were not re-extracted from the raw data for the cases where these errors were discovered too late and resulted in the wrong masses and wrong predicted structures, etc. A further case resulting in the most “non detects” for positive mode was the formation of adducts other than [M+H]⁺, resulting in the loss of 13 substances expected as [M]⁺ and another (Abamectin) observed almost exclusively as [M+Na]⁺ and [M+NH₄]⁺. Although MetFrag can handle different adduct settings, for the purpose of simplicity for the method establishment here (and due to the low number of adducts observed resulting in very small datasets), it was decided to evaluate the [M+H]⁺/[M+D]⁺ and [M-H]⁻/[M-D]⁻ cases only in the material presented here. Alternative adducts were not investigated in negative mode due to the intensity issues, which made it difficult to draw any form of conclusion. As measurements were performed on several mixes rather than individual compounds, it is also worth noting that these mixtures were chosen

partially for analytical convenience and many substances present in some mixes would require a more specialized chromatography for optimal measurement (e.g., many steroids and amines) and it was not expected that all substances would be observed in these experiments. This compromise was necessary to obtain the data presented here, as flooding a complete chromatographic system with deuterated solvents leads to an approximately 50 times cost increase per run above regular solvents (see [discussion](#) below).

The results achieved exceeded expectations in many ways and many high-quality normal and HDX spectra were obtained. As an example, the observed spectra (normal and HDX mode) for isophorone diamine, DTXSID6027503, are shown in Fig. 6 (a small compound has been chosen for clarity). The fragmentation is explained in the figure and caption.

MetFragHDX score validation

As described in the “Methods” section, four scoring terms were considered to account for the additional information arising from HDX experiments in MetFrag (see Eq. 6). The final selection of MS/MS pairs (as described above) was used in the evaluation of the scoring terms (note that a total of 498 spectra were used in positive mode as one compound was measured on both columns). The results are given in Table 1. The improvement in rank was much clearer for set 3, where the Top 1 ranks increased from 49 (10%) using the original MetFrag scoring alone to 78 (16%) by incorporating HDX information

Table 1 Absolute number (%) of top 1, 3, 5, and 10 ranks for MetFragHDX Score combinations for set 2 (57 and 63 MS/MS spectra) and set 3 (498 and 147 spectra) in positive and negative modes respectively. Results for all score terms and MetFrag only are shown for set 2;

various combinations for set 3. Although some of the individual scores do not have good ranking performance, the combination of all 4 terms results in a clear improvement. The combination of all four terms outperformed the tested combinations of 2–3 terms

Set 2 (QTOF)	Positive (n = 57)				Negative (n = 63)			
	Top 1	Top 3	Top 5	Top 10	Top 1	Top 3	Top 5	Top 10
MetFrag,PairHD,OSN,MetFragHD	4 (7%)	9 (16%)	15 (26%)	24 (42%)	2 (3%)	13 (21%)	19 (30%)	31 (49%)
MetFrag	4 (7%)	8 (14%)	11 (19%)	13 (23%)	1 (2%)	4 (6%)	5 (8%)	14 (22%)
Set 3 (Orbitrap)	Positive (n = 498)				Negative (n = 147)			
	Top 1	Top 3	Top 5	Top 10	Top 1	Top 3	Top 5	Top 10
MetFrag,PairHD,OSN,MetFragHD	78 (16%)	189 (38%)	251 (50%)	320 (64%)	20 (14%)	64 (44%)	90 (61%)	106 (72%)
MetFrag,PairHD,OSN	74 (15%)	192 (39%)	254 (51%)	321 (64%)	20 (14%)	61 (41%)	86 (59%)	106 (72%)
MetFrag,MetFragHD,PairHD	56 (11%)	145 (29%)	197 (40%)	255 (51%)	15 (10%)	48 (33%)	74 (50%)	86 (59%)
MetFrag,MetFragHD,OSN	76 (15%)	191 (38%)	255 (51%)	322 (65%)	21 (14%)	67 (46%)	89 (61%)	107 (73%)
MetFrag,MetFragHD	59 (12%)	152 (31%)	202 (41%)	258 (52%)	18 (12%)	49 (33%)	68 (46%)	82 (56%)
MetFrag,PairHD	51 (10%)	146 (29%)	200 (40%)	250 (50%)	16 (11%)	49 (33%)	69 (47%)	84 (57%)
MetFrag,OSN	74 (15%)	193 (39%)	253 (51%)	320 (64%)	21 (14%)	62 (42%)	86 (59%)	107 (73%)
PairHD,OSN	30 (6%)	109 (22%)	154 (31%)	224 (45%)	12 (8%)	46 (31%)	68 (46%)	90 (61%)
MetFragHD,PairHD	56 (11%)	133 (27%)	189 (38%)	238 (48%)	13 (9%)	42 (29%)	61 (41%)	78 (53%)
MetFrag	49 (10%)	130 (26%)	177 (36%)	238 (48%)	18 (12%)	47 (32%)	61 (41%)	80 (54%)
PairHD	26 (5%)	82 (16%)	121 (24%)	165 (33%)	8 (5%)	33 (22%)	54 (37%)	68 (46%)
OSN	12 (2%)	52 (10%)	87 (17%)	137 (28%)	8 (5%)	28 (19%)	50 (34%)	71 (48%)
MetFragHD	55 (11%)	130 (26%)	180 (36%)	235 (47%)	13 (9%)	40 (27%)	60 (41%)	72 (49%)

for the positive mode spectra. The results in Table 1 were also visualized to gain an overall view of the candidate ranking improvement. While in some cases using only three of the four terms yielded similar ranking results, in the end, all four terms were retained as each contributes valuable information for the interpretation of the results. Furthermore, the MetFrag output is designed in such a way that users can access all individual scoring terms in the results export and are thus able to re-score the results (or exclude specific terms) at any stage using their own weighting scheme.

Observations on environmental sample

The same chromatographic methods (normal and HDX) were applied to an environmental sample to investigate how transferable these methods would be to “real world” samples. A well-characterized sample that was the focus of the joint EU project SOLUTIONS (<https://www.solutions-project.eu/>) was chosen (see “Methods”). Screenshots of the full scan chromatograms are given in the ESM (ESM Figs. S5 and S6, in positive and negative modes, respectively). The targeted analytical results performed on this sample [22] were used to confirm the results observed for the mixes (see ESM Table S5a). As an example, the MS/MS spectra for metformin are shown in Fig. 7 below, with the expected reaction and corresponding chromatographic peaks in ESM Fig. S7. For comparison, the corresponding normal and HDX spectra for metformin from the standard mixes (as opposed to the sample) are given in ESM Fig. S8; the spectral similarity between the HDX spectrum from the sample and the mix (without

performing any form of additional spectral processing or cleanup) was 0.87, mainly due to the presence of additional peaks in the sample spectra.

In total, 107 target compounds that were reported were deemed to be detectable with the non-target Orbitrap method used here (many at low concentrations, see ESM Table S5). Of these 107, 90 pairs of normal and HDX peaks were found (68 in positive mode, 22 in negative mode), excluding messy or unclear peaks. MS/MS pairs existed for 28 of these (21 positive, 7 negative). For the remaining pairs, either no MS/MS was observed in normal conditions (6), under HDX conditions (27), or both (46). This is partially influenced by the data-dependent acquisition used (i.e., no inclusion list was used to try to record MS/MS spectra for these compounds, which would be a realistic scenario for performing non-target analysis on a sample with unknown compounds). These results are summarized in ESM Table S5a. The average intensities (for peaks where pairs were observed) were 3.5×10^7 , 2.4×10^7 , 3.3×10^6 , and 1.3×10^6 for positive normal, positive HDX, negative normal, and negative HDX, respectively. The average retention time shift over both modes was 0.20 min.

As for the standard mixes, a significant loss in intensity was again observed for the negative mode HDX measurements (see ESM Fig. S6), except for substances occurring after the isocratic gradient at 13 min, which once again sharpened dramatically and substances eluted much earlier in HDX conditions. While the positive mode data appears visually similar (ESM Fig. S5), this is not the case for negative mode (ESM Fig. S6), where most of the visible peaks between 0.4 and 14 min in the normal chromatogram are no longer (or only

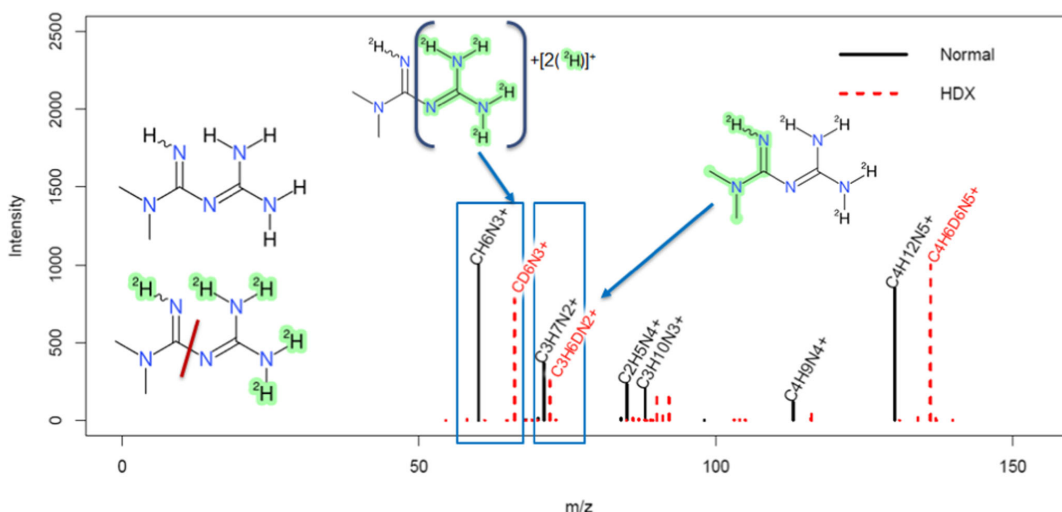


Fig. 7 Metformin (DTXSID2023270) in the Novi Sad sample; black in normal conditions and red dashed as observed under HDX conditions. The shift of the major fragments clearly shows the origins of the

fragments (see red line indicating the major “split” in the inset). Green highlighting in the fragments indicates the remaining backbone as represented in MetFrag

very slightly) visible in the HDX chromatogram, while the unresolved lump towards the end, due to dialkyl tetralin sulfonate (DATS, DTXSID70891725) surfactants, among others, has sharpened to a family of peaks between 14.5 and 16 min. The chromatography associated with individual masses in this homologous series is demonstrated in ESM Fig. S9. The corresponding fragmentation spectra in normal and HDX mode for C11-DATS ($C_{17}H_{26}O_3S$, precursor m/z 309.1530, identification level 3 [39]) is given as a head to tail plot in Fig. 8.

This retention time shift was also observed for the target compound perfluorooctanoic acid (DTXSID8031865), which was observed at RT = 15.5 min in normal mode and 13.7 min in HDX conditions. To investigate whether this is a phenomenon driven by the properties of these type of substances (a long apolar part followed by a polar head group), the sulfophenyl alkyl carboxylate (SPACs, DTXSID90891722) surfactants were also investigated, as these have polar functional groups on both ends of the molecule, due to the presence of the carboxyl group at the end of the alkyl chain. While these surfactants also suffered from the intensity loss in negative mode, they elute much earlier and did not appear to display large retention time shifts under HDX conditions (see ESM Fig. S10), although no MS/MS was obtained. Subsequently, surfactant series detected in wastewater [23], available here: https://comptox.epa.gov/dashboard/chemical_lists/eawagsurf, were screened by formula using RChemMass (<https://github.com/schymane/RChemMass>). Significant

shifts were observed for tentatively identified (level 3) groups of AS surfactants (RT 22–25 min to 14–15 min), DATS (RTs 21–24 min to 12–15 min), LAS (>24 min to 14–16 min). Less conclusive shifts, but clear sharpening of the elution profile in HDX mode, was observed for the AES and SAS classes, see ESM Table S5b.

Discussion

This article describes the integration of hydrogen-deuterium exchange (HDX) experiments into MetFrag to assist in the identification of unknown compounds in non-target high-resolution mass spectrometry experiments. The initial algorithms were implemented and tested on a small subset of stably labeled deuterated substances to ensure correct handling of deuterium. The full method was then applied to small test sets of hydrogen-deuterium exchange experiments before being evaluated extensively on a large set of environmental standards and finally applied to an environmental sample. Thus, the methods presented here have been validated on two separate LC-MS systems, one Orbitrap-based, and another QTOF-based. The experimental results were, in many ways, better than anticipated. For the standard mixes, very little deviation from the expected exchange behavior was observed and, despite intensity losses in negative mode observed for the Orbitrap data, generally very comparable MS/MS were

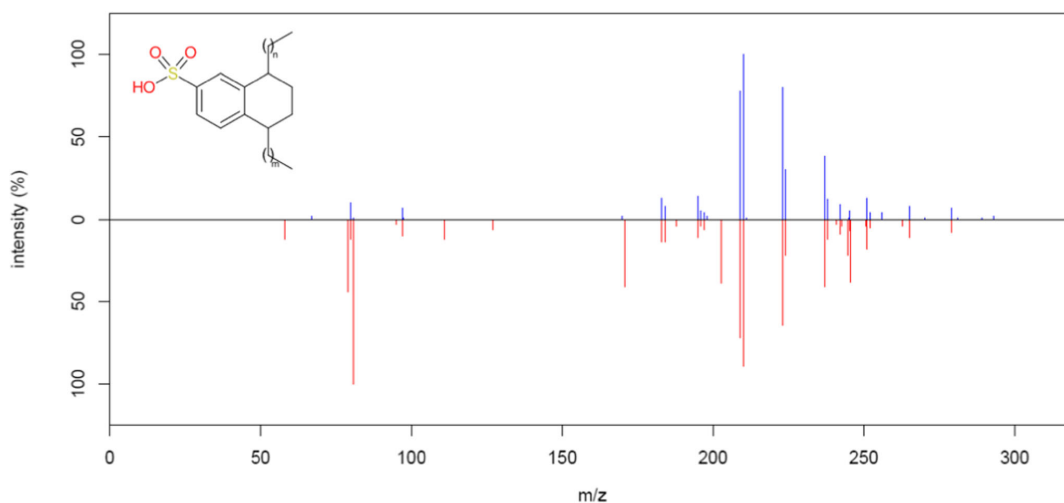


Fig. 8 Head to tail plot of MS/MS fragments from C11-DATS (where $m + n = 5$) in the Novi Sad sample. Blue: normal; red: HDX fragmentation. As only 1 D can be exchanged, which is lost during ionization, no D is observed in the structure of the ion. Shifts in the peaks in the lower masses are still observed due to the presence of D in the collision cell interacting with the aromatic structure, likely arising from other (deuterated) precursor ions included within the isolation window. Note

that the high-intensity precursor peaks (m/z 309.1530) have been excluded from both spectra to allow for better visualization of the fragmentation patterns. A lower intensity (~10%) precursor mass of m/z 308.6758 was observed in the full scan data for the HDX measurements, which would have been included in the isolation window for the HDX MS/MS data and could have been the source of deuterium. This mass was only visible at 2% in the MS/MS spectrum

obtained. However, despite this, the ranking improvements were not as great as hoped on the large set of ChemSpider candidates, with an increase from 10 to 16% of the candidates ranked correctly in first place. This contrasts with the influence of metadata on candidate ranking in MetFrag observed in the CASMI2016 results, which was run on a subset of 208 spectra from this same dataset, also using ChemSpider candidates [6]. In CASMI2016, MetFrag alone ranked 11% (24 of 208) correct in first place, compared with 78% (162 of 208) using MetFrag, retention time, and reference information [6] (where reference information was the largest contributor to the improvement in ranking [5]). This shows that metadata is still very much needed for rapid prioritization in high-throughput tentative identification for well-known substances. However, as discussed above, reference information is not always applicable, and in these cases, HDX experiments can provide additional information for candidate selection and has the clear advantage of being based on experimental information.

As demonstrated in this study (and also by previous studies utilizing this approach), HDX improves compound identification by narrowing down the number of potential candidates based on both MS1 and MS/MS data. The application with an LC system fully flushed with deuterated solvent is considerably more expensive than normal LC-HRMS, in our case about 15 vs 0.30 Euros per run for the solvent. Considering the overall cost of running non-target screening and the associated data evaluation, which may amount to many 100s of Euros, this extra cost can be considered acceptable for the additional information gained, as long as the instrument time and sample volume is available for the additional runs. In many cases, it is complementary to the MS/MS or retention time information typically used. With the integration into MetFrag, a semi-automated evaluation of data from HDX experiments is possible, while in previous studies, the data had to be evaluated and interpreted manually.

The way the data processing was performed in this study took advantage of the fact that the substance identity was “known,” which was critical for the method development. The expected HDX species were predicted and the corresponding data could thus be extracted easily. In true untargeted experiments, the “undeuterated” precursor masses in MS1 must be matched to the “deuterated” precursor masses without knowledge of the correct structure up front. This can be achieved by looking for a mass difference of $X \times (2.014102 - 1.007825) = 1.006277(X)$ units within a given retention time window, which could be determined using experiments on known standards. The number of deuteriums, X , can then be deduced from the mass difference and used in MetFrag to rank the candidates. As demonstrated in Fig. 5, the deuterated substance retention times can shift slightly and—in some cases—quite dramatically. The results presented here indicate that large retention time shifts will not be expected for rather fast gradient separations typically used in screening methods.

However, compounds eluting under isocratic conditions at low aqueous eluent fractions might be severely affected. Observations so far have occurred in a reproducible fashion over standard and sample measurements, such that some simple rules will help define appropriate retention time windows for these cases. Additional verification on different sample matrices and with further dual functionality standards would be needed to see exactly when the large retention time shifts are expected, for which substance classes and whether this effect varies in different sample matrices.

For a broader application to non-target screening, care must be taken that isotope peaks are not incorrectly assigned as potential deuterated masses in full scan data processing, as the mass difference between the ^{13}C isotope peak of the undeuterated species and a potential monodeuterated species is 0.00292 Da, which is, e.g., 7 ppm difference at m/z 400. In terms of MS/MS acquisition, a narrow isolation window (~ 1 Da) is essential, such that isotope peaks are not present in the fragmentation spectrum to confuse interpretation. In terms of full scan data processing, this will require high-quality peak grouping to correctly assign isotope peaks to features (componentization), in both the normal and deuterated experiments. For cases that behave as expected (e.g., 100% of H exchanged for D as expected), this should be relatively straightforward, as the isotope peaks will also be shifted by 100%. However, for cases of incomplete exchange, things can rapidly become more complicated. If only partial exchange occurs (e.g., 30%), then the $M+1$ peaks will be a mixture of $[\text{M}+\text{D}]^+$ and $^{13}\text{C}-[\text{M}+\text{H}]^+$, which requires a resolution $R = 35,000$ at $m/z = 100$, $R = 70,000$ at $m/z = 200$, etc. to resolve the isotopologues. It would be possible to resolve these peaks up to approximately $m/z = 400$ ($R = 140,000$) using the Orbitrap instrument applied in these experiments, but not generally with a QTOF. For molecules with a large number of exchangeable hydrogens and high mass (e.g., glycosides with several sugars), complex spectra will be obtained, and a low level of “normal” hydrogen in the deuterium-flooded LC systems becomes relevant (e.g., at 99% deuterium purity and 40 labile hydrogens, the probability that all these 40 hydrogens are exchanged is only 66%). Similar issues would be observed using post-column HDX, as these also yield mixed spectra, rather than the very clean spectra observed here. It is possible to do back-calculations to account for this (as is routinely done in proteomics experiments, for instance), but adds complications to the data interpretation and is beyond the scope of the current article. Additionally, future studies will need to investigate additional adducts, the combination of positive and negative ionization results to extract the molecular ion, as well as incomplete exchange.

In this manuscript, we have made use of the CompTox Chemicals Dashboard as a host for lists of chemical structures, both undeuterated and HDX versions. Each of these lists required manual registration of the chemical structures

(deuterated and undeuterated) into the underlying DSSTox database in order to be exposed via the Dashboard [26]. If the HDX approach proves to be of general value in analysis, the development of “HDX versions” of chemicals at registration may be possible, requiring the generation of deuterium-labeled forms of the chemicals to save as “related substances” by default. In many ways, this is similar to the generation of “MS-Ready” forms of the chemicals [40] that utilizes transformations of input chemicals to provide desalted, non-stereospecific forms to support mass spectrometry analyses. The generation of HDX forms of the chemicals could be done via the jar provided in the ESM or via the implementation of a set of transformation rules (e.g., D-exchange of OH, SH, NH, NH₂, etc.) to provide the HDX-related substance to support this type of analysis. Alternatively, a “HDX download file” could be provided of the predicted HDX forms of the entire CompTox database, if external users would find this useful.

Due to the methodological and experimental efforts, it is considered unlikely that HDX experiments will be applied to NTS of environmental samples on a regular basis (in contrast to stable isotope labelling in certain metabolomics experiments); however, in special cases, it may offer crucial help in identification. These cases include the screening for toxicologically relevant compounds such as amines or phenols where HDX can be expected to provide detailed structural information, as demonstrated in this study.

Funding information ELS is supported by the Luxembourg National Research Fund (FNR) for project 12341006. The QExactive Plus LC-HRMS used at UFZ is part of the major infrastructure initiative CITEPro (Chemicals in the Terrestrial Environment Profiler) funded by the Helmholtz Association. ELS, JH, CR, SN, and MK acknowledge funding by the SOLUTIONS project (grant agreement 603437), supported by the EU Seventh Framework Programme. CR was also supported by European Commission H2020 project PhenoMeNal Grant EC654241. SN acknowledges institutional funding by the Leibniz Association.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants and/or animals The authors declare that no human participants or animals were used in this study.

Disclaimer The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the US Environmental Protection Agency. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Frainay C, Schymanski E, Neumann S, Merlet B, Salek R, Jourdan F, et al. Mind the gap: mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites*. 2018;8:51. <https://doi.org/10.3390/metabo8030051>.
2. Blaženović I, Kind T, Ji J, Fiehn O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*. 2018;8:31. <https://doi.org/10.3390/metabo8020031>.
3. Freund DM, Hegeman AD. Recent advances in stable isotope-enabled mass spectrometry-based plant metabolomics. *Curr Opin Biotechnol*. 2017;43:41–8. <https://doi.org/10.1016/j.copbio.2016.08.002>.
4. Mahieu NG, Patti GJ. Systems-level annotation of a metabolomics data set reduces 25 000 features to fewer than 1000 unique metabolites. *Anal Chem*. 2017;89:10397–406. <https://doi.org/10.1021/acs.analchem.7b02380>.
5. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform*. 2016;8(1):3. <https://doi.org/10.1186/s13321-016-0115-9>.
6. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, et al. Critical assessment of small molecule identification 2016: automated methods. *J Cheminform*. 2017;9(1):22. <https://doi.org/10.1186/s13321-017-0207-1>.
7. Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. *J Cheminform*. 2017;9:32. <https://doi.org/10.1186/s13321-017-0219-x>.
8. Lam W, Ramanathan R. In electrospray ionization source hydrogen/deuterium exchange LC-MS and LC-MS/MS for characterization of metabolites. *J Am Soc Mass Spectrom*. 2002;13:345–53. [https://doi.org/10.1016/S1044-0305\(02\)00346-X](https://doi.org/10.1016/S1044-0305(02)00346-X).
9. Novak T, Helmy R, Santos I. Liquid chromatography–mass spectrometry using the hydrogen/deuterium exchange reaction as a tool for impurity identification in pharmaceutical process development. *J Chromatogr B*. 2005;825:161–8. <https://doi.org/10.1016/j.jchromb.2005.05.039>.
10. Muz M, Krauss M, Kutsarova S, Schulze T, Brack W. Mutagenicity in surface waters: synergistic effects of carboline alkaloids and aromatic amines. *Environ Sci Technol*. 2017;51:1830–9. <https://doi.org/10.1021/acs.est.6b05468>.
11. Acter T, Kim D, Ahmed A, Ha J-H, Kim S. Application of atmospheric pressure photoionization H/D-exchange mass spectrometry for speciation of sulfur-containing compounds. *J Am Soc Mass Spectrom*. 2017;28:1687–95. <https://doi.org/10.1007/s13361-017-1678-z>.
12. Ohashi N, Furuuchi S, Yoshikawa M. Usefulness of the hydrogen–deuterium exchange method in the study of drug metabolism using liquid chromatography–tandem mass spectrometry. *J Pharm Biomed*. 1998;18:325–34. [https://doi.org/10.1016/S0731-7085\(98\)00092-2](https://doi.org/10.1016/S0731-7085(98)00092-2).
13. Shah RP, Garg A, Puttur SP, Wagh S, Kumar V, Rao V, et al. Practical and economical implementation of online H/D exchange in LC-MS. *Anal Chem*. 2013;85:10904–12. <https://doi.org/10.1021/ac402339s>.
14. Kostyukevich Y, Acter T, Zherebker A, Ahmed A, Kim S, Nikolaev E. Hydrogen/deuterium exchange in mass spectrometry. *Mass Spectrom Rev*. 2018;37:811–53. <https://doi.org/10.1002/mas.21565>.
15. Ahmed A, Kim S. Atmospheric pressure photo ionization hydrogen/deuterium exchange mass spectrometry—a method to differentiate isomers by mass spectrometry. *J Am Soc Mass Spectrom*. 2013;24:1900–5. <https://doi.org/10.1007/s13361-013-0726-6>.
16. Zherebker A, Kostyukevich Y, Kononikhin A, Roznyatovsky VA, Popov I, Grishin YK, et al. High desolvation temperature facilitates the ESI-sourceH/D exchange at non-labile sites of hydroxybenzoic

5.3 Supporting non-target identification by adding hydrogen deuterium exchange MS/MS capabilities to MetFrag

- acids and aromatic amino acids. *Analyst*. 2016;141:2426–34. <https://doi.org/10.1039/C5AN02676H>.
17. Acter T, Cho Y, Kim S, Ahmed A, Kim B, Kim S. Optimization and application of APCI hydrogen–deuterium exchange mass spectrometry (HDX MS) for the speciation of nitrogen compounds. *J Am Soc Mass Spectrom*. 2015;26:1522–31. <https://doi.org/10.1007/s13361-015-1166-2>.
 18. Strehmel N, Böttcher C, Schmidt S, Scheel D. Profiling of secondary metabolites in root exudates of *Arabidopsis thaliana*. *Phytochemistry*. 2014;108:35–46. <https://doi.org/10.1016/j.phytochem.2014.10.003>.
 19. Ruttkies C, Strehmel N, Scheel D, Neumann S. Annotation of metabolites from gas chromatography/atmospheric pressure chemical ionization tandem mass spectrometry data using an *in silico* generated compound database and MetFrag: annotation of metabolites from high-resolution GC/APCI-MS/MS data. *Rapid Commun Mass Spectrom*. 2015;29:1521–9. <https://doi.org/10.1002/rcm.7244>.
 20. Brack W, Altenburger R, Schüttmann G, Krauss M, López Herráez D, van Gils J, et al. The SOLUTIONS project: challenges and responses for present and future emerging pollutants in land and water resources management. *Sci Total Environ*. 2015;503–504: 22–31. <https://doi.org/10.1016/j.scitotenv.2014.05.143>.
 21. Hashmi MAK, Escher BI, Krauss M, Teodorovic I, Brack W. Effect-directed analysis (EDA) of Danube River water sample receiving untreated municipal wastewater from Novi Sad, Serbia. *Sci Total Environ*. 2018;624:1072–81. <https://doi.org/10.1016/j.scitotenv.2017.12.187>.
 22. König M, Escher BI, Neale PA, Krauss M, Hilscherová K, Novák J, et al. Impact of untreated wastewater on a major European river evaluated with a combination of *in vitro* bioassays and chemical analysis. *Environ Pollut*. 2017;220:1220–30. <https://doi.org/10.1016/j.envpol.2016.11.011>.
 23. Schymanski EL, Singer HP, Longrée P, Loos M, Ruff M, Stravs MA, et al. Strategies to characterize polar organic contamination in wastewater: exploring the capability of high resolution mass spectrometry. *Environ Sci Technol*. 2014;48:1811–8. <https://doi.org/10.1021/es4044374>.
 24. NORMAN Network NORMAN suspect list exchange. In: NORMAN Suspect List Exchange. <https://www.norman-network.com/?q=node/236>. Accessed 13 Mar 2019.
 25. US Environmental Protection Agency. EAWAGSURF: Eawag surfactants list: surfactants screened in Swiss wastewater 2014. 2019. https://comptox.epa.gov/dashboard/chemical_lists/EAWAGSURF. Accessed 13 Mar 2019.
 26. Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *J Cheminform*. 2017;9:61. <https://doi.org/10.1186/s13321-017-0247-6>.
 27. Mayfield J CDK Depict Web Interface. <http://simolecule.com/cdkdepict/depict.html>. Accessed 30 Oct 2018.
 28. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 2012;30:918–20. <https://doi.org/10.1038/nbt.2377>.
 29. Stravs MA, Schymanski EL, Singer HP, Hollender J. Automatic recalibration and processing of tandem mass spectra using formula annotation: recalibration and processing of MS/MS spectra. *J Mass Spectrom*. 2013;48:89–99. <https://doi.org/10.1002/jms.3131>.
 30. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliakova N, et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminform*. 2017;9:33. <https://doi.org/10.1186/s13321-017-0220-4>.
 31. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, et al. SPLASH, a hashed identifier for mass spectra. *Nat Biotechnol*. 2016;34:1099–101. <https://doi.org/10.1038/nbt.3689>.
 32. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. *J Chem Inf Comput Sci*. 2003;43:493–500. <https://doi.org/10.1021/ci025584y>.
 33. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen E. Recent developments of the Chemistry Development Kit (CDK)—an open-source Java library for chemo- and bioinformatics. *Curr Pharm Des*. 2006;12:2111–20. <https://doi.org/10.2174/138161206777585274>.
 34. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform*. 2010;11:148. <https://doi.org/10.1186/1471-2105-11-148>.
 35. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem substance and compound databases. *Nucleic Acids Res*. 2016;44:D1202–13. <https://doi.org/10.1093/nar/gkv951>.
 36. Pence HE, Williams A. ChemSpider: an online chemical information resource. *J Chem Educ*. 2010;87:1123–4. <https://doi.org/10.1021/ed100697w>.
 37. Reed DR, Kass SR. Hydrogen–deuterium exchange at non-labile sites: a new reaction facet with broad implications for structural and dynamic determinations. *J Am Soc Mass Spectrom*. 2001;12:1163–8. [https://doi.org/10.1016/S1044-0305\(01\)00303-8](https://doi.org/10.1016/S1044-0305(01)00303-8).
 38. Kuck D. Scrambling versus specific processes in gaseous organic ions during mass spectrometric fragmentation: elucidation of mechanistic origins by isotope labelling – an overview. *J Label Compd Radiopharm*. 2007;50:360–5. <https://doi.org/10.1002/jlcr.1405>.
 39. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ Sci Technol*. 2014;48:2097–8. <https://doi.org/10.1021/es5002105>.
 40. McEachran AD, Mansouri K, Grulke C, Schymanski EL, Ruttkies C, Williams AJ. “MS-Ready” structures for non-targeted high-resolution mass spectrometry screening studies. *J Cheminform*. 2018;10:45. <https://doi.org/10.1186/s13321-018-0299-2>.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



working in informatics at OntoChem ID Solutions GmbH.

Christoph Ruttkies studied bioinformatics at Martin Luther University in Halle-Wittenberg and worked on the development of computational methods for the identification of metabolites based on tandem mass spectrometry data (i.e., MetFrag) during his PhD at the Leibniz Institute of Plant Biochemistry. He was also part of the European DevOps team in the H2020 project PhenoMeNal, working on a cloud-based metabolomics data analysis platform, and is now



Emma Schymanski is Associate Professor and Head of the Environmental Cheminformatics group at the Luxembourg Centre for Systems Biomedicine, University of Luxembourg, and a Luxembourg National Research Fund (FNR) ATTRACT Fellowship awardee. Her research combines open science, cheminformatics, and computational mass spectrometry approaches to elucidate the unknowns in complex samples and relate these to environmental

causes of disease, along with supporting several European and worldwide activities to improve the exchange of data, information, and ideas between scientists.



Steffen Neumann studied computer science and bioinformatics at Bielefeld University, and his group at the Leibniz Institute of Plant Biochemistry focuses on the development of tools and databases for metabolomics and computational mass spectrometry. They develop algorithms for data processing of metabolite profiling experiments (available in several Open Source Bioconductor packages), and address the identification of unknowns in mass spectrometry data with efforts in the

MassBank consortium and the MetFrag system, which allows the identification of compounds where no reference spectra are available.



Nadine Strehmel studied chemistry at the Technical University of Berlin, did her PhD thesis on metabolic biomarkers at the Max Planck Institute of Molecular Plant Physiology and her PostDoc study on root exudate metabolism at the Leibniz Institute of Plant Biochemistry, and currently heads the Mass Spectrometry Laboratory at the Governmental Institute of Legal Medicine and Forensic Sciences. She is very familiar with non-targeted metabolite profiling ex-

periments, in particular the identification of so-far unknown components from high-resolution mass spectrometry profiles.



Antony Williams is a computational chemist at the National Center of Computational Toxicology working on delivery of the center's data to the scientific community (via the CompTox Chemicals Dashboard at <https://comptox.epa.gov/dashboard>). An analytical scientist by training, he has over two decades of experience in cheminformatics and chemical information management and has worked extensively on complex data management issues with a focus

on internet-based projects to deliver free-access community-based chemistry websites and services (e.g., <http://www.chemspider.com>).



Juliane Hollender is Head of the Department of Environmental Chemistry at the Swiss Federal Institute of Aquatic Science and Technology (Eawag) as well as Adjunct Professor at the ETH Zurich in the Department of Environmental Systems Science. Her research concentrates on the fate of organic micropollutants in the natural and engineered aquatic environment; she is especially interested in biological transformation of contaminants in the environment, bioaccumulation in

aquatic organisms as well as non-target analysis using high-resolution mass spectrometry to obtain a more comprehensive picture of the contamination of aquatic systems.



Martin Krauss is a senior scientist at the Department Effect-Directed Analysis, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany. His research interests include developing and applying target and non-target screening methods for environmental micropollutant analysis and advancing HRMS-based approaches for structure elucidation.

5.4 Improving MetFrag with statistical learning of fragment annotations

Christoph Ruttkies, Steffen Neumann, Stefan Posch. Improving MetFrag with statistical learning of fragment annotations. BMC Bioinformatics. 20: 376, 2019. 20 Citations⁴

<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2954-7>

Contributions

I designed the study under supervision of Stefan Posch and Steffen Neumann. I implemented additional features and developed the scoring terms. I processed all datasets, created the R-based figures and prepared the manuscript.

⁴<https://scholar.google.com> (accessed on 01/2021)

RESEARCH ARTICLE

Open Access

Improving MetFrag with statistical learning of fragment annotations

Christoph Ruttkies^{1*} , Steffen Neumann^{1,2} and Stefan Posch³**Abstract**

Background: Molecule identification is a crucial step in metabolomics and environmental sciences. Besides in silico fragmentation, as performed by MetFrag, also machine learning and statistical methods evolved, showing an improvement in molecule annotation based on MS/MS data. In this work we present a new statistical scoring method where annotations of m/z fragment peaks to fragment-structures are learned in a training step. Based on a Bayesian model, two additional scoring terms are integrated into the new MetFrag2.4.5 and evaluated on the test data set of the CASMI 2016 contest.

Results: The results on the 87 MS/MS spectra from positive and negative mode show a substantial improvement of the results compared to submissions made by the former MetFrag approach. Top1 rankings increased from 5 to 21 and Top10 rankings from 39 to 55 both showing higher values than for CSI:OKR, the winner of the CASMI 2016 contest. For the negative mode spectra, MetFrag's statistical scoring outperforms all other participants which submitted results for this type of spectra.

Conclusions: This study shows how statistical learning can improve molecular structure identification based on MS/MS data compared on the same method using combinatorial in silico fragmentation only. MetFrag2.4.5 shows especially in negative mode a better performance compared to the other participating approaches.

Keywords: Mass spectrometry, Statistical modeling, Identification

Background

The identification of small molecules such as metabolites is a crucial step in metabolomics and environmental sciences. The analytical tool of choice to achieve this goal is mass spectrometry (MS) where ionized molecules can be differentiated by their mass-to-charge (m/z) ratio. As a single m/z value is not sufficient for the unequivocal determination of the molecular structure, tandem mass spectrometry (MS/MS) is applied, which results in the formation of fragment ions of the entire molecule. These fragments result in fragment peaks that are characterized by their m/z and intensity value. The intensity correlates with the amount of ions detected with that particular m/z value. These m/z fragment peaks can be used to infer additional hints about the underlying molecular structure.

The interpretation of the generated data is complex and usually requires expert knowledge. Over the past years, several software tools have been developed to overcome the time-consuming manual analysis of the growing amount of MS/MS spectra in an automated way. The first approaches tried to reconstruct observed fragment spectra by performing in silico fragmentation in either a rule based (e.g. MassFrontier [1]) or combinatorial manner such as MetFrag [2, 3], MIDAS [4], MS-Finder [5] and MAGMa [6].

MetFrag was one of the first combinatorial approaches developed and performs in silico fragmentation of molecular structures. Given a single MS/MS spectrum of an unknown molecule, MetFrag first selects molecular candidates from databases given the neutral mass of the parent ion. In the next step, each of the retrieved candidates is treated individually and fragmented in silico using a bond-disconnection approach. The generated fragment-structures are assigned to the m/z fragment peaks of the

*Correspondence: christoph.ruttkies@ipb-halle.de

¹Department Biochemistry of Plant Interactions, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany
Full list of author information is available at the end of the article



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

MS/MS spectrum, based on the comparison of the theoretical mass of the generated structure and the m/z value of the acquired fragment peak. Given a set of assignments of m/z fragment peaks to fragment-structures for one candidate, MetFrag calculates a score that indicates how well the candidate matches the given MS/MS spectrum. These scores are used to rank all retrieved candidates. Ideally, the correct one is ranked in first place.

Statistical approaches have evolved, which are learning fragmentation processes on the basis of annotated experimental MS/MS data. CFM-ID [7] is using Markov-chains to model transitions of fragment-structures for the prediction of MS/MS spectra. Generated spectra can be aligned with the spectrum of interest and report the candidates with the best matching spectral prediction. FingerID [8] uses MS/MS spectra to predict molecular fingerprints. These Fingerprints are bit-wise representations of molecular structures where each position in the fingerprint encodes a structural property of the underlying molecule. FingerID uses support vector machines (SVM) and is enhanced by CSI:FingerID (CSI:FID) [9], integrating fragmentation trees which are calculated by SIRIUS [10]. CSI:IOKR [11] replaces the SVM prediction by an input-output kernel regression approach. Recent analysis in one of the latest CASMI (Critical Assessment of Small Molecule Identification) contests (2016) [12] reveal that techniques supported by statistical learning (i.e. CSI:FID and CSI:IOKR) are the most promising and powerful methods used to perform structure elucidation if only the MS/MS data is considered.

In this work we introduce a new statistical approach to evaluate candidates for MS/MS spectra. Using training data, probabilities of the predicted fragment-structures given the observed m/z peaks are estimated with a Bayesian approach. These probabilities are integrated as new scoring terms for MetFrag to rank candidates. The new scoring schema is tested on the challenge data sets of the CASMI contest 2016. The method shown here complements the different machine learning and statistical approaches that perform MS/MS spectra prediction (CFM-ID), prediction of molecular fingerprints (CSI:FID, CSI:IOKR) and now combining in silico fragmentation and statistical scoring for the evaluation of retrieved molecular candidates. The new scoring functions are available with the new MetFrag version 2.4.5.

Methods

This section introduces the notation and the Bayesian model approach used to evaluate how likely a fragment-structure is in the presence of an m/z fragment peak. The resulting probabilities are defined across the domain of all possible fragment-structures and all m/z fragment peaks, but can be reduced to become tractable. The resulting probability distribution will be used in the candidate

score $S_{RawPeak}^C$ indicating whether a candidate can explain the m/z fragment peaks with fragment-structures seen in the training spectra. In analogy, neutral losses will also be considered. The parameter estimation to model the probability distribution is at the heart of our approach. We describe how they are estimated from training data, taking care to clearly separate training data from evaluation data. Finally we describe the evaluation using the CASMI 2016 challenge data and comparison to the results obtained by other approaches and state-of-the art small molecule identification programs.

First, we introduce notations required for our approach. A summary of the notation used in the following and their description can be found in Additional files 4 and 5: Tables S1 and S2. Consider a set of N centroided MS/MS spectra $\underline{m} = \{\underline{m}_n | n = 1, \dots, N\}$ where $\underline{m}_n = (m_{n1}, \dots, m_{nK_n})$ consists of K_n m/z fragment peaks m_{nk} . Furthermore, for each spectrum \underline{m}_n a set of candidates \underline{c}_n of length C_n is given, typically retrieved from a database. For a given candidate $c_{nc} \in \underline{c}_n$, MetFrag performs an in silico fragmentation and assigns each observed m/z fragment peak m_{nk} to one of the generated fragment-structures, denoted f_{nck} in the following. This can be interpreted as explaining the m/z fragment peak m_{nk} with the fragment-structure f_{nck} . On the basis of the in silico fragmentation, assignments of m/z fragment peaks to fragment-structures $(\underline{m}_n, \underline{f}_{nc}), c = 1, \dots, C_n$, are determined. As there is not necessarily a matching fragment-structure for every m/z fragment peak m_{nk} , we introduce \perp in case an m/z fragment peak m_{nk} cannot be annotated, and denote $f_{nck} = \perp$ in this case.

As stated in the introduction, we want to evaluate candidates for an MS/MS spectrum by a statistical scoring approach to be integrated into MetFrag. Therefore, we apply a scoring term based on the probability $P(\underline{f}_{nc} | \underline{m}_n)$. The distribution $P(\underline{f} | \underline{m})$ models the occurrence of fragment-structures in \underline{f} in the correct candidate for a given list \underline{m} of m/z fragment peaks in an observed spectrum. In the following we assume the independence of the assignments of m/z fragment peaks to fragment-structures yielding

$$P(\underline{f} | \underline{m}) = \prod_{k=1}^K P(f_k | m_k),$$

with $\underline{m} = (m_1, \dots, m_K)$ and $\underline{f} = (f_1, \dots, f_K)$. From a chemical point of view, we know that certain m/z fragment peaks occur concurrently with other m/z fragment peaks (or at least with a higher certainty) due to multi-stage fragmentation pathways that lead to a further fragmentation of a generated fragment-structure. However, for the sake of model simplification we do not consider this information when assuming independence of assignments of m/z fragment peaks to fragment-structures.

A fragment-structure can be regarded as a connected charged molecular structure consisting of atoms connected via bonds. A graph can be used as data structure to represent a fragment-structure, as atoms and bonds can be represented by graph nodes and edges, respectively. However, to reduce the computational costs for comparing graphs by determining graph isomorphisms, especially when working with thousands or even hundreds of thousands of fragment-structures, we use molecular fingerprints as a bit-string representation of a molecular structure. Each bit of the fingerprint describes the presence or absence of a molecular feature within the structure. As different fragment-structures may share the same fingerprint, this approach reduces the domain size and also generalizes very similar fragment-structures that would explain the same m/z fragment peak. There are different molecular fingerprint functions available, e.g., the MACCSFingerPrint [13] and the LingoFingerprint [14]. A fragment-structure fingerprint is defined as $f_k = \text{MolFing}(f_k)$, calculated by the fingerprint function *MolFing*.

We regard two fragment-structures f and f' to be equal, if \tilde{f} and \tilde{f}' are equal, although f and f' might be structurally different. This reduces the comparison to constant time as the fingerprint length is independent of the size of the fragment-structure. The distribution can now be re-defined as

$$P(\tilde{f}|m) = \prod_{k=1}^K P(\tilde{f}_k|m_k).$$

The comparison of two m/z fragment peaks m and m' can not be performed as a simple test for equality by $m = m'$. This is impractical for MS measurements as they show a certain degree of deviation depending on the mass accuracy of the instrument. For this reason, the m/z range covered by training and test spectra is discretized into non-equidistant bins $[b_i, b_{i+1}]$. The boundaries are calculated as $b_{i+1} = b_i + 2 \cdot (mzppm(b_i) + mzabs)$ with b_0 set to the minimum mass value of this range. The values $mzabs$ and $mzppm(b_i)$ represent the absolute (in m/z) and relative mass (in ppm) deviation given by the MS setup.

Two m/z fragment peaks m and m' are considered to be equal if they fall into the same bin. In the following each m/z fragment peak m is discretized to the central value of its bin.

Domains and Parameters

As a next step, the two domains M of m/z values m and F of all fragment-structure fingerprints \tilde{f} need to be defined. For M one could consider all bins resulting from discretization. However, this is impractical as the major part

of this domain is not observed for a given data set. Likewise, the domain F can be defined to contain all possible fragment-structure fingerprints. Using the MACCSFingerprint with 166 bits would result in $2^{166} \approx 9.35 \cdot 10^{49}$ different fingerprints. In practice this space needs to be reduced to be tractable, and again only a fraction will be observed for a given problem. For a spectral training data set of N MS/MS spectra and C_n candidates each, we define a reduced peak domain \tilde{M}_{tr} and a reduced fingerprint domain \tilde{F}_{tr} as

$$\begin{aligned} \tilde{M}_{tr} &= \{m_{nk} | n \in 1, \dots, N, k = 1, \dots, K_n\} \subseteq M \\ \tilde{F}_{tr} &= \{\tilde{f}_{nck} | n \in 1, \dots, N, c = 1, \dots, C_n, k = 1, \dots, K_n\} \subseteq F, \end{aligned}$$

which are the m/z fragment peaks and fragment-structure fingerprints observed in this data set.

Furthermore, we define \mathcal{D}_{train} as a list of all assignments of m/z fragment peaks to fragment-structures in the training data, i.e.

$$\mathcal{D}_{train} = ((m_{nk}, f_{nck}) | n = 1, \dots, N, c = 1, \dots, C_n, k = 1, \dots, K_n).$$

Besides the MS/MS spectra given in this training data set we also need to address observations of an additional centroided MS/MS query spectrum m_q that is not part of the training data set. The processing of m_q is illustrated in Fig. 1. The domains are extended by the observations retrieved from this single query spectrum with C_q candidates and K_q m/z fragment peaks, i.e.

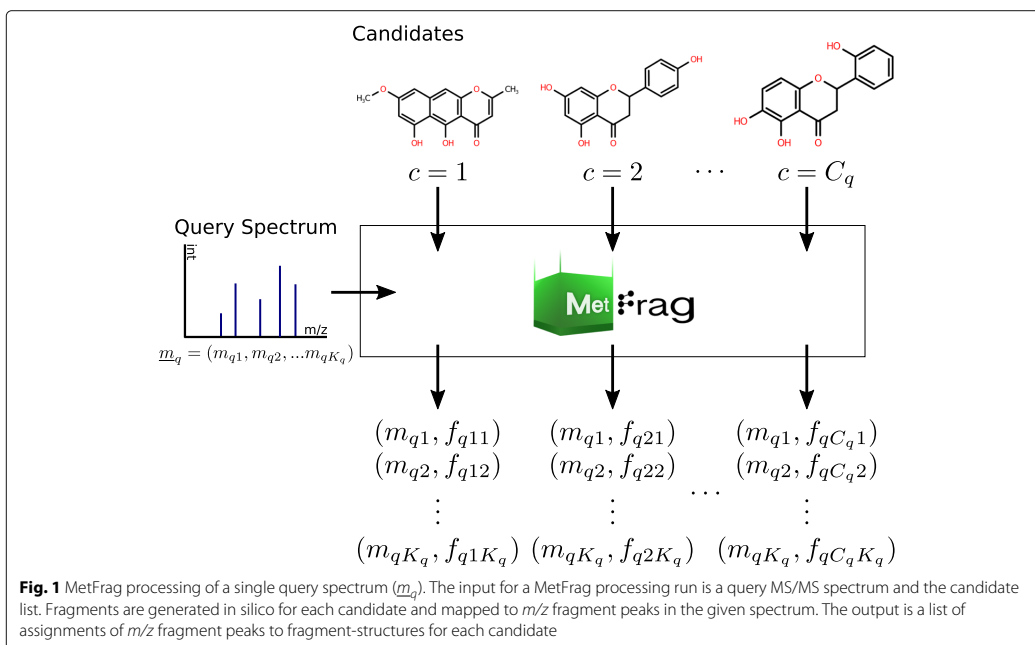
$$\begin{aligned} \tilde{M} &= \tilde{M}_{tr} \cup \{m_{qk} | k = 1, \dots, K_q\} \\ \tilde{F} &= \tilde{F}_{tr} \cup \{\tilde{f}_{qck} | c = 1, \dots, C_q, k = 1, \dots, K_q\}. \end{aligned}$$

To define the distribution $P(\tilde{f}|m)$ with $m \in \tilde{M}$ and $\tilde{f} \in \tilde{F}$, we introduce the notation $\theta_{m\tilde{f}} := P(\tilde{f}|m)$, which is the probability of fragment-structure fingerprint \tilde{f} given an observed mass m . The complete set of parameters is given as

$$\underline{\theta} = (\theta_{m\tilde{f}}), \quad \text{for } m \in \tilde{M}, \tilde{f} \in \tilde{F}.$$

Parameter estimation

The parameters are initially not known and need to be estimated from the training data. In the process of parameter estimation \mathcal{C}_n is set to only contain the known correct candidate ($C_n = 1$) for the generation of \mathcal{D}_{train} as this results in mainly correct predicted fragment-structure assignments as ground truth. The generation



of \mathcal{D}_{train} is illustrated in Fig. 2 where only the correct candidate for each spectrum is processed. One paradigm for parameter estimation is the maximum likelihood principle

$$\hat{\theta}^{ML} = \underset{\theta}{\operatorname{argmax}} P(\mathcal{D}_{train}|\theta),$$

which results in

$$\hat{\theta}_{m\tilde{f}}^{ML} = \frac{N_{m\tilde{f}}}{\sum_{\tilde{f}' \in \tilde{F}} N_{m\tilde{f}'}} ,$$

with $N_{m\tilde{f}} = \sum_{(m, \tilde{f}) \in \mathcal{D}_{train}} \delta(\tilde{f}_t, \tilde{f}) \delta(m_t, m)$

$N_{m\tilde{f}}$ is the absolute frequency of the assignments of m/z fragment peaks to fragment-structures (m, \tilde{f}) in the training data set \mathcal{D}_{train} .

If such an assignment (m, \tilde{f}) resulting from the query spectrum is not contained in the training data, a probability $\hat{\theta}_{m\tilde{f}}^{ML} = 0$ is estimated. As a consequence the probability $P(\tilde{f}|\underline{m})$ for the query will be zero.

Due to the limitation of the available training data, this situation will arise quite often. To avoid this problem, we use the Bayes paradigm including a priori distribution for the parameters to be estimated. In addition, as we only consider the correct candidate for each spectrum in \mathcal{D}_{train} it is not possible to reliably estimate parameters in case $\tilde{f} = \perp$, which is the probability for an m/z fragment peak without an assigned fragment-structure. Within the Bayesian approach we model this probability with the prior distribution and set $N_{m\perp} = 0$.

In the following we will use the mean posterior (MP) principle

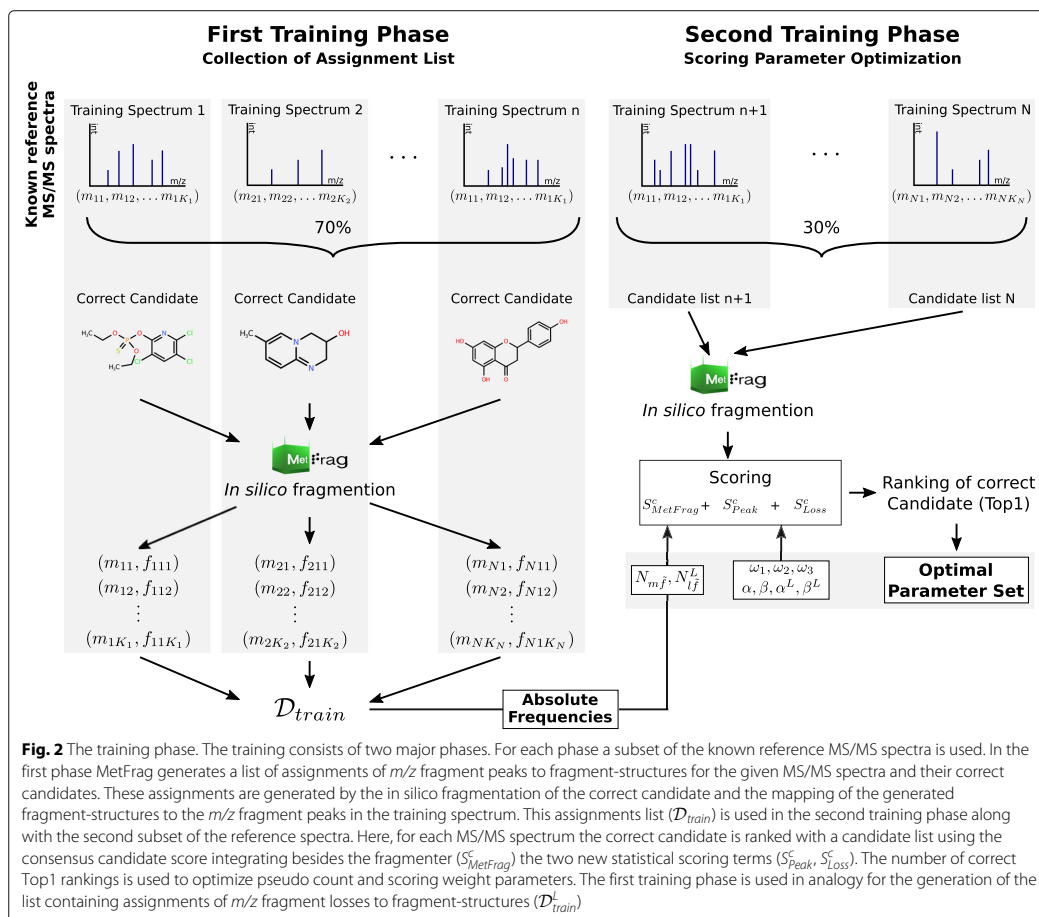
$$\hat{\theta}_{m\tilde{f}}^{MP} = E_{P(\theta|\mathcal{D}_{train}, \pi)}[\theta]$$

where

$$P(\theta|\mathcal{D}_{train}, \pi) = \frac{P(\theta|\pi)P(\mathcal{D}_{train}|\theta)}{P(\mathcal{D}_{train}|\pi)}$$

is the a posteriori distribution of parameters θ . As a prior distribution $P(\theta|\pi)$ on the parameters we use a product Dirichlet distribution with hyper parameters $\pi_{m\tilde{f}}$, $m \in \tilde{M}, \tilde{f} \in \tilde{F}$ defined as

$$\pi_{m\tilde{f}} = \begin{cases} \alpha, & \tilde{f} \neq \perp \\ \beta, & \tilde{f} = \perp \end{cases}$$



where α and β are also called pseudo counts. The parameter estimation is given by

$$\hat{\theta}_{mf}^{MP} = \frac{N_{mf} \tilde{\pi} + \pi_{mf} \tilde{\pi}}{\sum_{f' \in \tilde{F}} (N_{mf'} \tilde{\pi} + \pi_{mf'})}$$

Fragment losses

Fragment losses can provide additional evidence for a molecular structure as the difference between two m/z fragment peaks provides hints about a substructure that was lost but not observed directly by an m/z fragment peak (neutral loss). However, we want to include this information in the evaluation of candidates for a given MS/MS spectrum. We define l_{nkh} to be the m/z fragment

loss between two different m/z fragment peaks m_{nk} and m_{nh} from the spectrum \underline{m}_n , where

$$l_{nkh} = m_{nk} - m_{nh}, \quad m_{nk} > m_{nh}.$$

For each pair of assignments of m/z fragment peaks to fragment-structures (m_{nk}, f_{nck}) and (m_{nh}, f_{nch}) with f_{nch} being a genuine substructure of f_{nck} ($f_{nck} \neq f_{nch}$), we introduce f_{nckh} as a loss fragment-structure. This fragment-structure is a substructure of f_{nck} , that is generated if all bonds and atoms present in f_{nch} are removed ($f_{nckh} = f_{nck} \setminus f_{nch}$). If f_{nckh} is connected, we define (l_{nkh}, f_{nckh}) to be an assignment of an m/z fragment loss to a fragment-structure.

In analogy to the pairs of m/z fragment peaks and fragment-structures (m_{nk}, f_{nck}) , we define the domains for

the m/z fragment losses and loss fragment-structures for the N MS/MS training spectra as

$$\begin{aligned}\tilde{L}_{tr} &= \{l_{nkh} | n \in 1, \dots, N, k = 1, \dots, K_n, h = 1, \dots, K_n\} \\ \tilde{F}_{tr}^L &= \left\{ \tilde{f}_{nckh} | n \in 1, \dots, N, c = 1, \dots, C_n, \right. \\ &\quad \left. k = 1, \dots, K_n, h = 1, \dots, K_n \right\}\end{aligned}$$

for a given training data set

$$\begin{aligned}\mathcal{D}_{train}^L &= ((l_{nkh}, f_{nckh}) | n = 1, \dots, N, c = 1, \dots, C_n, \\ &\quad k = 1, \dots, K_n, h = 1, \dots, K_n)\end{aligned}$$

of assignments of m/z fragment losses to fragment-structures.

In addition, both domains need to be extended for the additional query MS/MS spectrum \underline{m}_q

$$\begin{aligned}\tilde{L} &= \tilde{L}_{tr} \cup \{l_{qkh} | k = 1, \dots, K_q, h = 1, \dots, K_q\}, \\ \tilde{F}^L &= \tilde{F}_{tr}^L \cup \{\tilde{f}_{qckh} | c = 1, \dots, C_q, k = 1, \dots, K_q, h = 1, \dots, K_q\}.\end{aligned}$$

We consider the distribution $P(\tilde{f} | \tilde{L})$ for assignments of fragment-structures to m/z fragment losses with $l \in \tilde{L}$ and $\tilde{f} \in \tilde{F}^L$, and denote $\phi_{l\tilde{f}}^L := P(\tilde{f} | l)$. In analogy to the estimation of the parameters θ_{mf} , we can now formulate the estimation of $\phi_{l\tilde{f}}^L$ including a Dirichlet a priori distribution with the additional hyper parameters $\psi_{\tilde{f}}$:

$$\psi_{l\tilde{f}} = \begin{cases} \alpha^L, & \tilde{f} \neq \perp \\ \beta^L, & \tilde{f} = \perp \end{cases}$$

This yields the mean posterior estimates

$$\begin{aligned}\hat{\phi}_{l\tilde{f}}^{MP} &= \frac{N_{l\tilde{f}}^L + \psi_{\tilde{f}}}{\sum_{\tilde{f}' \in \tilde{F}^L} (N_{l\tilde{f}'}^L + \psi_{\tilde{f}'})}, \\ \text{with } N_{l\tilde{f}}^L &= \sum_{(l, \tilde{f}) \in \mathcal{D}_{train}^L} \delta(\tilde{f}_t, \tilde{f}) \delta(l_t, l)\end{aligned}$$

analogous to the parameter estimation for the assignments of m/z fragment peaks to fragment-structures, where $N_{l\tilde{f}}^L$ is the absolute frequency of the m/z fragment loss and fragment-structure pair (l, \tilde{f}) observed in the training data set \mathcal{D}_{train}^L .

Evaluation of the assignments of fragment-structures to m/z fragment peaks and losses in MetFrag candidate scoring

To evaluate a given candidate c retrieved from a compound database for an MS/MS query spectrum \underline{m}_q based on the statistical models, we define a score for both the models of the assignments of m/z fragment peaks/losses to fragment-structures. In addition, the MetFrag fragmenter score $S_{MetFrag}^c$ as defined in [3] is also integrated in this candidate evaluation. We define the score S_{Fin}^c as

the final or consensus score for a candidate c to be the weighted sum of these three scoring terms

$$\begin{aligned}S_{Fin}^c &= \omega_1 \cdot S_{MetFrag}^c + \omega_2 \cdot S_{Peak}^c + \omega_3 \cdot S_{Loss}^c \\ \omega_i &\geq 0, \quad \sum_{i=1,2,3} \omega_i = 1.\end{aligned}$$

To define S_{Peak}^c and S_{Loss}^c , we first introduce the raw score of a candidate as

$$S_{RawPeak}^c = \frac{1}{-\log P(\tilde{f}_{nc} | \underline{m}_n, \hat{\theta}^{MP})}$$

using the log likelihood based on the estimated parameters $\hat{\theta}^{MP}$ for the assignment of an m/z fragment peak to a fragment-structure $(\underline{m}_n, \tilde{f}_{nc})$ for candidate c . With $\tilde{f}_{nc} = (\tilde{f}_{nc1}, \dots, \tilde{f}_{ncK_n})$ and $\underline{m}_n = (m_{n1}, \dots, m_{nK_n})$ the log likelihood decomposes as

$$\log P(\tilde{f}_{nc} | \underline{m}_n, \hat{\theta}^{MP}) = \sum_{k=1}^{K_n} \log P(\tilde{f}_{nck} | m_{nk}, \hat{\theta}^{MP}).$$

Furthermore, the raw score is normalized to the interval $[0, 1]$ by

$$S_{Peak}^c = \frac{S_{RawPeak}^c}{\max_{c' \in C_q} S_{RawPeak}^{c'}}.$$

Using identical ranges for the different scoring terms as for the MetFrag fragmenter score simplifies their integration into the weighted sum of the final score. The score for including the assignments of m/z fragment losses to fragment-structures S_{Loss}^c is defined in analogy.

Method evaluation

For the evaluation of the presented approach we used the challenge data set and evaluation procedures of the CASMI 2016 contest. In this contest candidate lists were provided by the organizers along with the spectra to be used by all participants. After the contest, several participants which used statistical learning (e.g. CSI:FID, CSI:IOKR, CFM-ID) coordinated which compounds were used in the training steps to improve the comparability between methods. They exchanged the InChIKeys (InChI: International Chemical Identifier) [15] of the spectra used in training their approaches, although it was not guaranteed that two participants used exactly the same MS/MS spectrum for a compound identified by a common InChIKey if they used different spectral databases. This evaluation is based on 87 of the 208 spectra provided originally in the challenge, as the remaining 121 spectra were removed as they were included in the training data of at least one participant. The results for this subset of the challenge spectra were published in [12] and used here in Table 2 for comparison against MetFrag2.4.5. We used the same set of InChIKeys to obtain the training spectra for

this paper. The training data is available from the github repository accompanying the paper.

Preparation of the training data set

The training data set includes MS/MS spectra provided by the contest organizers consisting of 312 CASMI training spectra. Participants were allowed to use additional training spectra retrieved from spectral databases e.g. the MassBank of North America (MoNA) [16] and the Global Natural Products Social Molecular Networking (GNPS) [17] spectral library. The InChIKeys of the molecules of these additional spectra were provided by the participants.

We used the provided InChIKeys to retrieve the additional training spectra by querying the MoNA and GNPS spectral databases. For MoNA, retrieved MS/MS spectra from one institution were merged in case more than one spectrum was present for a molecule based on the first block the InChIKey. Thus for one InChIKey several merged spectra can be present in case they originate from different sources. Spectra originating from GNPS spectral database were merged independently of their source. The spectra merging was performed by averaging m/z fragment peaks within a specified mass range (given by MS setup of the MS/MS spectra) and retaining the peak of maximum intensity. This resulted in 5622 spectra (4728 positive and 884 negative) which were used for training. To reduce the spectral complexity only the 40 most abundant (based on intensity) m/z peaks in each spectrum were used. The same applies to test spectra used for evaluation.

Training of parameters

In the training phase the optimal parameters used to calculate the candidates' consensus score need to be determined. This parameter set consists of the absolute frequencies N_{mf} and N_{lf}^L of the assignments of m/z fragment peaks and losses to fragment-structures, the hyper parameters α , β , α^L and β^L , and the score weights ω_1 , ω_2 and ω_3 . The whole training phase described in this paragraph is illustrated in Fig. 2.

Training was separated into two phases where in the first phase the N_{mf} and N_{lf}^L parameters were determined using only the correct candidate for each training spectrum. Based on these absolute frequencies the optimal hyper parameters and weight scores are determined in the second phase.

If we had used the same data set for the estimation of all parameters, \mathcal{D}_{train} and \mathcal{D}_{train}^L would have contained the same pairs of m/z fragment peaks/losses and fragment-structures for the correct candidate to be ranked in the second phase. The correct candidate would then be favoured during candidate ranking. This is not representing a realistic case when a query spectrum of an

unobserved molecule is processed where we expect also m/z fragment peak and loss assignments not previously observed in the optimization phase.

For this reason the complete training data set was split randomly into two disjunct groups of spectra. The splitting was performed by dividing the unique list of InChIKeys (first block) with a ratio of 70:30 and collecting each corresponding spectrum to a group based on the InChIKey of the underlying molecule. The larger group is used in the first phase to calculate the N_{mf} and N_{lf}^L .

In the first phase the correct candidate of each spectrum was processed by MetFrag's in silico fragmentation. The m/z fragment peaks explained by a fragment-structure were corrected to the mass of the molecular formula of the assigned fragment-structure. This is required to be independent of the different mass accuracies of MS/MS spectra acquired under different instrument conditions. Thus the list of assignments of m/z fragment peaks/losses to fragment-structures \mathcal{D}_{train} and \mathcal{D}_{train}^L contained assignments with the corrected m/z values used for the calculation of N_{mf} and N_{lf}^L .

In the second training phase candidates were retrieved from a local PubChem [18] mirror (June 2016) using the monoisotopic mass of the correct candidate of each spectrum and a relative mass deviation dependent on the experimental conditions of the underlying MS measurement. To reduce runtime the correct and at most 500 randomly sampled candidates were processed from the retrieved list of candidates. The rank of the correct candidate was determined and the overall number of Top1 ranks was used as optimization criterion.

For the hyper parameters the optimization was performed by a grid search over an initial domain including a set of all combinations of the values 0.0025, 0.0005 and 0.0001 resulting in a total of $3^4 = 81$ sets of hyper parameters. If the optimal number of Top1 ranks was located at the border of this hyper parameter domain the search space was extended by increasing or decreasing the parameter by a factor of 5 or 1/5 respectively. This procedure was continued until an optimum was found with an improvement of less than 1% compared to the previous optimum of Top1 ranks. For the score weights a set of 1000 parameter combinations was sampled equally distributed on the simplex. Consensus scores and the rankings of the correct candidates were calculated for all combinations of hyper parameters and weights resulting in initially 81.000 combinations.

Subsequent to this training procedure, the absolute frequencies N_{mf} and N_{lf}^L were recalculated using the entire training data set to increase the observation domain of assignments of m/z fragment peaks/losses to fragment-structures used for the processing of the challenge data set.

Fingerprint function

To investigate the effect of the fingerprint function *MolFingerprint* on the results, the complete training phase was performed four times with different fingerprint functions for the same training spectra. For comparison the Lingo- [14], the MACCS- [13], the Circular- [19], and the GraphOnlyFingerprint were used. For calculation of the different fingerprints CDK (version 2.1) [20] implementations were used. The fingerprint with the best training result was selected for the processing of the challenge data set.

Processing of the CASMI challenge data set

After the training phase and the selection of the fingerprint function, the *in silico* fragmentation and scoring was performed for the 87 challenge spectra using the provided candidate lists. Candidates that included non-connected substructures or non-natural isotopes (like deuterium) were discarded from the candidate lists. The candidate ranking was performed after the removal of multiple stereoisomers in compliance with the contest rules and evaluation. Stereoisomers were detected based on the first block of the candidates' InChIKey representing the molecular skeleton and only the best scoring stereoisomer was regarded for candidate ranking. The results were evaluated and compared on the basis of the average Top1, Top3, and Top10 rankings and the median and mean average rankings of the correct candidate as in [12].

Stability of parameter optima and ranking results

Splitting of the training data set for the two phases was performed randomly. As the resulting parameters depend

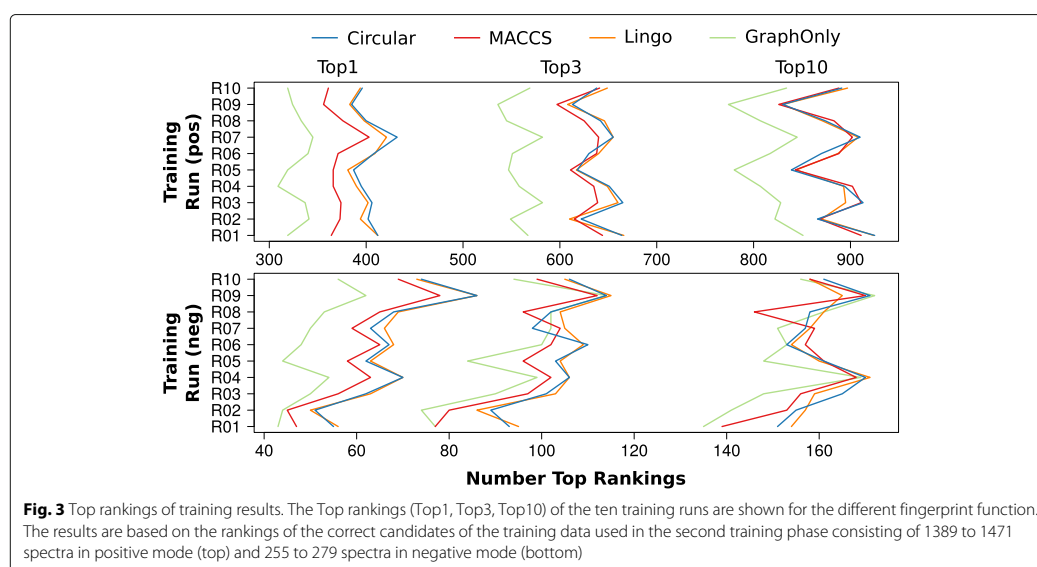
on the splitting, we performed ten independent trials with different splits of the training data. The resulting parameters and their performance on the challenge data set were reported to investigate the effect of randomization.

Results

Comparison of different fingerprint functions

The ranking results obtained in the training phase on the basis of the different fingerprint functions (*MolFingerprint*) are shown in Fig. 3. The fingerprints used are the Lingo-, MACCS-, Circular-, and GraphOnlyFingerprint. The training results are based on the spectra processed in the second phase during training consisting of 1389 to 1471 spectra in positive and 255 to 279 spectra in negative mode depending on the run and the spectra splitting.

Comparable results are obtained with the Circular- and LingoFingerprint across both ion modes and across the different rankings as shown in Fig. 3 by the similar curve for the Top1, Top3 and Top10 rankings. Similar means of the rankings across the ten runs confirm this observation with 402.3, 639.8, and 881.2 for the mean Top1, Top3 and Top10 rankings using the Circular- and 398.4, 640.0 and 881.9 using the LingoFingerprint. These two fingerprint functions show superior results for the Top1 rankings compared to MACCS with 371.0 and GraphOnly 328.6. For Top3 and Top10 rankings and positive mode the MACCSFingerprint gives comparable results. Top3 and Top10 rankings in negative mode are comparable for all fingerprint functions.



The CircularFingerprint shows with the runs R07 in positive and R09 in negative mode the overall highest number of Top1 rankings with 518 of the 1686 training spectra. Due to this performance the CircularFingerprint is used for subsequent investigations and the evaluation of the challenge data set.

Randomization of training data sets

In this section we evaluate the impact of the randomization of the training data on parameter optimization. Table 1 shows the optimal parameter sets and the performance achieved on the training data using the CircularFingerprint. The overall ranking results vary across the ten runs for the Top1, Top3 and Top10 numbers in both positive and negative ion mode as expected. Boxplots of the parameter sets are shown in Fig. 4. The variation of optimal hyper parameters as well as weights shows a similar pattern for both positive and negative ion mode where a larger variation can be observed in negative mode. Particularly the pseudo counts for annotated m/z fragment peaks show a broader variation with $5e-04$ to $2e-05$ (α)

and $1e-03$ to $2e-05$ (α^L) compared to positive mode with $1e-04$ as optimum for α and an interval of $2e-03$ to $1e-04$ for α^L .

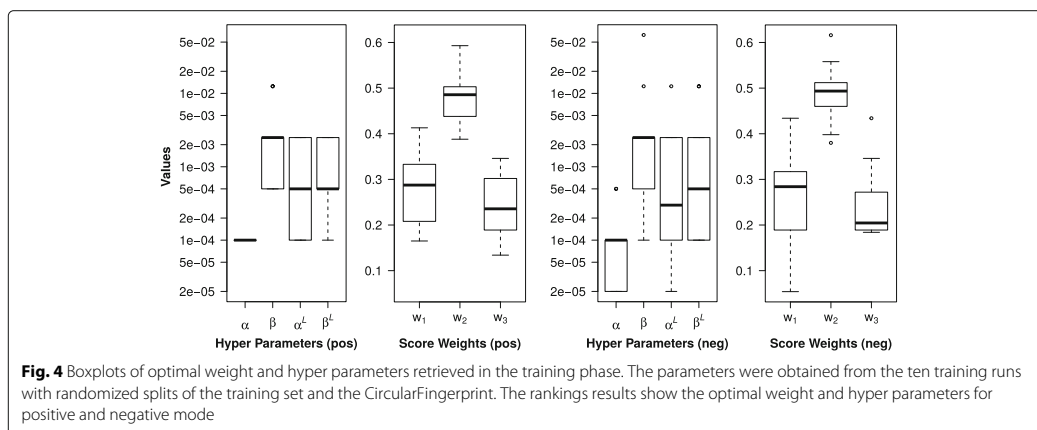
The largest of the weights combining the three scores is ω_2 which gives the score S_{Peak}^c the largest influence in the overall assessment. The median of ω_2 is 0.4855 in positive and 0.4935 in negative mode. The impact of the original MetFrag score $S_{MetFrag}^c$ and S_{Loss}^c are distinctively lower and comparable to each other. The weight ω_1 for the MetFrag score has a median of 0.2875 in positive and 0.2840 in negative mode. The weights for ω_3 are 0.2355 respectively 0.2045.

In the following we analyze the robustness and the homogeneity of the results on the challenge data set with regard to varying parameters across the parameter space evaluated during optimization. This also helped to obtain a better explanation on the deviation of optimized parameters. Specifically we compare the distribution of the Top1 rankings considering (i) the ten optimal parameter sets from the ten randomizations, (ii) the parameter sets within the convex hull constituted by these ten optimal

Table 1 Ranking results in the training phase based on the CircularFingerprint

Top1	Top3	Top10	Top1 (%)	α	β	α^L	β^L	ω_1	ω_2	ω_3	# Spectra
Negative Mode											
55	93	151	20.8	0.00002	0.00250	0.00050	0.00050	0.268	0.460	0.272	265
51	89	155	19.5	0.00002	0.06250	0.01250	0.00050	0.434	0.380	0.186	261
62	101	165	22.9	0.00050	0.01250	0.00010	0.01250	0.309	0.508	0.184	271
70	106	170	25.8	0.00050	0.00250	0.00002	0.01250	0.317	0.494	0.189	271
62	103	161	23.8	0.00010	0.00010	0.00010	0.00250	0.170	0.616	0.214	260
67	110	153	24.0	0.00010	0.00250	0.00250	0.00010	0.300	0.493	0.207	279
63	98	157	22.9	0.00010	0.00050	0.00010	0.00050	0.054	0.512	0.434	275
68	102	158	25.0	0.00002	0.00250	0.00250	0.00250	0.240	0.558	0.202	272
86	114	171	31.2*	0.00010	0.00250	0.00250	0.00010	0.413	0.398	0.189	276
74	106	161	29.0	0.00010	0.00010	0.00002	0.00010	0.189	0.465	0.346	255
Positive Mode											
412	664	925	28.0	0.00010	0.00250	0.00010	0.00250	0.333	0.438	0.229	1471
402	622	866	28.2	0.00010	0.00050	0.00010	0.00250	0.208	0.483	0.309	1426
406	665	913	29.0	0.00010	0.01250	0.00250	0.00250	0.333	0.438	0.229	1399
395	651	894	27.6	0.00010	0.00250	0.00250	0.00250	0.309	0.503	0.188	1432
387	618	839	27.4	0.00010	0.00250	0.00050	0.00050	0.413	0.398	0.189	1413
408	630	870	28.6	0.00010	0.00050	0.00050	0.00050	0.165	0.584	0.251	1428
432	655	910	30.6*	0.00010	0.01250	0.00250	0.00050	0.378	0.488	0.134	1410
400	642	874	28.2	0.00010	0.00250	0.00250	0.00050	0.210	0.488	0.302	1420
385	613	830	27.7	0.00010	0.00250	0.00010	0.00010	0.266	0.388	0.346	1389
396	638	891	27.7	0.00010	0.00050	0.00050	0.00010	0.165	0.593	0.242	1428

The optimization of the parameters was performed on the training data set with ten different random splits of the MS/MS training spectra to be used for first and second training phase. Optimization was performed separately for positive and negative mode. *Runs with the best results based on the relative correct Top1 rankings (neg: R09, pos: R07)

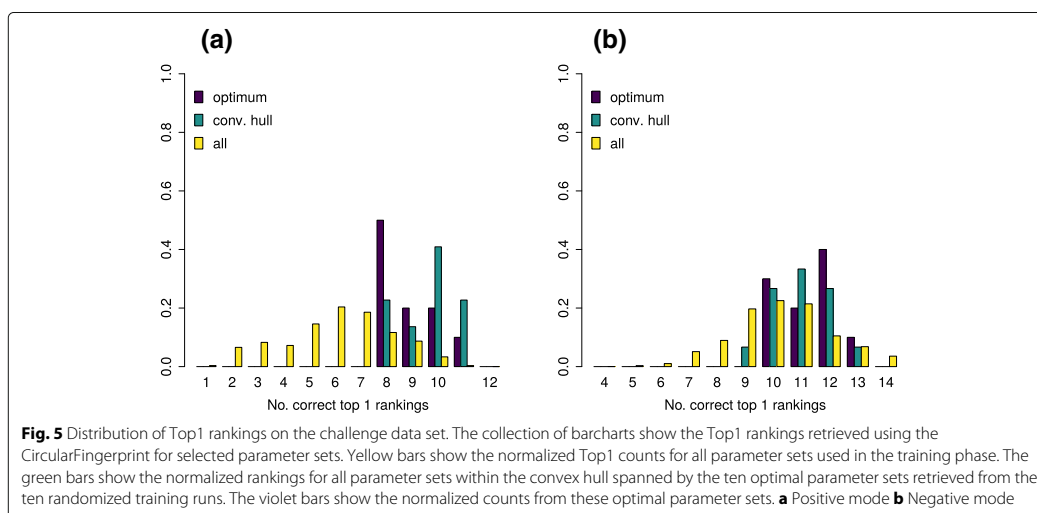


parameter sets in the six dimensional parameter space, and (iii) the complete parameter space evaluated during training of the parameters. The convex hull over the ten optimal parameter sets was calculated using the six degrees of freedom (α , β , α^L , β^L , ω_1 , ω_2) from the seven parameters with the Python *Numpy* package.

Figure 5 shows in yellow the distribution of the Top1 rankings of the CASMI challenge data set for the complete parameter space. Top1 ranking vary from 1 to 12 for the positive and from 4 to 14 for the negative challenge spectra, where the maximum of the distributions are six and ten for positive and negative mode, respectively. If parameter sets are restricted to the convex hull the distribution is clearly shifted to better performance,

where Top1 rankings vary between 8 to 11 for positive and 10 to 13 for negative mode. This range of Top1 rankings is almost identical to the one resulting from the ten optimal parameter sets. The only exception are nine Top1 rankings for parameter sets within the convex hull in negative mode. In positive mode about 76% of the investigated parameters show worse results than achieved by the parameters contained in the convex hull. For negative mode this proportion is reduced to around 15% which can again be explained by the smaller number of available training data.

For the subsequent comparison to other methods on the challenge data set we use the parameter sets resulting in the best relative Top1 ranking performance in the training



phase. The corresponding runs are highlighted in Table 1 and are R07 for positive and R09 in negative mode.

Comparison with MetFrag2.3

The main goal of the integration of the proposed approach into MetFrag was to improve the candidate ranking augmenting the fragmenter score with statistical scores. The MetFrag versions 2.3 and 2.4.5 use exactly the same in silico fragmentation approach. MetFrag2.4.5 scoring was extended with the statistical scoring terms which make the difference in the comparison of both version. The results of MetFrag version 2.4.5 show a drastic improvement of the rankings for the CASMI challenge data compared to its older version 2.3 with regard to all performance measures as given in the first two columns of Table 2. The correct Top1 rankings show a more than four fold increase from 5 to 21 Top1 rankings. The improvement is especially distinct for positive mode with 9 Top1 rankings where MetFrag2.3 resulted in one single query correctly ranked at first position. The number of Top1 hits in negative mode is also increased three fold from 4 to 12. The improvement is also illustrated by the reduced mean and median ranks. Where the mean rank halved to 34.6 the median rank was even reduced by two third to 5. All three scores contribute substantially to these improvements and Top1 rankings vary smoothly with the weight scores (see Additional file 1: Figure S1).

Comparison with other CASMI participants

The MetFrag2.4.5 results were compared to the results obtained by all other participants of CASMI 2016, i.e., CFM_retrain, CSI_IOKR_AR, and CSI:FID_leaveout (abbreviated by CFM-ID, CSI:IOKR, and CSI:FID), MS-Finder and MAGMa. Table 2 shows the original data from Table 7 of [12] with the ranking results for the 87 Challenge MS/MS spectra. The additional MetFrag2.4.5 column summarizes the results achieved using the new MetFrag statistical scoring terms.

In positive mode, MetFrag2.4.5 obtains nine Top1 rankings and shows a similar performance as CFM-ID (9)

and CSI:IOKR (10). CSI:FID (13) outperforms all other approaches with regard to Top1 rankings in positive mode, however did not submit results for negative mode spectra. Figure 6b shows the overlap of the Top1 ranked challenges in positive mode for MetFrag2.4.5 and CSI:FID. There are only five challenges ranked first by both tools and thus a large degree of divergence between the correct predictions.

For the negative mode spectra MetFrag2.4.5 considerably outperformed all participants with 12 Top1 rankings. These are five more queries than MS-Finder could rank in first position and even twice as many than the other statistical approaches CFM-ID and CSI:IOKR.

Considering the complete test data set MetFrag2.4.5 outperforms all participants with regard to Top1, Top3, and Top10 rankings including the declared winner of the contest CSI:IOKR (Top1: 21, Top3: 38, Top10: 55 vs. Top1: 16, Top3: 26, Top10: 46). The improved results are also confirmed by the smaller median and mean rankings of 5 and 34.6 compared to 10 and 97.9. We note that considering the median, CSI:FID shows a better performance than MetFrag2.4.5, however did only submit results for positive mode.

Figure 6a shows the overlap of correctly identified Top1 challenges of the participants which use statistical approaches. Interestingly, there is a relatively large number of challenges that are identified by only one of the approaches. With 10 challenges MetFrag2.4.5 shows the highest amount of unique queries ranked correctly in first place, which is predominantly caused by the eight Top1 negative mode challenges.

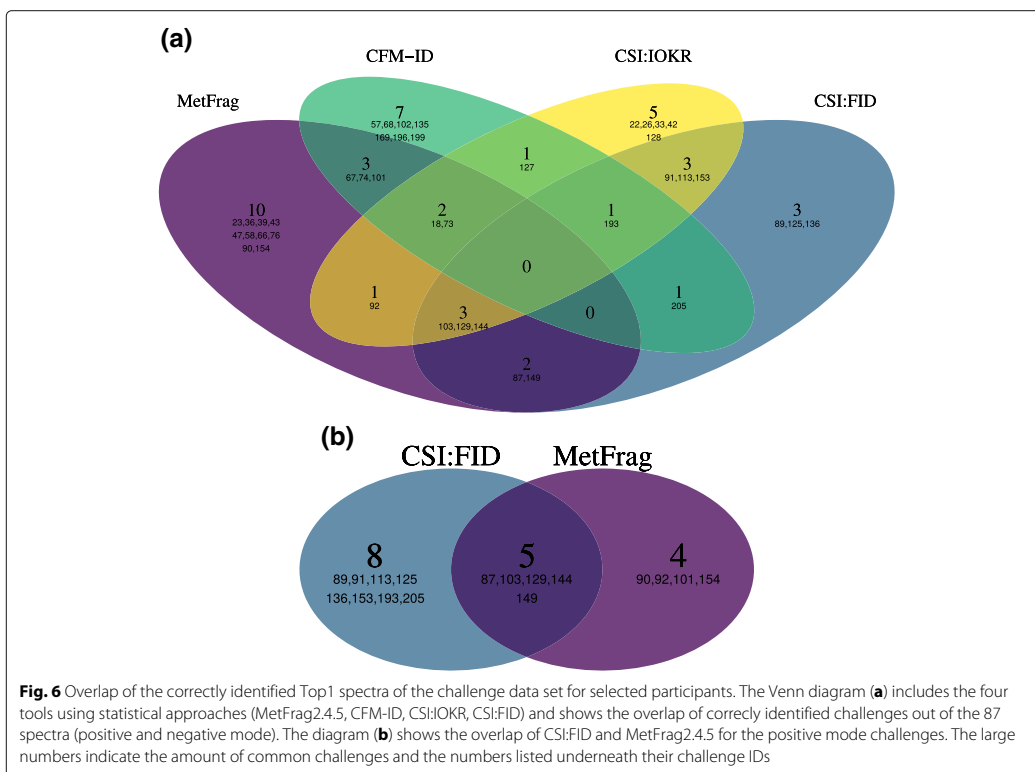
Discussion

The results obtained by the combination of MetFrag's in silico fragmentation approach and statistical fragment annotation learning have shown an overall improvement of the ranking results of the relevant CASMI 2016 test set. Different fingerprint functions have been tested to avoid the expensive graph isomorphism problem to find matching fragments. The training phase revealed a dependency

Table 2 Results for the 87 MS/MS test spectra from the CASMI 2016 Challenge taken from Table 7 in [12] augmented with the results of the proposed approach (MetFrag 2.4.5). For the participants of the challenge the best result is given

	MetFrag 2.4.5	MetFrag 2.3	CFM-ID	CSI:IOKR	CSI:FID	MS-Finder	MAGMa
Top 1 Pos.	9	1	9	10	13	3	2
Top 1 Neg.	12	4	6	6	—*	7	4
Top 1	21	5	15	16	13*	10	6
Top 3	38	16	24	26	23*	25	16
Top 10	55	39	40	46	32*	38	35
Mean rank	34.6	68.4	64.1	97.9	41.5*	28.7	76.8
Med. rank	5	14.5	12.5	10	3*	17.5	23.5

*CSI:FID did not submit results for negative mode spectra



between the number of correct top hits and the fingerprint used. While MACCS- and especially Lingo- and the CircularFingerprint showed the best and also comparable results, the GraphOnlyFingerprint showed a significantly lower number of correct top rankings on the training set. We attribute the inferior performance of the GraphOnlyFingerprint primarily to the lack of representing bond orders and hence encoding less chemical information than all other fingerprint types evaluated. Due to the best performance in the training phase the CircularFingerprint was selected for further investigation on the test set.

Ten different hyper and weight parameter sets resulting from optimization with ten randomized splits of the training data were used to investigate the robustness and the distribution of these parameters across the different training sets. While the optima of the seven parameters varied slightly between the different splits, the parameter sets still showed a clear trend across all ten runs. Especially the effect of the S_{Peak}^c score weight ω_2 was predominantly higher compared to ω_1 and ω_3 for both positive and negative ion mode. The assumption

that the observed parameter variation is an indication for a relatively broad and homogenous parameter optimum was confirmed by the investigation of the ranking results retrieved using parameters located in the convex hull spanned by the ten optima. These distributions also indicate a high robustness of the performance with varying parameter sets across these parameter optima.

An important outcome of this study is the significant improvement of the ranking results retrieved adding the presented Bayesian approach to MetFrag's native in silico fragment annotation. While the improvement gain for the Top3 and Top10 rankings are less pronounced, this comparison impressively demonstrates the benefit including statistical approaches for MS based compound identification. This corresponds to the outcome of CASMI 2016 where a comparison of different statistical and non-statistical approaches was made [12].

The proposed Bayesian approach follows a different mechanism than the existing statistical compound identification methods predicting molecular fingerprints

(CSI:FingerID, CSI:IOKR) or MS/MS spectra (CFM-ID). The comparison of the different approaches on the CASMI 2016 test set used in this study shows on the one hand that the presented approach compares well to the existing ones and on the other hand that a relatively large number of challenges are identified by only one of the approaches (Fig. 6a). From the latter finding it may be concluded that there are different preferences for certain types of spectra of the CASMI 2016 contest. The comparison also revealed that for MetFrag2.4.5 the performance is comparable between positive and negative mode (9 vs. 12). CSI:IOKR shows lower performance ranking result for the negative mode spectra compared to positive mode (6 vs. 10). We assume the combination of in silico fragmentation and statistical scoring has a positive effect in case only limited training data is available. Only a small fraction of negative mode training data was available for this contest and resulted in generally worse results of the statistical approaches in negative mode.

Conclusions

In this work new statistical scoring terms are introduced to MetFrag. This model assesses the assignments of m/z fragment peaks/losses to fragment-structures derived from in silico fragmentation of a candidate and assumes independence of the individual assignments. The model parameters are estimated using the mean posterior approach. Hyper parameters of the statistical model as well as score weights are optimized by a grid search. The performance is evaluated on a subset of the CASMI 2016 contest challenge spectra for which the spectrum was not among the training data set of any participant. The results show that with the integration of the two new statistical scoring terms MetFrag could be improved four fold regarding the number of Top1 rankings. In addition it showed a better performance than the declared winner of the contest CSI:IOKR regarding the number of correctly ranked Top1, Top3 and Top10 candidates. The new scoring terms are now available in the command line tool (version 2.4.5) as AutomatedPeakFingerprintAnnotationScore and AutomatedLossFingerprintAnnotationScore and also in the web interface (<https://msbi.ipb-halle.de/MetFrag>) as “Statistical Scoring” trained on extended data set than used in this work. The additional scoring terms complement current scoring terms based on experimental data and can also be combined with additional meta information if available as described in [3].

We also want to stress that once the method is trained on spectra in the training phase, it can be applied and used for annotation on any data set. The data set can vary whereas the training data set is fixed once the method was trained, which is similar to all other machine learning and statistical methods mentioned in this work.

Additional files

Additional file 1: Figure S1 - Weight Parameter Scan for the test dataset. (PDF 767 kb)

Additional file 2: Figure S2 - Maximum spectral similarities. (PDF 196 kb)

Additional file 3: Figure S3 - Rankings of the correct candidates (test) vs. max. spectral similarity. (PDF 204 kb)

Additional file 4: Table S1 - Notation summary. (PDF 109 kb)

Additional file 5: Table S2 - Notation summary (Scores). (PDF 70.4 kb)

Abbreviations

CASMI: Critical assessment of small molecule identification; CSI:FID: CSI:fingerID; InChI: International chemical identifier; MP: Mean posterior; MS/MS: Tandem mass spectrometry; m/z : Mass-to-charge ratio; mzabs: Absolute mass deviation; mzppm: Relative mass deviation; SVM: Support vector machines

Acknowledgements

We thank all CASMI 2016 participants for generating and providing all result sets of their used software and methods. We acknowledge Emma Schymanski (Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg) for valuable discussions and proof-reading the manuscript. CR and SN acknowledge support from the Leibniz Association's Open Access Publishing Fund.

Authors' contributions

SP, SN, CR contributed to method development, manuscript preparation and revision, discussion. CR implemented all necessary changes to MetFrag and performed data analysis to generate presented results. All authors read and approved the final version of the manuscript.

Funding

CR acknowledge funding from the European Commission for the FP7 project SOLUTIONS under Grant Agreement No. 603437 and for the H2020 project PhenoMeNal under Grant Agreement No. 654241. Funding bodies played no role in study design, data analysis and interpretation, nor manuscript development.

Availability of data and materials

The m/z peak and candidate lists used in this study is available on the official CASMI website, <http://www.casmi-contest.org/2016/index.shtml>. A complete list of the used MassBank and GNPS training spectra and the ranking data sets generated during the current study are available on GitHub, https://github.com/c-ruttkies/metfrag_statistical_annotation. Further information on how to use the new scoring terms with the commandline version of MetFrag can be found on the project website <http://ipb-halle.github.io/MetFrag/projects/metfragcl>. The source code is published on GitHub (<https://github.com/ipb-halle/MetFragRelaunched> (branch: feature/statistical_scoring)).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

SN is Associate Editor for BMC Bioinformatics.

Author details

¹Department Biochemistry of Plant Interactions, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle (Saale), Germany. ²German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, 04103 Leipzig, Germany. ³Institute of Computer Science, Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06099 Halle (Saale), Germany.

Received: 16 November 2018 Accepted: 17 June 2019

Published online: 05 July 2019

References

1. MassFrontier. <http://www.highchem.com/>. Accessed 19 June 2018.

2. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*. 2010;11:148.
3. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J Cheminformatics*. 2016;8(1):1.
4. Wang Y, Kora G, Bowen BP, Pan C. Midas: A database-searching algorithm for metabolite identification in metabolomics. *Anal Chem*. 2014;86(19):9496–503.
5. Tsugawa H, Kind T, Nakabayashi R, Yukihiro D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M. Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal Chem*. 2016;88(16):7946–58.
6. Ridder L, van der Hoof JJJ, Verhoeven S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom*. 2014;3(Special Issue 2):0033.
7. Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*. 2015;11:98.
8. Heinonen M, Shen H, Zamboni N, Rousu J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*. 2012;28(18):2333–41.
9. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci*. 2015.
10. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A*. 2015;112(41):12580–85.
11. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J. Fast metabolite identification with input output kernel regression. *Bioinformatics*. 2016;32(12):28–36.
12. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, Allen F, Vaniya A, Verdegem D, Böcker S, Rousu J, Shen H, Tsugawa H, Sajed T, Fiehn O, Ghesquière B, Neumann S. Critical assessment of small molecule identification 2016: automated methods. *J Cheminformatics*. 2017;9(1):22.
13. McGregor MJ, Pallai PV. Clustering of large databases of compounds: Using the mdl “keys” as structural descriptors. *J Chem Inform Comput Sci*. 1997;37(3):443–8.
14. Vidal D, Thormann M, Pons M. Lingo, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model*. 2005;45(2):386–93.
15. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. Inchi, the iupac international chemical identifier. *J Cheminformatics*. 2015;7(1):23.
16. MassBank of North America. <http://mona.fiehnlab.ucdavis.edu/>. Accessed 8 Dec 2016.
17. Wang MX, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu WT, Criesemann M, Boudreau PD, Esquenazi E, Sandoval-Calderon M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu CC, Floros DJ, Gavilan RG, Kleigrew K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw CC, Yang YL, Humpf HU, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, Klitgaard A, Larson CB, Boya CA, Torres-Mendoza D, Gonzalez DJ, Silva DB, Marques LM, Demarque DP, Pociute E, O'Neill EC, Briand E, Helfrich EJN, Granatosky EA, Glukhov E, Ryffel F, Houson H, Mohimani H, Kharbush JJ, Zeng Y, Vorholt JA, Kurita KL, Charusanti P, McPhail KL, Nielsen KF, Vuong L, Elfeki M, Traxler MF, Engene N, Koyama N, Vining OB, Baric R, Silva RR, Mascuch SJ, Tomasi S, Jenkins S, Macherla V, Hoffman T, Agarwal V, Williams PG, Dai JQ, Neupane R, Gurr J, Rodriguez AMC, Lamsa A, Zhang C, Dorrestein K, Duggan BM, Almaliti J, Allard PM, Phapale P, Nothias LF, Alexandrov T, Litaudon M, Wolfender JL, Kyle JE, Metz TO, Peryea T, Nguyen DT, VanLeer D, Shinn P, Jadhav A, Muller R, Waters KM, Shi WY, Liu XT, Zhang LX, Knight R, Jensen PR, Palsson BO, Pogliano K, Linington RG, Gutierrez M, Lopes NP, Gerwick WH, Moore BS, Dorrestein PC, Bandeira N. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol*. 2016;34(8):828–37. n/a.
18. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, Han L, He J, He S, Shoemaker BA, et al. Pubchem substance and compound databases. *Nucleic Acids Res*. 2015;44(D1):1202–13.
19. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*. 2010;50(5):742–54.
20. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliaskova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G, Evelo CT, Guha R, Steinbeck C. The chemistry development kit (cdk) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J Cheminformatics*. 2017;9(1):33.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



5.5 LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

Michael Witting, **Christoph Ruttkies**, Steffen Neumann, Phillipe Schmitt-Kopplin. LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome. PLOS ONE. 12. e0172311, 2017. 22 Citations⁵ <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0172311>

Contributions

I designed the study together with Michael Witting who did the experimental lab work, data preprocessing and extraction of the MS/MS peak lists. I developed and implemented the statistical models. Michael Witting and I prepared the manuscript. The work was supervised and coordinated by Steffen Neumann and Phillipe Schmitt-Kopplin.

⁵<https://scholar.google.com> (accessed on 01/2021)

RESEARCH ARTICLE

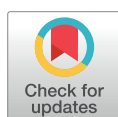
LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

Michael Witting^{1,2}*, Christoph Ruttkies³, Steffen Neumann³, Philippe Schmitt-Kopplin^{1,2}

1 Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstaedter Landstrasse, Neuherberg, Germany, **2** Chair of Analytical Food Chemistry, Technische Universität München, Alte Akademie 10, D-85354 Freising-Weihenstephan, Germany, **3** Leibniz Institute of Plant Biochemistry, IPB Halle, Department of Stress and Developmental Biology, Weinberg, Halle, Germany

* These authors contributed equally to this work.

* michael.witting@helmholtz-muenchen.de



Abstract

Lipid identification is a major bottleneck in high-throughput lipidomics studies. However, tools for the analysis of lipid tandem MS spectra are rather limited. While the comparison against spectra in reference libraries is one of the preferred methods, these libraries are far from being complete. In order to improve identification rates, the *in silico* fragmentation tool MetFrag was combined with Lipid Maps and lipid-class specific classifiers which calculate probabilities for lipid class assignments. The resulting LipidFrag workflow was trained and evaluated on different commercially available lipid standard materials, measured with data dependent UPLC-Q-ToF-MS/MS acquisition. The automatic analysis was compared against manual MS/MS spectra interpretation. With the lipid class specific models, identification of the true positives was improved especially for cases where candidate lipids from different lipid classes had similar MetFrag scores by removing up to 56% of false positive results. This LipidFrag approach was then applied to MS/MS spectra of lipid extracts of the nematode *Caenorhabditis elegans*. Fragments explained by LipidFrag match known fragmentation pathways, e.g., neutral losses of lipid headgroups and fatty acid side chain fragments. Based on prediction models trained on standard lipid materials, high probabilities for correct annotations were achieved, which makes LipidFrag a good choice for automated lipid data analysis and reliability testing of lipid identifications.

OPEN ACCESS

Citation: Witting M, Ruttkies C, Neumann S, Schmitt-Kopplin P (2017) LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome. PLoS ONE 12(3): e0172311. doi:10.1371/journal.pone.0172311

Editor: Monika Oberer, Karl-Franzens-Universität Graz, AUSTRIA

Received: July 25, 2016

Accepted: February 2, 2017

Published: March 9, 2017

Copyright: © 2017 Witting et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data in the open .mzML format and abundance matrix are available from the MetaboLights repository as MTBLS291 (<http://www.ebi.ac.uk/metabolights/MTBLS291>).

Funding: Christoph Ruttkies acknowledges funding from DFG grant NE1396/5-1.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Metabolite and lipid identification represents the current bottleneck in metabolomics and lipidomics. The diversity of the lipidome is huge, with estimates of up to 100,000 different possible lipid structures. This is based on the combinatorial composition of different defined building blocks, which include fatty acids, long-chain bases, glycerol-phosphate, various head groups

and many more [1]. Lipids fulfill several cellular functions, including storage of energy, building blocks of membranes, and signaling.

Several efforts have been made to catalog lipid diversity. Lipidat was one of the first electronic lipid databases [2], and contained 11,000 records, LIPIDBANK (initiated in 1989) contains just over 7,000 records as of 2013, and can still be browsed on the (<http://lipidbank.jp/>) [3].

The Lipid Maps database and classification system structure database (LMSD) [4] is a widely used resource for a systematic classification of lipids. It divides lipids into the eight major classes: fatty acyls (FA), glycerolipids (GL), glycerophospholipids (GP), sphingolipids (SP), sterol lipids (ST), prenol lipids (PR) saccharolipids (SL) and polyketides (PK), each with several subclasses. Lipid Maps contains currently 40,360 structures (accession date 4/2/15) and is accessible via the web (www.lipidmaps.org). LipidHome was developed at the European Bioinformatics Institute (EBI) and is a database of theoretical lipids with 20,297 species and 36 million theoretical sub species [5]. SwissLipids as another resource contains 244,000 known and theoretically lipids [6].

Multiple lipids have similar physicochemical properties which complicates their analysis. Nowadays, two different types of lipid analyses are commonly performed: Liquid chromatography-mass spectrometry (LC-MS) based lipidomics, or shotgun lipidomics. The latter uses direct infusion of the raw lipid extract into a mass spectrometer (MS) and acquisition of multi-stage mass spectra for precise quantification of lipid species using low and high resolution MS [7–10]. High resolution MS, MS/MS and ITMS³ have been employed for structural characterization and quantification of lipids from mouse cerebellum and hippocampus [10]. In contrast to shotgun lipidomics, LC-MS based lipidomics uses chromatographic separation of lipid species followed by mass spectrometric detection, which allows differentiation of isomeric lipid species. Common lipid profiling methods use C8 or C18 reversed phase columns and an acetonitrile/isopropyl alcohol (ACN/iPrOH) gradient. This method allows detection of phospho- and glycerolipids as well as other lipids in a single run [11, 12] and is typically coupled to high-resolution accurate mass Q-ToF or Orbitrap instruments for non-targeted profiling of as many lipids as possible. The high mass resolution and accuracy helps to annotate MS features with known lipids from different databases or to calculate molecular formulas of possible lipid species. However, several lipids, even from different lipid classes, can have identical molecular formulas, e.g. phosphatidylcholine (PC) and/or phosphatidylethanolamine (PE) species, such that a definite identification is impossible from the mass and even molecular formula alone. Searching LipidMaps for molecules having the same molecular formula or exact mass can result in up to 115 candidates with one single molecular formula. Taking into account possible adducts during the ionization process the number increases further. For example the sodium adduct ($[M+Na]^+$) of PC(18:0/20:1) and the $[M+H]^+$ adduct of PC(18:0/22:4) have a mass difference of 0.0024, which reflects a deviation of 2 ppm. A more extreme example is the formic acid adduct of PC(16:0/20:1) $[M+HCOO]^-$ and the $[M-H]^-$ of PS(17:0/22:0) having exactly the same molecular formula. Today's ultrahigh resolution mass spectrometer, like Orbitrap or FT-ICR-MS instruments can reach mass errors below 1 ppm, but even with these ultrahigh resolution MS instruments it is only possible to accurately calculate a molecular formula, whereas no information about the lipid class, structure or fatty acid side chain composition is available [13]. Thus, tandem mass spectrometry (MS/MS) is needed to provide further information for a more reliable annotation. Fragmentation in positive ion mode can help to reveal the lipid class by neutral losses of lipid head groups, whereas negative ion mode resolves the fatty acid composition and position [14]. Data dependent acquisition (DDA) offers the possibility to collect several hundred MS/MS spectra for identification of metabolites or lipids during chromatographic runs [15].

Interpretation of the resulting MS/MS spectra, especially in high-throughput studies, is rather limited and manual analysis of several hundreds to thousands of MS/MS spectra is not feasible. To speed up identification, comparison against reference spectral databases is possible, but the lipid coverage in these databases is sparse. Lipid Maps currently contains only few hundreds low resolution MS/MS spectra, while MassBank has 3,158 records on both low and high resolution instruments covering 707 unique lipids [16].

In silico fragmentation has been suggested as a possible solution to analyze MS/MS spectra without the need of reference spectral databases [17]. LipidBlast is a spectral library that includes a 212,516 *in silico* generated tandem mass spectra covering 119,200 compounds from 26 lipid classes [18]. More recently, Greazy, an approach for identification of phospholipids from MS/MS data was presented which includes the estimation of false discovery rates (FDR). The modul LipidLama, integrated in Greazy, uses kernel density estimation to fit non-parametrized models to distinguish false and true lipid assignments. The cutoff score for a putative correct lipid assignment can then be defined by using a pre-defined FDR of e.g. 5% [19].

In this study we present a workflow to improve the reliability of *in silico* MS/MS annotations of lipids. To achieve this, we introduce bayesian classifiers based on parametrised distributions and maximum-likelihood estimation to calculate a reliability score for a result to be a correct annotation among its lipid class, which is based on training data obtained from lipid standard materials and true positive manual identifications. This workflow consists of the annotation of precursor masses with possible lipid structures using MassTRIX [20–22], followed by MetFrag batch processing of candidates retrieved via the putative neutral masses derived from ion species annotation results. The performance was evaluated using MS/MS spectra obtained previously with UPLC-Q-ToF-MS/MS and data dependent acquisition (DDA) [23]. The lipid classes relevant for this paper are depicted in Fig 1, which included ceramides, different glycerophospholipids classes and glycolipids. Results from this training allowed the development of the central new feature in LipidFrag, the classifiers to predict the probability of a reliable MetFrag annotation for an unknown lipid class (Fig 2A). This is used to differentiate between good and poor identification results and to predict the underlying lipid main class of the precursor in high-throughput MS/MS experiments like in this case study performed with the lipid extract of *C. elegans*.

Methods

Chemicals

HPLC-grade methyl-tert-butyl ether (MTBE) and LC-MS-grade methanol (MeOH), iso-propanol (iPrOH), acetonitrile (ACN), ammonium formate and formic acid were obtained from Sigma-Aldrich. Water was purified using a Merck Millipore Integral water purification system with a resistance of 18 M Ω and TOC < 5 ppb.

Lipid standard material preparation

Phosphatidylcholine preparation from chicken egg (840051P, Avanti Polar Lipids), Escherichia coli polar lipid extract (100600P, Avanti Polar Lipids), phosphatidyl serines from porcine brain (840032P, Avanti Polar Lipids), ceramide from porcine brain (860052P, Avanti Polar Lipids) and ceramide from chicken egg (860051P, Avanti Polar Lipids) were obtained from Avanti Polar Lipids (Otto Nordwald GmbH, Germany) and dissolved in MeOH at a concentration of 1 mg/mL. Additionally, L-alpha-Phosphatidylinositol sodium salt from Glycine max (P0639), Triglyceride mix (17811-AMP), 1,3-Dioleoyl-2-palmitoyl-glycerol (D1657), Glycerol tritricosanoate (T1412), Glycerol trioleate (T7140) and 1,2-Dilinoleoyl-3-palmitoyl-rac-glycerol (D0301) were obtained from Sigma-Aldrich (Taufkirchen, Germany) and dissolved in either

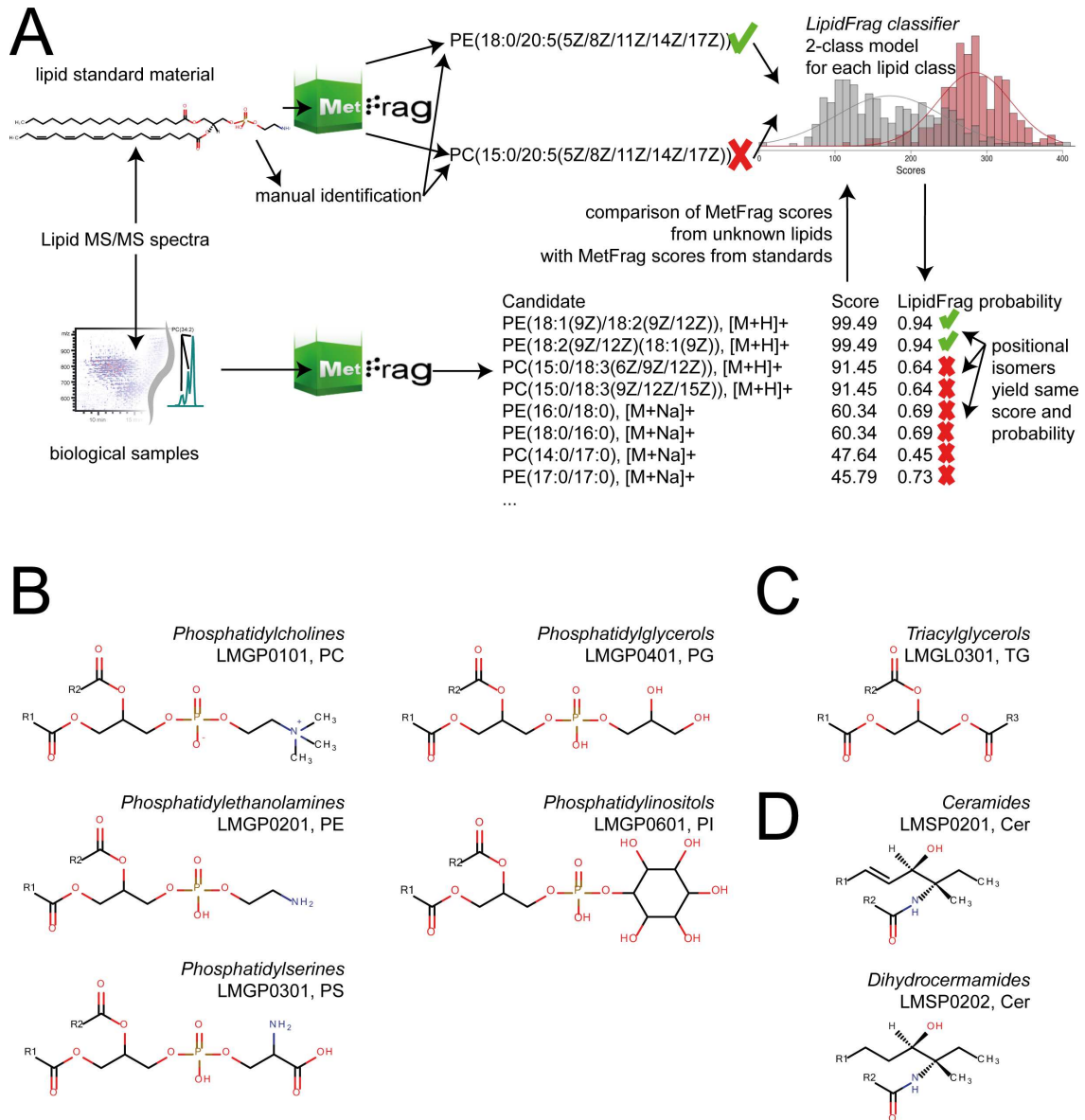


Fig 1. LipidFrag workflow and related lipid sub-classes. (A) Schematic drawing of LipidFrag workflow. MS/MS spectra from known lipids derived from lipid standard materials and from unknown lipids are subjected to MetFrag *in silico* fragmentation, whereby all possible precursor structures are taken into consideration. During training phase true positive identity and decoy candidates are used to calculate a 2-class classifier by which reliable results from unknown lipids can be identified. (B) Structures of detected phospholipid classes, phosphatidylethanolamines (PE, LMGPO201), phosphatidylcholines (PC, LMGPO101), phosphatidylglycerols (PG, LMGPO401), phosphatidylserines (PS, LMGPO301) and phosphatidylinositols (PI, LMGPO601) (C) Structure of triacylglycerols (TG, LMGL0301) (D) Structure of ceramides (Cer, LMSP0201) and dihydroceramides (Cer, LMSP0202).

doi:10.1371/journal.pone.0172311.g001

5.5 LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

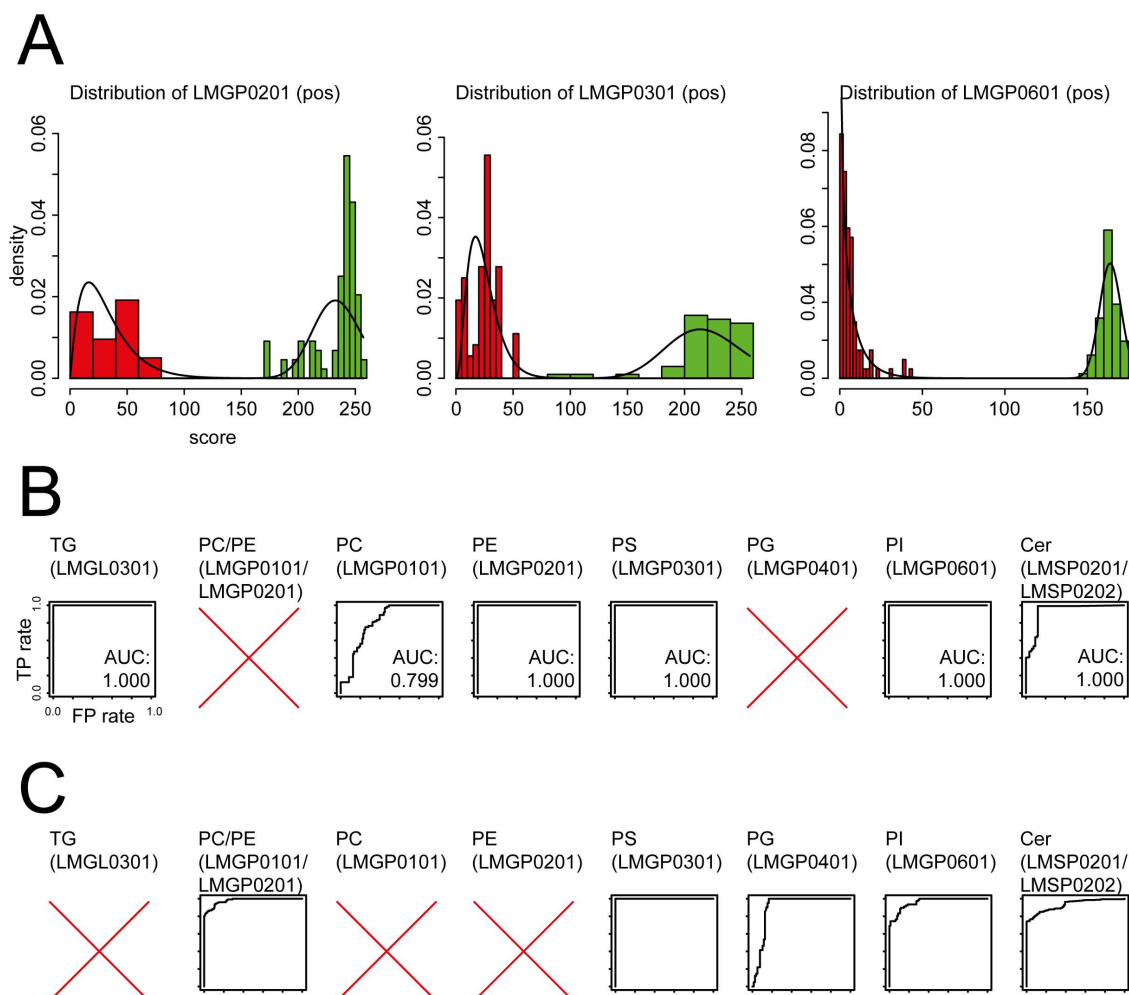


Fig 2. Visualization of input data and results obtained by LipidFrag. (A) Examples of histograms showing distribution of raw MetFrag score for the back- and foreground training dataset. (B) Receiver-Operator characteristics (ROC) derived from 10-fold cross-validation of MS/MS spectra from lipid standard materials detected negative ion mode. (C) Receiver-Operator characteristics (ROC) derived from 10-fold cross-validation of MS/MS spectra from lipid standard materials detected positive ion mode. In both panels, plots having no AUC value indicate that this lipid class was not detected in this ion mode concluding that there was not training data for classifiers available. All axes have the same scale.

doi:10.1371/journal.pone.0172311.g002

MeOH, MTBE, CHCl₃ or solvent mixtures, depending on solubility. Different samples for analysis were prepared and diluted in ACN/iPrOH/water (65/30/5, v/v/v) to 10 µg/mL for analysis. The following lipid classes and standard samples were analysed and are named throughout the paper as indicated in brackets: Phosphatidylcholines (PC, LMGP0101), phosphatidylethanolamines (PE, LMGP0201), phosphatidylglycerols (PG, LMGP0401),

phosphatidylserines (PS, LMGP0301), phosphatidylinositols (PI, LMGP0601), ceramides (Cer, LMSP0201/ LMSP0202), and triacylglycerols (TG, LMGL0301).

Lipid extraction from *C. elegans*

Lipids were extracted from young adult *C. elegans* using a modified method from Matyash et al. [24], described in [23]. The worms were washed off the plates and their metabolism was quenched with 500 μ L -20°C MeOH. Samples were flash frozen in liquid nitrogen and stored at -80°C prior to extraction. Samples were then thawed on ice and 1.7 ml MTBE was added and samples were vortexed vigorously. *C. elegans* were lysed for 30 minutes in an ice cold ultrasonic bath, after which 420 μ L of water was added and samples were sonicated for further 15 minutes. Phases were separated by centrifugation at 4°C and 14,000 rpm for 15 minutes. The upper organic phase was transferred to a 4 ml glass vial and the remaining lower phase was re-extracted with additional 650 μ L MTBE for 15 minutes. After centrifugation the organic layers were combined and evaporated in a SpeedVac vacuum concentrator at 30°C for 0.5-1h. The residue was redissolved in 500 μ L ACN/iPrOH/water (65/30/5, v/v/v).

UPLC-Q-ToF-MS lipid profiling

Lipid profiling was performed as previously described [23]. Briefly, separation was achieved on a Waters Cortecs C18 column, 150mm x 2.1 mm ID, 1.6 μ m using a Waters Acquity UPLC (Waters, Eschborn, Germany) coupled to a Bruker maXis UHR-Q-ToF-MS (Bruker Daltonic, Bremen, Germany). Flow rate was 0.25 ml/min and column temperature was set to 50°C . Eluent A consisted of 60% ACN and 40% water, eluent B of 90% iPrOH and 10% ACN, both with 10 mM ammonium formate and 0.1% formic acid. Detection was carried out in positive and negative ion mode with data dependent acquisition with a scan rate of 5 Hz and selection of 2 precursors. Masses were excluded from DDA after 3 spectra and released from exclusion after 0.15 min. An absolute threshold of 1500 was used for selection.

MS data processing

MS data was imported to Genedata Expressionist for Mass Spectrometry 8.2 (Genedata, Basel, Switzerland) for internal re-calibration, retention time alignment and peak picking. Files were exported to .xlsx format and further data handling was carried out in MS Excel. Lipids were annotated with a new in-house version of MassTRIX to also cover the adducts $[\text{M}+\text{NH}_4]^+$ and $[\text{M}+\text{HCOO}]^-$, as well as $[\text{M}+\text{H}]^+$, $[\text{M}+\text{Na}]^+$ and $[\text{M}-\text{H}]^-$ and an absolute error of 0.005 Da [22].

MS/MS spectra were exported from the calibrated and aligned chromatograms from Genedata Expressionist for MS 8.2 as .mgf file. Only spectra associated with a detected feature were kept and converted to MetFrag batch files (available at <http://msbi.ipb-halle.de/msbi/lipidfrag>) using a custom Perl script. The neutral mass and formula for the batch file were obtained by annotation with MassTRIX, for all possible annotation results. Finally, spectra in batch files were de-isotoped using the CAMERA package with a custom R script [23]. The raw data in the open.mzML format and abundance matrix are available from the MetaboLights repository as MTBLS291 (<http://www.ebi.ac.uk/metabolights/reviewerfec5e44e-fae6-46de-b55d-d2f22d425286>).

Manual lipid identification

Manual lipid identification was performed using known lipid fragmentation pathways. Information from both ionization modes was combined and matched via identical retention times, where available. For phospholipids, fragments used for identification included head group

fragments and their respective neutral loss, loss fatty acid side chains and their carboxylate fragment. Position of fatty acids was inferred from intensity distributions of the respective [M-sn1], [M-sn2], sn1-fatty acid and sn2-fatty acid fragments. In the case of triacylglycerols neutral losses of fatty acid side chain as ammonium salt and the respective fragments were used. Ceramide species were identified based on typical sphingolipid fragments, e.g. loss of N-bound fatty acid and sphingoid base fragments.

Since exact position and stereochemistry of double bonds cannot be deduced from these experiments, all possible isomers were reported as potential identification for further processing with LipidFrag.

LipidFrag identification

Batch query files were processed with the MetFrag command line tool (version 2.4 available at <https://c-ruttkies.github.io/MetFrag/>). Lipid Maps (LMSDFDownload18Mar14) was used as structure database. Candidates were considered within 20 ppm of the theoretical mass, and measured MS/MS peaks were matched against *in silico* fragments, generated with tree depth 3, with an error window of 0.01 Da + 15 ppm. The ion mode for the generated fragments were set according to the acquisition of the processed MS/MS peak list and the minimum peak intensity was set to 1000 arbitrary units. The resulting ranked candidate lists were filtered by the first part of the molecules' InChIKey to eliminate stereo isomers and stored as CSV files, with the calculated MetFrag scores stored in the CSV columns. CSV files for MS/MS peak lists containing less than two informative MS/MS peaks were excluded from the evaluation. The score calculated by MetFrag was used to rank the known candidates of the standard spectra. Here, we always used the pessimistic (worst case) ranking result when candidates, including the correct one, shared equal MetFrag scores. Hence all potential isomers, e.g. double bond positional isomers, which usually have identical MetFrag scores, are covered and reported.

The original MetFrag scoring function considers the bond dissociation energy (BDE) of bonds which are cleaved during the *in silico fragmentation*. As the cleavage of C-C bonds of the fatty acid chains is unlikely to occur under the given conditions in the mass spectrometer, the BDE of this bond type was set to the arbitrarily high value of 10e9, which effectively eliminates fragments generated by a C-C cleavage.

Lipid class specific classifiers for reliability calculation

A new feature of LipidFrag is the use of classifiers for reliability calculation of the MetFrag result. The distribution of the MetFrag raw scores depends on both the query spectra and the compound classes of the candidates, as shown in [S1 Fig](#), [S2 Fig](#) and [S3 Fig](#). Generally, in metabolomics this structural compound class classification is neither always obvious nor easy to obtain for small molecules, but for lipids there is the structural categorization initiated by the International Lipid Classification and Nomenclature Committee (ILCNC), available on the Lipid Maps website [4]. With this nomenclature, the structures are hierarchically ordered and encoded as positions in the Lipid Maps ID. This classification was used here to obtain well-defined ranges of MetFrag raw scores for particular lipid classes. Therefore, a training step was implemented to predict the reliability of MetFrag results based on the training of classifiers with MS/MS spectra of the lipids standard material for different lipid sub classes ([S2–S6 Tables](#)). For this task one classifier was created for each lipid subclass, where raw scores of correctly identified structures from the lipid standard materials served as foreground data. The same spectra were queried with deliberately wrong precursor candidates in the same mass range (up to 150 ppm), originating from the other lipid sub classes respectively, to obtain a decoy database and subsequently the MetFrag scores for the background data set. This

approach was inspired from proteomics, where foreground and background training data are used to assign significance values to peptide identifications [25].

Gamma distributions were used to model the scores for the foreground and background data. The model parameters for the distributions were calculated by maximum-likelihood estimation on the fore- and background dataset. For each lipid class a separate classifier was trained, because the MetFrag scores exhibit large differences between the classes.

Eq (1) shows the calculation of the foreground class probabilities (FCP) of a MetFrag result with the bayesian approach, where $P(\text{score} | \text{Foreground}, \Theta)$ is the likelihood of the foreground model represented by a gamma distribution of the lipid sub class for the present score and $P(\text{score} | \text{Background}, \Theta)$ is the corresponding likelihood of the present score in the background model which is also represented by a gamma distribution. The estimated parameters of the distributions are represented by Θ .

For testing, a 10-fold cross-validation was applied. FCPs of the lipid classes were used to calculate the true positive and false negative rates on the test dataset to determine a Receiver Operating Characteristic curve (ROC) and the Area under Curve (AUC) as quality measure of the different classifiers.

$$FCP = \frac{p(\text{score} | \text{Foreground}, \theta)}{p(\text{score} | \text{Foreground}, \theta) + p(\text{score} | \text{Background}, \theta)} \quad (1)$$

Reliability of MetFrag results

After training, the classifiers were used to predict the reliability of MetFrag candidate identifications for the *C. elegans* MS/MS spectra, where the correct candidate is unknown. Given a candidate list processed by MetFrag as SDF or CSV file, LipidFrag calculates the FCP for each candidate lipid in this result list by first selecting the appropriate classifier based on the candidate's Lipid Maps ID. The selected classifier together with the calculated MetFrag raw score is used to calculate the FCP value. Those results, where no candidate exceeds a defined FCP threshold (of e.g. 0.95) have to be treated as unreliable or not identified.

LipidBlast identification

For comparison lipid annotations were performed using the LipidBlast *in silico* tandem MS library [18]. The provided LipidBlast fork (v2 Hiroshi Tsugawa fork) was downloaded and converted by Lib2NIST tool (v1.0.4.38 (beta), options: "Include Synonyms": Yes, "MW from chem. formula": Yes, "MS/MS spectra only": Yes, "2008 MS Search compatible": Yes) to NIST format and used as spectral library for LipidBlast annotation of all standards used for LipidFrag available in MGF format obtained from Genedata Expressionist for MS 8.2. The NIST MS Pep Search GUI (v0.91, options: defaults except for "Q-TOF": Yes, "Min. match factor": 100, "Presearch mode": Standard, "Load libraries in memory": No, "Max. number of output hits": 10, "Presearch mode": Standard, "Precursor ion tolerance": 0.02, "Fragment peak m/z tolerance": 0.02) was used to process input spectra in batch mode. LipidMaps identifiers provided for the correct identifications were mapped to common names annotated by LipidBlast for comparison with the LipidFrag annotations. The pessimistic rankings (among the top 10 reported candidates) were calculated based on the Rev-Dot (reverse dot) scores and compared with the LipidFrag results.

Availability of LipidFrag

LipidFrag comes with several R scripts available at <https://github.com/c-ruttkies/LipidFrag> together with the used data for training and an example for lipid class prediction. After

prediction and model parameter training LipidFrag uses a MetFrag result CSV file retrieved by using MetFrag and the Lipid Maps database for candidate retrieval and predicts the underlying lipid class. The calculated FCP value is an indication of the reliability of the lipid identification. The newest version of the MetFrag commandline tool is available at <http://c-ruttikies.github.io/MetFrag>.

Results

LipidFrag uses the result scores of a lipid candidate list retrieved from MetFrag, which performs *in silico* fragmentation of lipids. Then the matching classifier is selected based on the lipid sub class of a currently considered lipid candidate in the candidate list. Using the bayesian equation, LipidFrag calculates the posterior probability of the MetFrag score under the assumption to come from the foreground distribution of the selected bayesian classifier. This probability value can then be used as prediction of the lipid class of the regarded MS/MS spectrum, and secondly, as a measure of reliability of the current MetFrag lipid annotation to filter out false positive lipid assignments.

Analysis of lipid standard materials

For the positive ion mode spectra, classifiers were built for the following lipid sub classes: PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PI (LMGP0601), Cer (LMSP0201/ LMSP0202) and TG (LMGL0301). As the scores for the Cer species (LMSP02) show a bimodal distribution in positive ion mode, two separate classifiers were trained for the available ceramide sub classes (LMSP0201 and LMSP0202) for the foreground data. Compared to a single classifier for the whole Ceramide main class, this captures the multimodal score ranges of the lipid sub classes in a better way (S1 Fig). For the negative ion mode spectra, the lipid sub classes: PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PG (LMGP0401), PI (LMGP0601) and Cer (LMSP0201/ LMSP0202) were used for training. As candidates for the LMGP0101 sub class show similar MetFrag scores on LMGP0201 sub class MS7MS spectra a combined classifier was trained. This resulted in six different classifiers for positive and five for negative ion mode. With these classifiers, used for positive and negative ion mode, LipidFrag is able to cover already over one third of the lipid species in the Lipid Maps database.

The classifiers were extensively cross-validated on the lipid standards spectra to generate receiver operating characteristic (ROC) curves and the corresponding area under curve (AUC) values as measure of identification performance. These values are partly shown in Table 1, for the full results see S1 Fig. For clarity, results are grouped into three lipid types: ceramides, glycerophospholipids and glycerolipids, and presented separately in the following paragraphs. Mean ranks shown in Table 1 are calculated with and without a FCP threshold to highlight the performance using the LipidFrag classifiers. To reduce the false negative rate a FCP threshold of 0.6 was set within LipidFrag. With this value the number of false positive assignments could be reduced from 91% to 57% for positive ion mode and from 93% to 27% for negative mode.

Ceramides

Ceramides have quite distinct molecular formulas compared to other lipid classes (i.e. glycerophospholipids); therefore, the overlap with other classes and the number of potential candidates is low. Major differences between different ceramide species are the length of the sphingoid base, the number of hydroxyl groups in the sphingoid base, the length of the N-linked fatty acid and total number of double bonds. The fragmentation of ceramides has been studied extensively by Hsu et al. [26], focusing mainly on the $[M-H]^-$ ions, whereas here ceramides were observed

Table 1. Results of the MetFrag identification and the classifier testing.

Negative ion mode							
Metric	LMGL0301 (TG)	LMGP0101, LMGP0201 (PC, PE)	LMGP0201 (PE)	LMGP0301 (PS)	LMGP0401 (PG)	LMGP0601 (PI)	LMSP0201, LMSP0202 (Cer)
FCP+	—	0.871	—	0.979	0.888	0.834	0.817
FCP-	—	0.098	—	0.009	0.164	0.154	0.236
AUC	—	0.979	—	1.0	0.901	0.961	0.931
Mean Rank	—	2.2	—	1.8	1.8	2.6	1.3
Mean Rank (FCP > = 0.6)	—	2.3 (68%)	—	1.8 (55%)	1.8 (94%)	2.4 (78%)	1.2 (63%)
Cand	—	31.3	—	15.8	15.3	14.6	2.3
positive ion mode							
Metric	LMGL0301 (TG)	LMGP0101 (PC)	LMGP0201 (PE)	LMGP0301 (PS)	LMGP0401 (PG)	LMGP0601 (PI)	LMSP0201, LMSP0202 (Cer)
FCP+	1.000	0.551	0.994	0.969	—	1.000	0.908
FCP-	0.000	0.442	0.000	0.000	—	0.000	0.095
AUC	1.0	0.799	1.0	1.0	—	1.0	0.935
Mean Rank	3.1	5.8	1.7	1.9	—	1.0	1.17
Mean Rank (FCP > = 0.6) (FP-Rate)	1.7 (16%)	3.0 (45%)	1.7 (49%)	1.9 (9%)	—	1.0 (100%)	1.0 (68%)
Cand	33.9	26.5	26.5	14.9	—	15.3	2.2

The mean values of the FCPs retrieved from the cross-validation for the foreground (FCP+, higher is better) and the background (FCP-, lower is better) scores are shown. An AUC of 1.0 represents the best possible classification result for the corresponding lipid main/sub class. Additionally, the mean rank of the correct candidate (Rank) using MetFrag and LipidFrag with a FCP threshold of 0.6 together with the discarded proportion of false positives (FP-Rate) and the mean number of candidates retrieved (Cand) are given.

doi:10.1371/journal.pone.0172311.t001

predominantly as $[M+HCOO]^-$ adducts in negative ion mode. Both positive and negative ion modes were used to characterize the ceramides. In total, 17 ceramides were identified manually from obtained MS/MS, with 11 found in both ion modes, 2 in negative and 4 in positive ion mode only.

LipidFrag shows the best results for ceramides in positive ion mode, indicated by the AUCs of 0.935 for the sphingenine and sphinganine lipids (LMSP0201 /LMSP0202). In negative ion mode the AUC is also good with a value of 0.931. The mean rank of the correct solution is 1.17 in positive and 1.3 in negative ion mode, which is also due to the low number of candidates (see Table 1).

Glycerophospholipids

Different classes of glycerophospholipids were subjected to fragmentation, including PC (LMGP0101), PE (LMGP0201), PS (LMGP0301), PG (LMGP0401) and PI (LMGP0601). The molecular formulas of PCs and PEs overlap considerably, which can lead to ambiguous results if only the accurate mass of the precursor is used for the annotation with potential structures. Ekroos et al. studied the use of fragmentation and fatty acid scanning using an ion trap MS for elucidation of the fatty acid composition of PCs [14]. Fragmentation is very class and ion mode specific, e.g. PCs yield mainly m/z 184.07 as the head group fragment in positive ion mode, whereas in negative ion mode fragments originating from $[M+HCOO]^-$ adducts provide information about fatty acid composition and their positions. Diagnostic fragments indicating fatty acid composition were only detected for very high abundant species in positive ion mode. Several studies have shown that the carboxylate anion from the sn2 fatty acid is up to

three times higher compared to sn1 [27]. PEs in contrast show mainly the diacylglycerol fragment derived from the neutral loss of the head group in positive mode and side chain fragments of very low intensity (usually below 2%). Therefore, MS3 of the diacylglycerol fragment is needed for side chain identification in positive ion mode. In negative ion mode, fragmentation of PE species yields carboxylate anions from sn1 and sn2 fatty acids similar to PCs.

Most of the glycerophospholipids show very good identification results with LipidFrag. This is indicated with the mean rank values 2.24, 1.8, 1.8 and 2.6 for the available PC/PE (LMGP0101/LMGP0201), PS (LMGP0301), PG (LMGP0401) and PI (LMGP0601) species in negative ion mode. The AUCs of 0.979, 1.0, 0.901 and 0.961 also show excellent classification results (Table 1 and Fig 2).

In positive ion mode the PE (LMGP0201), PS (LMGP0301) and PI (LMGP0601) species show similar results with mean ranks of 1.7, 1.9 and 1.0 and the AUCs of 1.0. Though, the PC (LMGP0101) species show a similar performance with a mean rank of 1.69 when using a FCP filter with threshold 0.6 (see S8 Table) the filter sorted out 58 of the 71 spectra caused by the limited fragmentation which also indicated by a lower AUC of 0.799 (Table 1).

Glycerolipids

Glycerolipids (LMGL0301) were detected in positive ion mode mainly as $[M+NH_4]^+$ adducts. From this adduct, typical fragmentation is the neutral loss of fatty acid side chains plus ammonia yielding a diacylglycerol-like fragment [28]. This loss can occur for all side chains and lead to a pattern that allows the identification of composition, but rarely provides sufficient information to determine the position of fatty acids in the intact lipid.

Five different TG standards were employed as training data, showing previously known fragmentation pathways. These five compounds had different fatty acid compositions and therefore different retention times. However, in *C. elegans* samples many possible isomers and isobars are co-eluting with many different fatty acid combinations that can be deduced from fragmentation data (Fig 3). The TG species are observable in positive ion mode and the relating classifier shows a good result with an AUC of 1.0. However, the mean rank indicates a lower performance for the identification results with 3.1, as the typical loss of a fatty acid side chain during fragmentation is not only explained by the correct candidate, but also by structurally very similar TG species. The fragment peaks of these types of losses seem to be very specific for the lipid main class, indicated by the high AUC, but this does not help to distinguish between different TG lipids sharing the same molecular formula.

Handling of mixed spectra

One problem not only for LipidFrag are non-pure spectra arising from co-isolation of co-eluting isomeric/isobaric lipids during the MS measurement. In order to test how well LipidFrag can deal with this, we created such spectra *in silico* using measured spectra as template. Overlap especially occurs for glycerolipids in the later elution range of the chromatogram, but might also occur for other lipids. Although the used UPLC method can separate major isobars of the glycerophospholipids [23], overlap might also occur with major interference most likely coming from isomers/isobars within same lipid class. Interference from different lipid classes having the same molecular formula can be neglected because polarity, and hence retention time is very different (e.g. PE(18:0/20:2) has a logP of 13.12, whereas the isobaric PC(18:0/17:0) has a logP of 11.47).

We used one measured lipid MS/MS as target and added interfering MS/MS peaks at the intensity ratios of 10:1, 2:1 and 1:1 and evaluated the MetFrag raw score of the true candidate. Mixtures included binary, ternary and even quaternary mixes of isobaric lipids (S1 Table).

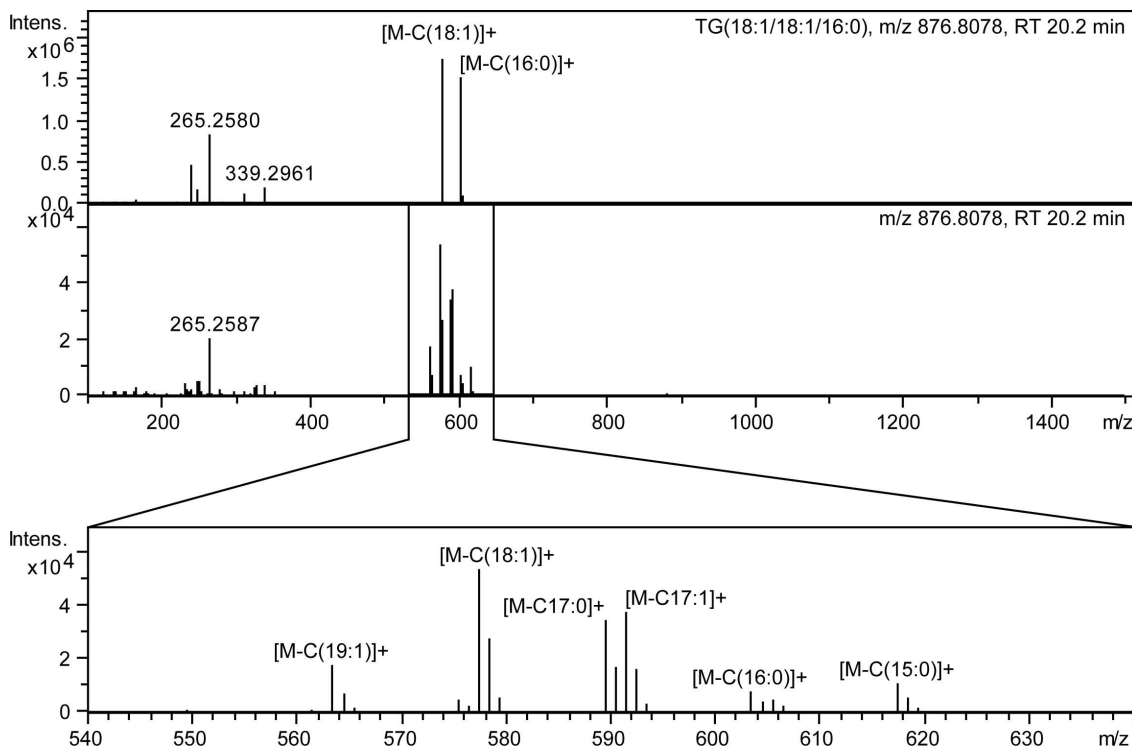


Fig 3. Example of co-elution and overlapping of different TG species in *C. elegans*. Analysis of spectra derived from TGs is complicated in real samples due to overlap of several isomeric and isobaric species. The upper panel shows the MS/MS spectrum of TG(18:1/18:1/16:0) standard and the lower of the same chromatographic peak in a *C. elegans* lipid extract.

doi:10.1371/journal.pone.0172311.g003

Results indicate that mixtures with an equal amount of target and interference cause a drop in the score and rank of the true candidate (S4 Fig) as expected.

The target substances still rank in the upper quarter. Results from one particular example in *C. elegans* samples having two isomeric PC species in on MS/MS spectrum are discussed in a later section (see Analysis of *C. elegans* samples).

Analysis of publicly available MSMS spectra

To test the performance of the LipidFrag approach on an independent second dataset we used 415 negative ion mode lipid MS/MS spectra retrieved from Bio-MassBank [29] where a Lipid Maps ID was available for the correct candidate. Although these spectra were measured on a different instrument with higher mass error than the data used for classifier training, they served as an additional validation of the workflow. Altogether, the spectra were annotated by the submitters to be from ten different sub classes (LMGL0301, LMGP0101, LMGP0102, LMGP0103, LMGP0105, LMGP0201, LMGP0202, LMGP0203, LMGP0601 and LMSP0301). Table 2 shows the ranking results obtained from LipidFrag. The mean ranks within the lipid sub classes were 4.4, 6.0, 2.9, 3.9, 2.3, 2.8, 1.0, 1.0, 2.0, 1.8, respectively. Only two classifiers were available for the spectra originating from PC/PE (LMGP0101 /LMGP0201) and PI

5.5 LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

Table 2. LipidFrag rankings on the 415 Bio-MassBank spectra.

Lipid sub class	Mean rank	Median rank	Mean candidates	Median candidates	Number MS/MS
LMGL0301 (TG)	4.4	2.0	15.0	15.0	7
LMGP0101 (PC)	6.0	3.5	22.0	23.0	118
LMGP0102 (PC)	2.9	3.0	9.2	8.0	36
LMGP0103 (PC)	3.9	2.5	15.7	14.0	18
LMGP0105 (PC)	2.3	2.0	4.2	4.0	30
LMGP0201 (PE)	2.8	2.0	17.2	19.0	53
LMGP0202 (PE)	1.0	1.0	7.0	7.0	12
LMGP0203 (PE)	1.0	1.0	12.5	14.5	24
LMGP0601 (PI)	2.0	2.0	11.3	11.0	9
LMSP0301 (SM)	1.8	1.0	18.9	14.0	108
All	3.3	2.0	16.6	16.0	415

For each lipid sub class the number of MS/MS spectra available and the retrieved mean and median rank as well as the mean and median number of candidates are given.

doi:10.1371/journal.pone.0172311.t002

(LMGP0601) species. For the 180 MS/MS spectra 157 have been identified with the correct lipid sub class based on the foreground class probability (FCP) which is a true positive rate of ~87% for the low resolution spectra where a classifier was available. The LMGP0601 classifier calculated a sub class FCP which reached this threshold for all cases (9) and the LMGP0101/LMGP0201 classifier for 148 out of the 171 cases.

Comparison with LipidBlast annotations

The results of LipidBlast compared with the mean ranks of LipidFrag are shown in Table 3. The values indicate that results are comparable between both software tools. Nevertheless, there are slight deviations for some lipid classes and LipidFrag usually annotates more spectra (FCP threshold 0.6) for both ion modes.

In positive ion mode on average LipidFrag could annotate 69 and LipidBlast 49.7 spectra across all lipid classes. Considering the mean ranks, LipidBlast showed better results for TG (LMGL0301) (1.0 to 3.1) species. No results were annotated for PI spectra as the predictions are missing in the current spectral database mirror of LipidBlast. Developers of LipidBlast indicated that predictions are in progress for several missing lipid classes and will be added to the library in the near future. LipidFrag showed better mean ranks for PE (LMGP0201) (1.7 to 1.8)

Table 3. Comparison of LipidFrag with LipidBlast results.

Negative ion mode								
Mean Rank	TG (LMGL0301)	PC/PE (LMGP0101/LMGP0201)	PC (LMGP0101)	PE (LMGP0201)	PS (LMGP0301)	PG (LMGP0401)	PI (LMGP0601)	Cer (LMSP02/LMSP0202)
LipidFrag	—	2.3 (112)	—	—	1.8 (35)	1.8 (41)	2.4 (62)	1.2 (155)
LipidBlast	—	1.2 (116)	—	—	1.0 (34)	1.0 (40)	2.3 (70)	1.0 (158)
positive ion mode								
Mean Rank	TG (LMGL03)	PC/PE (LMGP0101/LMGP0201)	PC (LMGP01)	PE (LMGP02)	PS (LMGP03)	PG (LMGP04)	PI (LMGP06)	Cer (LMSP0201/LMSP0202)
LipidFrag	3.1 (25)	—	1.7 (13)	1.7 (88)	1.9 (50)	—	1.0 (82)	1.0 (156)
LipidBlast	1.0 (13)	—	1.0 (9)	1.8 (75)	7.8 (43)	—	NA (0)	1.0 (158)

The table shows the mean ranks of the used lipid main/sub classes in the standard data set. The LipidFrag results are calculated by using a FCP threshold of 0.6 (as in Table 1). Besides the mean rankings also the number of annotated spectra are given.

doi:10.1371/journal.pone.0172311.t003

and PS (LMGP0301) (1.9 to 7.8) species. Equal mean ranks for both software tool could be assigned to the Ceramide classes (LMSP0201 and LMSP0202) with a value of 1.0. Both software tools filtered out a large proportion of the PC spectra (LipidFrag: 58 spectra, LipidBlast: 62 spectra) as this lipid class shows sparse fragmentation in positive ion mode resulting in less informative MS/MS spectra.

For negative ion mode LipidFrag and LipidBlast could annotate an almost equal number of MS/MS spectra with mean values of 81 and 83.6 across all lipid classes. LipidBlast performed slightly better the Ceramide (LMSP0201) (1.0 to 1.6) and the PI (LMGP0601) species, whereas LipidFrag showed better mean ranks for PC/PE (LMGP0101 /LMGP0201) (2 to 2.3) and PG (LMGP0401) (1.0 to 1.8) species.

Analysis of *C. elegans* samples

To demonstrate the applicability to biological data, lipids extracted from *C. elegans* were used, representing a realistic challenge for LipidFrag. The composition of the worm lipidome has been extensively reviewed [30]. Several lipid classes are present in the worm and different fatty acid combinations, including odd-numbered side chains, are possible in glycerol- and glycerophospholipids. Shotgun lipidomics was applied for analysis of a novel class of lipids only present in dauer larvae [31].

Coverage of lipids with MS/MS spectra

The total number of lipid features and those with at least one associated MS/MS spectrum are depicted in Fig 4. The green histogram shows all features, while the red one shows features with MS/MS spectra. Due to technical limitations of DDA, only a small fraction of the detected lipids is subjected to fragmentation, a problem well known from proteomics [32].

The DDA method used was able to fragment approximately a quarter (28%) of detected lipids. This number remains surprisingly constant, across different sample sets that have been analyzed with the same analytical method (data not shown). Different parameter settings for inclusion/exclusion lists and exclusion time have been tested. If the exclusion window is set too big, one or several features will be missed due to close elution of different isomeric and isobaric species, if too small, the same peak will be fragmented too often. Fig 4B and 4D show how often each peak with at least one MS/MS spectrum was fragmented across 5 technical replicates. Optimizing the analytical method for a better the coverage of peaks with MS/MS spectra is beyond the scope of this publication. However, 28% coverage corresponds to several thousand spectra (>3000), making the need for an automated analysis tool obvious.

Application of LipidFrag workflow to *C. elegans* MS/MS spectra

LipidFrag then was applied to MS/MS spectra obtained from *C. elegans* lipid extracts. Table 3 gives an overview on detected lipid features in positive and negative ion mode runs. Altogether 1,518 MS/MS spectra acquired in negative and 2,355 MS/MS spectra in positive ion mode were processed. Results with a foreground class probability (FCP) of ≥ 0.95 can be found across the whole intensity range, although higher intensities seem to lead to better results in positive ion mode (Fig 4A). More important than precursor intensity is to detect diagnostic fragments, which especially is the case in negative ion mode, where fatty acyl side chains can be directly detected. Good results in this mode were also obtained for most of the middle intensity range (Fig 4C). Table 4 gives an overview on the number of detected lipid features and their corresponding MS/MS information and LipidFrag results.

For the 3,873 (1,518 + 2,355) *C. elegans* spectra used, the MetFrag *in silico* fragmentation and scoring took altogether ~31 hours (user+system time) on a single core CPU, i.e. 30

5.5 LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

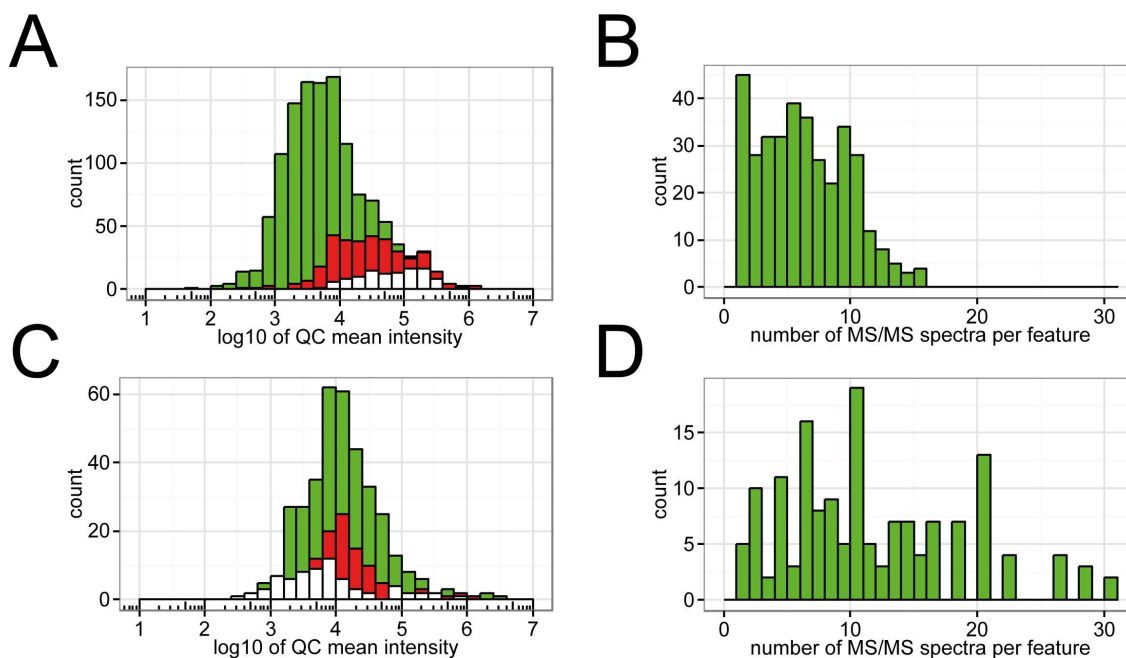


Fig 4. Histogram of intensities of features detected in positive ion mode. (A) The green histogram represents all features, in red are features with one or more associated MS/MS spectra and the white features having a FCP > 0.95 in LipidFrag. (B) Histogram of MS/MS spectra per feature in positive ion mode across all 5 technical replicates. (C) and (D) show the same for negative ion mode.

doi:10.1371/journal.pone.0172311.g004

seconds per spectrum. Using the calculated classifiers, which are based on the standard lipid spectra, the FCP calculation for all 3,873 *C. elegans* spectra took less than 10 minutes, or 0.15 seconds per spectrum.

For the positive ion mode, LipidFrag detected 452 spectra as TG (LMGL0301), 69 as PE (LMGP0201). Additional 3 PE and 1 PC (LMGP0101) species were added by decreasing the FCP threshold to 0.9. In negative ion mode, LipidFrag found 206 spectra with PC/PE (LMGP0101/LMGP0201) lipid sub class annotations having a FCP \geq 0.95. With a lower FCP threshold of 0.9, additional 47 PC/ PE species were annotated. Irrespective of the ion mode over 22% of the LipidFrag results have a FCP \geq 0.75 (S5 Fig).

Table 4. Overview on lipid MS1 features detected in *C. elegans* samples in the two respective ion modes with reliable LipidFrag results.

Ion mode	No. of cluster	With accurate mass annotation	With MS/MS	Manually identified in standards	Reliable LipidFrag (FCP cut-off)
Pos	1655	1297	685	65	<ul style="list-style-type: none"> • 108 (0.7) • 106 (0.8) • 100 (0.9) • 98 (0.95)
Neg	505	358	228	52	<ul style="list-style-type: none"> • 45 (0.7) • 43 (0.8) • 43 (0.9) • 40 (0.95)

doi:10.1371/journal.pone.0172311.t004

Fig 5 shows the spectrum of PE(18:0/20:5). The most prominent peaks show the corresponding fatty acids, with higher intensities for C20:5 bound at the sn2 position. A further diagnostic fragment $[M-sn2-H]^-$ at m/z 480 is detected, and with lower intensities also the $[M-sn1-H]^-$ at m/z 462 ion. Precursor mass together with these four peaks and their respective

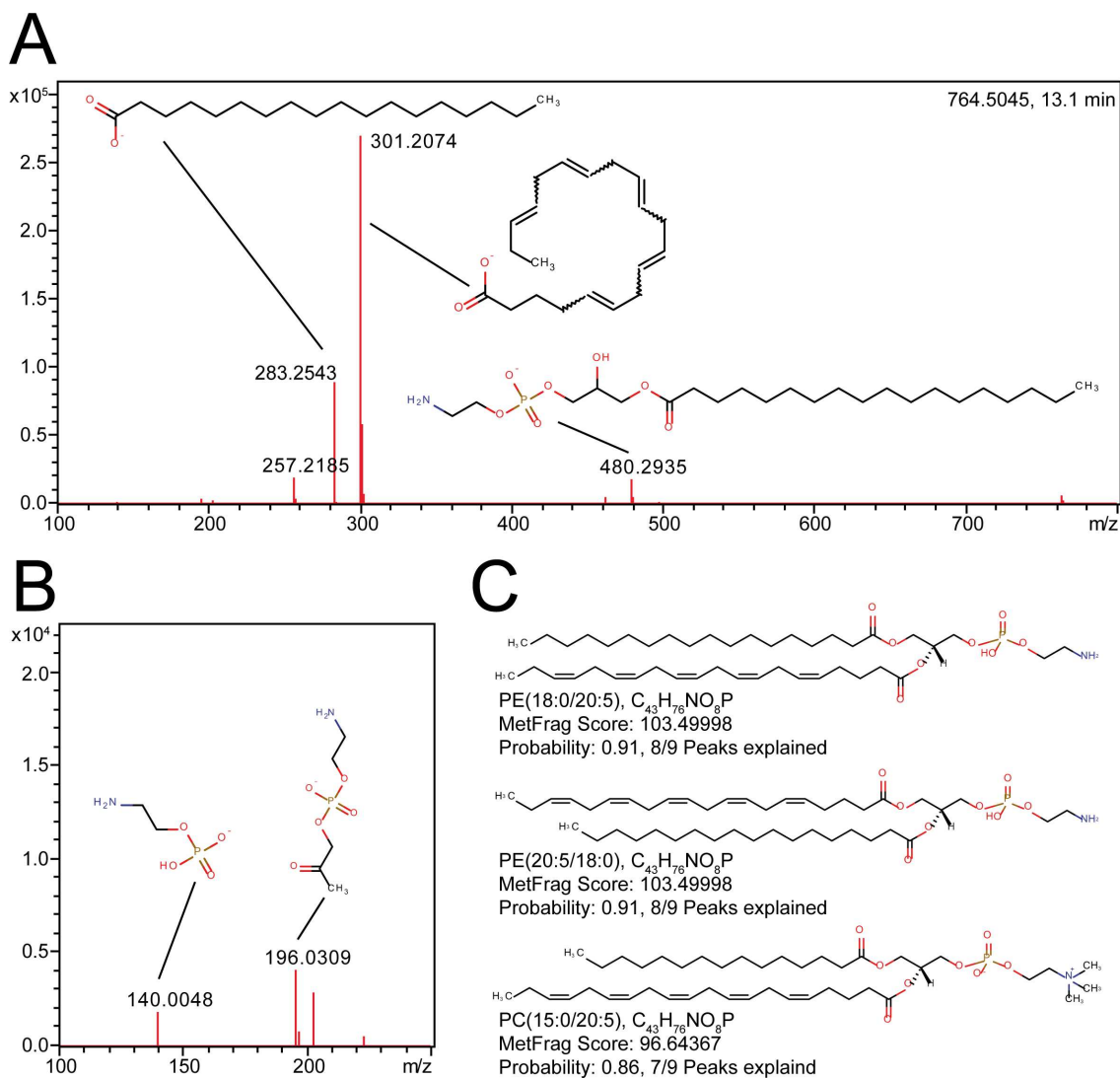


Fig 5. Example of a LipidFrag identification in *C. elegans* data. (A) MS/MS spectrum of m/z 764.5045 at 13.1 minutes detected in *C. elegans* with fragment structures annotated. (B) Close up of lower mass region (m/z 100–250). (C) Structures of the best three candidates obtained from MetFrag with result filtering using foreground class probabilities. Name, formula, MetFrag score and probability are indicated below each structure.

doi:10.1371/journal.pone.0172311.g005

ratios allow manual identification as PE(18:0/20:5). Furthermore, the head group was detected as fragment together with a fragment containing the head group and the glycerol backbone. MetFrag was able to explain 8 of 9 fragments for identification. Additionally, small fragments derived from C20:5 were found. LipidFrag calculated a FCP of 0.91 for the result being a PE. The fatty acid positional isomer showed a similar score and probability. Because the scoring does not take any intensity ratios into account, both isomers obtain the same score. At this point, manual interpretation of intensities is required to determine which annotation is correct. The isobaric PC(15:0/20:5) was ranked third, with a similar MetFrag raw score (103.49998 for the correct PE and 96.64367 for the PC) but a lower FCP of 0.86 and only 7 of 9 peaks correctly explained. Number of explained peaks was used as additional metric for correct identification, in case scores and probabilities are similar. A second sample is depicted in [S6 Fig](#) and in the [S6](#) and [S7 Tables](#). At the respective retention time two PC isomers are coeluting, leading to a superposition of different MS/MS spectra. For the precursor ion two different possible annotations were found by MassTRIX, $[M+HCOO]^-$ or $[M-H]^-$. LipidFrag was able to correctly annotate both isomers using $[M+HCOO]^-$ as precursor with high scores and FCPs (≥ 0.99), where 4 different isomers (two fatty acids and two positional isomers) were found on the first four ranks. Manual identification confirmed the automated results. In addition to the correct PC species, other PC and PE species were annotated due to several possibilities that arise from the merging of two lipids and other minor fatty acid fragments with very low intensities (e.g. C20:2) ([S6 Table](#)). On the other hand, results for the $[M-H]^-$ annotation yielded only low scores and FCPs (< 0.1) ([S7 Table](#)). At the current stage no further details, e.g. on position of double bonds can be given without using specialized analytical approaches [[33](#)].

Both demonstrated that using MetFrag scores alone results can be ambiguous where the addition of the LipidFrag classifiers into the workflow improved automatic annotation results, by removing many false positive results ([S8](#) and [S9 Tables](#)). Remaining spectra were either of low quality, low MetFrag scores or no training data was available due to missing standard materials for respective lipid class (e.g. glycosphingolipids, LMSP05). No classification could be made in the latter case.

Lipids in the used biological samples subjected to fragmentation by DDA were almost exclusively from the class of glycerophospholipids, di- and triacylglycerols. Lower concentration lipids, e.g. ceramides, were masked by these highly abundant classes. Using *C. elegans* lipid extracts it was shown that the developed approach can be applied to biological samples. Coverage of features with one or more associated MS/MS spectra has to be improved, e.g. using pseudo-targeted methods [[34](#)], data independent approaches and spectra reconstruction [[35](#)] or improved DDA [[36](#)]. Lastly, in order to achieve full lipidome coverage, several more classifiers for different lipid classes are needed, but not for all classes lipid standards (e.g. maradolipids detected in dauer larvae) are currently available.

Discussion

Although the number of tools for automatic identification of lipids is increasing, most research still performs manual inspection of MS/MS spectra or automated comparison against rather small reference libraries. *In silico* fragmentation offers an elegant, automatic way to tentatively identify metabolites and lipids if no standard is available, by reducing the number of possible candidates or even propose just a single reliable match. A workflow was developed and validated for analysis of lipid MS/MS spectra derived from data dependent acquisition on a UPLC-Q-ToF-MS system. This workflow is based on annotation of potential lipids to the precursor mass, isotope clean-up of MS/MS spectra and identification using the *in silico* fragmentation tool MetFrag in combination with a novel reliability calculation based on bayesian

classifiers. Lipid standard materials were used for validation purposes and the *in silico* analysis was compared against manual identification.

Cross-validation of the obtained results showed that the true correct identification can be easily separated from background spectra for most cases. Scores of correctly identified lipids and deliberately wrong candidates as decoys were used to generate fore- and background datasets to calculate the FCP giving a reliability of a result of an unknown to be correct. Using lipid standard materials, good performance of LipidFrag was shown, with high relative rankings of the correct candidate, high probabilities and high AUC values obtained from the cross-validation. Furthermore, comparison with LipidBlast, one of the most utilized tool for lipid spectra prediction, showed comparable results for both tools, with the main difference that the LipidFrag approach needs an initial training step for its classifiers but no *ab initio* information on fragmentation compared to LipidBlast. The workflow was applied to a lipid extract of *C. elegans*. From the obtained spectra, about 20% had high foreground class probabilities of ≥ 0.9 . Higher identification rates could be achieved in future investigations by measuring more lipid standards from different classes to train more classifiers. However even with only 11 classifiers, the application of LipidFrag to MS/MS spectra derived from lipid extracts from *C. elegans* was successful and showed the advantage of this workflow.

An advantage here is that MetFrag does not rely on previously known fragmentation pathways and is therefore also applicable to novel lipid classes, currently not present in databases. In this case, candidate structures can be scored by generating potential structures, e.g. using theoretical lipids from LipidHome or even structures from a molecular structure generator like MOLGEN as input database [37].

For the results retrieved from the *C. elegans* data, comparison of the LipidFrag annotation with high probabilities and the manual identification for randomly-selected spectra showed excellent agreement with most of the peaks correctly explained by the *in silico* fragmentation. For application to complete lipidomics studies, the results from LipidFrag can serve as a first filtering and interpretation for further manual investigation, especially for potential marker peaks. A major limitation is co-elution of isomeric species leading to mixed MS/MS spectra. Although the chromatographic method is able to resolve several isomeric lipids as shown previously [23], not all of them can be resolved, especially for lipids like TGs where several isomers exist. Where identified spectra as training data are available, e.g. through authentic standards, LipidFrag can help in high-throughput identification. With the standard MS setups, as employed in this study, lipid class and fatty acid composition can be deduced. Our selected example with the PE(18:0/20:5) species from the biological dataset showed that the MetFrag score alone cannot distinguish ambiguous results. Here, the wrong candidate had a similar score to the correct one, but their FCPs were significantly different and enhanced the annotation confidence. Manual interpretation of obtained data often allows to additionally identify fatty acid position based on intensity ratios of fatty acid fragments, which is not possible with MetFrag.

Result output could be simplified by the lipid annotation scheme proposed by Liebisch et al. [38], which combines different lipid isomers under a common identifier. For mass spectrometry using UHR-Q-ToF-MS, the fatty acid scan level and fatty acid positional isomer are relevant. The former represents lipid identification of the fatty acid composition, but their position is not determined. This level is well suited for LipidFrag identification. For example, all isomeric results can be collapsed under a common identifier, which would be easier to interpret. Unfortunately, the Liebisch annotation is currently not widespread in structural databases. LipidHome is an *in silico* database [5], using this nomenclature, whereas no structural representation of the chemical structure is available, which would be needed for MetFrag. Currently, no chemoinformatics representation exists to encode ambiguity in the position of double bonds [13].

The use of data dependent fragmentation in conjunction with non-targeted studies can further benefit from improved chromatographic methods with increased chromatographic resolution, especially in regions where several lipids co-elute. Different column chemistry, e.g. C30 stationary phase, helps with isomer separation. In the end, a trade-off between resolution, analysis time and throughput has to be found. Here, only one particular extract was used to test the workflow, but in a more extensive study it is likely that more MS/MS spectra from different lipid features would be obtained, based on natural sample inhomogeneity and differences between sample groups. The new approach of all ion fragmentation or data independent acquisition (DIA) offered by most MS vendors can increase the coverage, but tools for deconvolution and reconstruction of MS/MS spectra from this type of acquisition are still very limited today [39, 40]. Additionally, positively identified lipids can be uploaded to general repositories, e.g. MassBank, to improve data distribution.

Here, this method is applied to different *C. elegans* studies and allows comprehensive analysis of the nematodes' lipidome, but is also applicable datasets from different experiments.

Conclusion

Our newly developed workflow LipidFrag improves lipid identification from simple annotation to higher levels of accuracy. It utilizes *in silico* fragmentation of lipid candidate structures. Fragments explained by LipidFrag match known fragmentation pathways, e.g. neutral losses of lipid headgroups and fatty acid side chain fragments. These *in silico* fragmentation results are used to determine reliability scores calculated by bayesian classifiers, which helps to distinguish between true and false annotation results. For training of the classifiers authentic chemical standards from known lipid classes were used. This novel, additional filter step decreases interference from isomeric or isobaric results from different lipid classes having similar MetFrag scores. Extensive cross-validation and application to lipids from *C. elegans* showed its applicability. With inclusion of more and more future available lipid standards identification rates using LipidFrag will increase.

Supporting information

S1 Information. A website <http://msbi.ipb-halle.de/msbi/lipidfrag> has been created to provide additional material for this manuscript. All files are provided for both positive and negative ion mode. The peaklist archives contain the actual MetFrag query files of the standard and *C. elegans* MS/MS spectra. Furthermore, the result files are attached containing the MetFrag identifications and LipidFrag's calculated foreground class probabilities for the *C. elegans* peaklists.
(DOCX)

S1 Fig. Histograms of MetFrag score distributions for positive ion mode. Histograms of back- (red) and foreground (green) datasets with their respective modeled distributions from specific lipid sub-classes.
(TIF)

S2 Fig. Histograms of MetFrag score distributions for negative ion mode. Histograms show back- (red) and foreground (green) datasets with their modeled distributions from specific lipid sub-classes.
(TIF)

S3 Fig. Scatterplots of raw MetFrag scores from lipid standard material MS/MS spectra. The score are shown for negative (A) and positive (B) ion mode.
(TIF)

S4 Fig. MetFrag results from overlapping experiment. Rank as function of different mixtures is shown.

(TIF)

S5 Fig. LipidFrag results on *C. elegans* data. The maximal foreground class probabilities (FCPs) and their histograms calculated by LipidFrag are plotted in descending order for 2,355 MS/MS spectra in positive (A) and 1,518 MS/MS spectra in negative (B) ion mode originating from the *C. elegans* lipid extract.

(TIF)

S6 Fig. LipidFrag annotation example from *C. elegans* dataset. (A) Extracted ion chromatogram of an example lipid and one MS/MS spectrum acquired at 13.11 minutes. Under this peak two isomeric PC species are co-eluting. LipidFrag identified all four isomer (fatty acid isomers and positional isomers) with high scores and probabilities (S6 Table). (B) MS/MS spectrum at 13.11 showing a mixed spectrum of two isomeric PC species.

(TIF)

S1 Table. Target lipids and used interfering species for overlapping experiments.

(PDF)

S2 Table. Statistics on training MS/MS spectra from positive ion mode.

(PDF)

S3 Table. Statistics on training MS/MS spectra from negative ion mode.

(PDF)

S4 Table. Number of used MS/MS spectra for training in positive ion mode.

(PDF)

S5 Table. Number of used MS/MS spectra for training in negative ion mode.

(PDF)

S6 Table. LipidFrag results for *C. elegans* MS/MS spectrum shown in S6 Fig derived from [M+HCOO]⁻ annotation.

(PDF)

S7 Table. LipidFrag results for *C. elegans* MS/MS spectrum shown in S6 Fig derived from [M-H]⁻ annotation.

(PDF)

S8 Table. LipidFrag's improvement of ranks for training MS/MS spectra in positive ion mode.

(PDF)

S9 Table. LipidFrag's improvement of ranks for training MS/MS spectra in negative ion mode.

(PDF)

Acknowledgments

Christoph Ruttkies acknowledges funding from DFG grant NE/1396/5-1. *C. elegans* samples were kindly provided by Steve Garvis, ENS Lyon. We thank Emma Schymanski (Eawag, Dübendorf, Switzerland) for fruitful and intensive discussions and comments on the manuscript.

Author Contributions

Conceptualization: MW SN PS.

Data curation: MW CR.

Formal analysis: MW CR.

Investigation: MW.

Methodology: MW CR SN.

Project administration: SN PS.

Software: MW CR.

Validation: MW CR.

Visualization: MW CR.

Writing – original draft: MW CR SN PS.

Writing – review & editing: MW CR SN PS.

References

1. van Meer G. Cellular lipidomics. *EMBO J.* 2005; 24(18):3159–65. doi: [10.1038/sj.emboj.7600798](https://doi.org/10.1038/sj.emboj.7600798) PMID: [16138081](https://pubmed.ncbi.nlm.nih.gov/16138081/)
2. Caffrey M, Hogan J. LIPIDAT: A database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis. *Chemistry and Physics of Lipids.* 1992; 61(1):1–109. PMID: [1315624](https://pubmed.ncbi.nlm.nih.gov/1315624/)
3. Watanabe K, Yasugi E, Oshima M. How to Search the Glycolipid data in “LIPIDBANK for Web”; the Newly Developed Lipid Database in Japan. *Trends in Glycoscience and Glycotechnology.* 2000; 12(65):175–84.
4. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Research.* 2007; 35(suppl 1):D527–D32.
5. Foster JM, Moreno P, Fabregat A, Hermjakob H, Steinbeck C, Apweiler R, et al. LipidHome: A Database of Theoretical Lipids Optimized for High Throughput Mass Spectrometry Lipidomics. *PLoS ONE.* 2013; 8(5):e61951. doi: [10.1371/journal.pone.0061951](https://doi.org/10.1371/journal.pone.0061951) PMID: [23667450](https://pubmed.ncbi.nlm.nih.gov/23667450/)
6. Aimo L, Liechti R, Hyka-Nouspikel N, Niknejad A, Gleizes A, Götz L, et al. The SwissLipids knowledge-base for lipid biology. *Bioinformatics.* 2015; 31(17):2860–6. doi: [10.1093/bioinformatics/btv285](https://doi.org/10.1093/bioinformatics/btv285) PMID: [25943471](https://pubmed.ncbi.nlm.nih.gov/25943471/)
7. Yang K, Cheng H, Gross RW, Han X. Automated Lipid Identification and Quantification by Multidimensional Mass Spectrometry-Based Shotgun Lipidomics. *Analytical Chemistry.* 2009; 81(11):4356–68. doi: [10.1021/ac900241u](https://doi.org/10.1021/ac900241u) PMID: [19408941](https://pubmed.ncbi.nlm.nih.gov/19408941/)
8. Yang K, Han X. Accurate Quantification of Lipid Species by Electrospray Ionization Mass Spectrometry—Meets a Key Challenge in Lipidomics. *Metabolites.* 2011; 1(1):21. doi: [10.3390/metabo1010021](https://doi.org/10.3390/metabo1010021) PMID: [22905337](https://pubmed.ncbi.nlm.nih.gov/22905337/)
9. Han X, Gross RW. Shotgun lipidomics: Electrospray ionization mass spectrometric analysis and quantification of cellular lipidomes directly from crude extracts of biological samples. *Mass Spectrometry Reviews.* 2005; 24(3):367–412.
10. Almeida R, Pauling J, Sokol E, Hannibal-Bach H, Ejsing C. Comprehensive Lipidome Analysis by Shotgun Lipidomics on a Hybrid Quadrupole-Orbitrap-Linear Ion Trap Mass Spectrometer. *J Am Soc Mass Spectrom.* 2015; 26(1):133–48. doi: [10.1007/s13361-014-1013-x](https://doi.org/10.1007/s13361-014-1013-x) PMID: [25391725](https://pubmed.ncbi.nlm.nih.gov/25391725/)
11. Hu C, van Dommelen J, van der Heijden R, Spijksma G, Reijmers TH, Wang M, et al. RPLC-Ion-Trap-FTMS Method for Lipid Profiling of Plasma: Method Validation and Application to p53 Mutant Mouse Model. *Journal of Proteome Research.* 2008; 7(11):4982–91. doi: [10.1021/pr800373m](https://doi.org/10.1021/pr800373m) PMID: [18841877](https://pubmed.ncbi.nlm.nih.gov/18841877/)
12. Bird SS, Marur VR, Stavrovskaya IG, Kristal BS. Separation of Cis–Trans Phospholipid Isomers Using Reversed Phase LC with High Resolution MS Detection. *Analytical Chemistry.* 2012; 84(13):5509–17. doi: [10.1021/ac300953j](https://doi.org/10.1021/ac300953j) PMID: [22656324](https://pubmed.ncbi.nlm.nih.gov/22656324/)

13. Ishida M, Yamazaki T, Houjou T, Imagawa M, Harada A, Inoue K, et al. High-resolution analysis by nano-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for the identification of molecular species of phospholipids and their oxidized metabolites. *Rapid Communications in Mass Spectrometry*. 2004; 18(20):2486–94. doi: [10.1002/rcm.1650](https://doi.org/10.1002/rcm.1650) PMID: [15384179](https://pubmed.ncbi.nlm.nih.gov/15384179/)
14. Ekroos K, Ejsing CS, Bahr U, Karas M, Simons K, Shevchenko A. Charting molecular composition of phosphatidylcholines by fatty acid scanning and ion trap MS3 fragmentation. *Journal of Lipid Research*. 2003; 44(11):2181–92. doi: [10.1194/jlr.D300020-JLR200](https://doi.org/10.1194/jlr.D300020-JLR200) PMID: [12923235](https://pubmed.ncbi.nlm.nih.gov/12923235/)
15. Schwudke D, Liebisch G, Herzog R, Schmitz G, Shevchenko A. Shotgun Lipidomics by Tandem Mass Spectrometry under Data-Dependent Acquisition Control. In: Brown HA, editor. *Methods in Enzymology*. Volume 433: Academic Press; 2007. p. 175–91.
16. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*. 2010; 45(7):703–14. doi: [10.1002/jms.1777](https://doi.org/10.1002/jms.1777) PMID: [20623627](https://pubmed.ncbi.nlm.nih.gov/20623627/)
17. Wolf S, Schmidt S, Muller-Hannemann M, Neumann S. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*. 2010; 11(1):148.
18. Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast *in silico* tandem mass spectrometry database for lipid identification. *Nat Meth*. 2013; 10(8):755–8. <http://www.nature.com/nmeth/journal/v10/n8/abs/nmeth.2551.html#supplementary-information>.
19. Kochen MA, Chambers MC, Holman JD, Nesvizhskii AI, Weintraub ST, Belisle JT, et al. Greazy: Open-Source Software for Automated Phospholipid Tandem Mass Spectrometry Identification. *Analytical Chemistry*. 2016; 88(11):5733–41. doi: [10.1021/acs.analchem.6b00021](https://doi.org/10.1021/acs.analchem.6b00021) PMID: [27186799](https://pubmed.ncbi.nlm.nih.gov/27186799/)
20. Suhre K, Schmitt-Kopplin P. MassTRIX: mass translator into pathways. *Nucleic Acids Research*. 2008; 36(suppl 2):W481–W4.
21. Wägele B, Witting M, Schmitt-Kopplin P, Suhre K. MassTRIX Reloaded: Combined Analysis and Visualization of Transcriptome and Metabolome Data. *PLoS One*. 2012; 7(7):e39860. doi: [10.1371/journal.pone.0039860](https://doi.org/10.1371/journal.pone.0039860) PMID: [22815716](https://pubmed.ncbi.nlm.nih.gov/22815716/)
22. Witting M, Schmitt-Kopplin P. Chapter 17—Transcriptome and Metabolome Data Integration—Technical Prerequisites for Successful Data Fusion and Visualization. In: Carolina Simó AC, Virginia G-C, editors. *Comprehensive Analytical Chemistry*. Volume 63: Elsevier; 2014. p. 421–42.
23. Witting M, Maier TV, Garvis S, Schmitt-Kopplin P. Optimizing a ultrahigh pressure liquid chromatography-time of flight-mass spectrometry approach using a novel sub-2µm core-shell particle for in depth lipidomic profiling of *Caenorhabditis elegans*. *Journal of Chromatography A*. 2014; 1359(0):91–9. doi: [10.1016/j.chroma.2014.07.021](https://doi.org/10.1016/j.chroma.2014.07.021) PMID: [25074420](https://pubmed.ncbi.nlm.nih.gov/25074420/)
24. Matyash V, Liebisch G, Kurzchalia TV, Shevchenko A, Schwudke D. Lipid extraction by methyl-tert-butyl ether for high-throughput lipidomics. *Journal of Lipid Research*. 2008; 49(5):1137–46. doi: [10.1194/jlr.D700041-JLR200](https://doi.org/10.1194/jlr.D700041-JLR200) PMID: [18281723](https://pubmed.ncbi.nlm.nih.gov/18281723/)
25. Käll L, Storey JD, MacCoss MJ, Noble WS. Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin. *Journal of Proteome Research*. 2007; 7(1):40–4. doi: [10.1021/pr700739d](https://doi.org/10.1021/pr700739d) PMID: [18052118](https://pubmed.ncbi.nlm.nih.gov/18052118/)
26. Hsu F-F, Turk J. Characterization of ceramides by low energy collisional-activated dissociation tandem mass spectrometry with negative-ion electrospray ionization. *J Am Soc Mass Spectrom*. 2002; 13(5):558–70. doi: [10.1016/S1044-0305\(02\)00358-6](https://doi.org/10.1016/S1044-0305(02)00358-6) PMID: [12019979](https://pubmed.ncbi.nlm.nih.gov/12019979/)
27. Hsu F-F, Turk J. Charge-driven fragmentation processes in diacyl glycerophosphatidic acids upon low-energy collisional activation. A mechanistic proposal. *J Am Soc Mass Spectrom*. 2000; 11(9):797–803. doi: [10.1016/S1044-0305\(00\)00151-3](https://doi.org/10.1016/S1044-0305(00)00151-3) PMID: [10976887](https://pubmed.ncbi.nlm.nih.gov/10976887/)
28. Murphy RC, James PF, McAnoy AM, Krank J, Duchoslav E, Barkley RM. Detection of the abundance of diacylglycerol and triacylglycerol molecular species in cells using neutral loss mass spectrometry. *Analytical Biochemistry*. 2007; 366(1):59–70. doi: [10.1016/j.ab.2007.03.012](https://doi.org/10.1016/j.ab.2007.03.012) PMID: [17442253](https://pubmed.ncbi.nlm.nih.gov/17442253/)
29. <http://bio.massbank.jp/> 2015.
30. Witting M, Schmitt-Kopplin P. The *Caenorhabditis elegans* lipidome: A primer for lipid analysis in *Caenorhabditis elegans*. *Archives of Biochemistry and Biophysics*. 2016; 589:27–37. doi: [10.1016/j.abb.2015.06.003](https://doi.org/10.1016/j.abb.2015.06.003) PMID: [26072113](https://pubmed.ncbi.nlm.nih.gov/26072113/)
31. Papan C, Penkov S, Herzog R, Thiele C, Kurzchalia T, Shevchenko A. Systematic Screening for Novel Lipids by Shotgun Lipidomics. *Analytical Chemistry*. 2014; 86(5):2703–10. doi: [10.1021/ac404083u](https://doi.org/10.1021/ac404083u) PMID: [24471557](https://pubmed.ncbi.nlm.nih.gov/24471557/)
32. Michalski A, Cox J, Mann M. More than 100,000 Detectable Peptide Species Elute in Single Shotgun Proteomics Runs but the Majority is Inaccessible to Data-Dependent LC-MS/MS. *Journal of Proteome Research*. 2011; 10(4):1785–93. doi: [10.1021/pr101060v](https://doi.org/10.1021/pr101060v) PMID: [21309581](https://pubmed.ncbi.nlm.nih.gov/21309581/)

5.5 LipidFrag: Improving reliability of *in silico* fragmentation of lipids and application to the *Caenorhabditis elegans* lipidome

33. Pham HT, Maccarone AT, Thomas MC, Campbell JL, Mitchell TW, Blanksby SJ. Structural characterization of glycerophospholipids by combinations of ozone- and collision-induced dissociation mass spectrometry: the next step towards "top-down" lipidomics. *Analyst*. 2014; 139(1):204–14. doi: [10.1039/c3an01712e](https://doi.org/10.1039/c3an01712e) PMID: [24244938](https://pubmed.ncbi.nlm.nih.gov/24244938/)
34. Luo P, Dai W, Yin P, Zeng Z, Kong H, Zhou L, et al. Multiple Reaction Monitoring-Ion Pair Finder: A Systematic Approach To Transform Nontargeted Mode to Pseudotargeted Mode for Metabolomics Study Based on Liquid Chromatography–Mass Spectrometry. *Analytical Chemistry*. 2015; 87(10):5050–5. doi: [10.1021/acs.analchem.5b00615](https://doi.org/10.1021/acs.analchem.5b00615) PMID: [25884293](https://pubmed.ncbi.nlm.nih.gov/25884293/)
35. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Meth*. 2015; 12(6):523–6. <http://www.nature.com/nmeth/journal/v12/n6/abs/nmeth.3393.html#supplementary-information>.
36. Yan Z, Yan R. Improved Data-Dependent Acquisition for Untargeted Metabolomics Using Gas-Phase Fractionation with Staggered Mass Range. *Analytical Chemistry*. 2015; 87(5):2861–8. doi: [10.1021/ac504325x](https://doi.org/10.1021/ac504325x) PMID: [25654645](https://pubmed.ncbi.nlm.nih.gov/25654645/)
37. Gruner T, Kerber A, Laue R, Meringer M. MOLGEN 4.0. *MATCH Commun Math Comput Chem*. 1998; 37:205–8.
38. Liebisch G, Vizcaino JA, Köfeler H, Trötz Müller M, Griffiths WJ, Schmitz G, et al. Shorthand notation for lipid structures derived from mass spectrometry. *Journal of Lipid Research*. 2013; 54(6):1523–30. doi: [10.1194/jlr.M033506](https://doi.org/10.1194/jlr.M033506) PMID: [23549332](https://pubmed.ncbi.nlm.nih.gov/23549332/)
39. Ahmed Z, Mayr M, Zeeshan S, Dandekar T, Mueller MJ, Fekete A. Lipid-Pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms. *Bioinformatics*. 2014.
40. Castro-Perez JM, Kamphorst J, DeGroot J, Lafeber F, Goshawk J, Yu K, et al. Comprehensive LC–MSE Lipidomic Analysis using a Shotgun Approach and Its Application to Biomarker Detection and Identification in Osteoarthritis Patients. *Journal of Proteome Research*. 2010; 9(5):2377–89. doi: [10.1021/pr901094j](https://doi.org/10.1021/pr901094j) PMID: [20355720](https://pubmed.ncbi.nlm.nih.gov/20355720/)

5.6 Tackling CASMI 2012: Solutions from MetFrag and MetFusion

Christoph Ruttkies, Michael Gerlich, Steffen Neumann. Tackling CASMI 2012: Solutions from MetFrag and MetFusion. *Metabolites*. 2013;3(3):623–636. 2013. 7 Citations⁶
<https://www.mdpi.com/2218-1989/3/3/623/htm>

Contributions

I designed the study together with Michael Gerlich. Steffen Neuman prepared the provided datasets. I prepared the submission results for MetFrag. Michael Gerlich prepared the submission results for MetFusion. I prepared the manuscript together with Michael Gerlich.

⁶<https://scholar.google.com> (accessed on 01/2021)

Metabolites **2013**, *3*, 623-636; doi:10.3390/metabo3030623

OPEN ACCESS

metabolites

ISSN 2218-1989

www.mdpi.com/journal/metabolites

Article

Tackling CASMI 2012: Solutions from MetFrag and MetFusion

Christoph Ruttkies ^{*,†}, Michael Gerlich ^{*,†} and Steffen Neumann

Leibniz Institute of Plant Biochemistry, Department of Stress and Developmental Biology, Weinberg 3, DE-06120 Halle (Saale), Germany; E-Mail: sneumann@ipb-halle.de

[†] These authors contributed equally to this work.

* Authors to whom correspondence should be addressed; E-Mails: cruttkie@ipb-halle.de (C.R.); mgerlich@ipb-halle.de (M.G.); Tel.: +49-345-5582-1471 (C.R.); Fax: +49-345-5582-1409 (C.R.).

Received: 24 April 2013; in revised form: 29 July 2013 / Accepted: 30 July 2013 /

Published: 5 August 2013

Abstract: The task in the critical assessment of small molecule identification (CASMI) contest category 2 was to determine the identification of (initially) unknown compounds for which high-resolution tandem mass spectra were published. We focused on computer-assisted methods that tried to correctly identify the compound automatically and entered the contest with MetFrag and MetFusion to score candidate structures retrieved from the PubChem structure database. MetFrag was combined with the metabolite-likeness score, which helped to improve the performance for the natural product challenges. We present the results, discuss the performance, and give details of how to interpret the MetFrag and MetFusion output.

Keywords: mass spectrometry; metabolite identification; MetFrag; MetFusion; metabolite likeness; molecular formula

1. Introduction

The critical assessment of small molecule identification contest (CASMI) was organised in 2012 by Emma Schymanski and Steffen Neumann, to call upon the computational mass spectrometry community and demonstrate the performance of compound identification from mass spectrometry data.

At the Leibniz Institute of Plant Biochemistry (IPB), we are developing several tools for metabolite identification. The MetFrag system [1] is able to perform *in silico* fragmentation of candidate structures, which can be retrieved from compound databases or obtained through structure generation [2]. The

IPB is also part of the MassBank consortium [4], which collects a large number of reference spectra, particularly of soft electrospray ionisation (ESI) spectra. Our MetFusion system [5] integrates these two strategies to obtain a more reliable identification compared to each individual approach taken alone.

In the CASMI contest, our tools did not officially take part because one author was in the organisation team and some of the challenge spectra were obtained at the IPB. Nevertheless, we tried to approach the challenges in as unbiased a manner as possible, and did not use our inside knowledge to tune any parameters in order to obtain better results. We also restricted the participation to category 2 (“best structure identification for high resolution liquid chromatography/mass spectrometry (LC/MS) data”) and did not submit the molecular formulas to category 1 (“best molecular formula for high resolution LC/MS data”).

2. Methods

The spectra preprocessing steps and the elimination of redundant candidate structures are the same for both MetFrag and MetFusion.

2.1. Spectra Processing and Neutral Mass Heuristics

All of the challenges were measured in a single ionization mode, but with multiple ionization energies. If a challenge provided two or more spectra, the spectra were merged to create a corresponding composite spectrum. This processing step was recommended by the MassBank consortium [4] for a more reliable identification. Challenges 2, 10 and 12 each consisted of only one spectrum, so the spectra merging was not applied to them. We used the `mzClust` grouping algorithm in `xcms` (version 1.37.0) [6,7]. The composite spectrum contains the unique peaks where m/z values are averaged and the maximum intensity across all spectra is used. The R-code for the merging is shown in Appendix B.

To determine the neutral mass of a compound, we used a simple heuristic which located the lowest m/z in the isotope pattern as a monoisotopic peak and then removed the adduct, taking the polarity of the measurement into account to automatically deduce the neutral exact mass of the compounds for the candidate search.

2.2. Eliminating Redundant Candidates

Both MetFrag and MetFusion obtain candidate structures from chemical databases. They often contain redundant structures which increase the candidate lists without adding chemical diversity. In addition, mass spectrometry can, in general, not distinguish between the stereoisomers of a compound and the identification methods we use assign identical scores to isomers. Therefore, we eliminate redundant candidate structures with an InChIKey-based filtering.

The InChIKey is a string that is characteristic of the molecular structure, where the first block of 14 characters is determined by the molecular skeleton (or connectivity). More information regarding both InChI and InChIKey can be found elsewhere [8]. We calculate the InChIKey for each candidate and keep only candidates with a unique first InChIKey block.

2.3. *In silico* Fragmentation with MetFrag

We used MetFrag as described in Wolf *et al.* [1], with the composite spectra as explained in Section 2.1 to submit candidates for all challenges in CASMI category 2. We queried a local PubChem [3] mirror (created September 2010) for the candidate retrieval and filtered as explained in Section 2.2. For the candidate selection we used the putative neutral exact mass and a mass window of 5 ppm and 0.001 Da mass deviation for the fragment matching. For later resubmissions for Challenge 5, we adapted the mass window to 10 ppm and 0.002 Da for, due to the higher mass error. For this paper, we additionally used a molecular formula candidate search using the correct formulas which were not known during the contest but given in the solutions. This allows estimation of the MetFrag performance the correct molecular formulas are used as input.

The score calculated by MetFrag evaluates the match of *in silico*-generated fragments of the candidate molecules to the given challenge tandem mass spectra. The mass as well as the intensity of the peak matched by a fragment are considered in the score.

Compounds for challenges 1 to 6 were known to be natural products, as explained on the CASMI website. Because large compound databases, such as PubChem [9], contain many non-natural compounds, several filtering strategies have been developed for metabolomics data. While Kind and Fiehn [10] proposed filter criteria based on the molecular formula, Peironcely *et al.* [11] used machine learning to train a random forest model [12] on metabolite structures from the Human Metabolome Database (HMDB) [13] and structures from the ZINC database [14] to predict a metabolite-likeness score (MLS) based on structural fingerprints.

We used the MLS to prefer biological compounds for challenges 1–6. For those challenges, we used the adapted version of the final score:

$$Score_{final} = Score_{MetFrag} + \omega \cdot MLS$$

to obtain the ranking, where ω represents the weight of the MLS which we arbitrarily set to 0.5 to give it a lower influence in the final score than the MetFrag score. In the future we plan to optimise ω by learning from given data. The influence of the metabolite-likeness score on the rankings of candidates was investigated by comparing the rankings of results with $\omega = 0$ and $\omega = 0.5$.

2.4. MetFusion: Integration of MetFrag with Spectral Libraries

We also applied MetFusion [5] to generate submissions for all Category 2 challenges. We used the MassBank spectral library and PubChem compound database, which in this case was queried online in January and March 2013. For the candidate selection we used the putative neutral mass and a mass window of 10 ppm. A mass window of 10 ppm is sufficient as all Category 2 challenges promise an accuracy of <10 ppm. For the fragment matching, we applied a window of 0.002 Da and 10 ppm. As explained above, we used composite query spectra and the InChIKey-based candidate filtering.

MassBank provides separate search forms for either a precursor mass search or peak list search. The combination of both types of information is currently not available, although it would be possible to search MassBank with both an MS/MS spectrum and explicitly apply the precursor neutral mass as filter afterwards. This search strategy is used by, e.g., the Metlin database. Instead, MetFusion

invokes the peak list search, so MassBank will also return compounds with similar MS/MS spectra in order to possibly return also structurally similar compounds. MetFusion then implicitly combines the fragmentation similarity from MassBank with the exact mass hit from PubChem.

All challenges were queried against all available ESI spectra in MassBank [4]. For the resubmissions, we also included instruments with ion sources at atmospheric-pressure levels, namely chemical ionization (APCI) and photoionization (APPI). This instrument selection covers triple quadrupole (QqQ), quadrupole time-of-flight (QTOF) and Orbitrap devices i.e., both nominal and accurate mass spectra were queried.

Besides the peak list and instrument selection, the number of result hits and the intensity cut-off are the only parameters for the MassBank peak search. The result limit was set to 100 hits and the intensity cut-off was set to 5. The intensity cut-off determines which peaks are ignored due to having a lower intensity than the specified cut-off. MassBank internally applies a fixed 0.3 Da mass window when matching peaks. MassBank also utilizes the intensity information for spectra comparison, i.e., low intensity peaks have less weight in the resulting scores.

For the MassBank query results, we also performed an InChIKey-based filtering where among the duplicates only the entry with the best MassBank score, i.e., the highest spectral similarity, was kept. The MetFusion workflow and the scoring have been described earlier [5].

In the next section we also discuss the chemical similarity, e.g., between the correct solution and the most similar MassBank record. We used the Tanimoto similarity based on the fingerprints of the structures as implemented in the CDK [15]. A Tanimoto score of 0 indicates that no structural features are shared in both structures. Conversely, a Tanimoto score of 1 indicates that all investigated structural features (determined by the fingerprint) are present in both structures. A Tanimoto score ≥ 0.8 indicates reasonable structural similarity, whereas scores ≥ 0.95 indicate very high structural similarity.

The whole set of challenges was processed with the command line version of MetFusion. Results were stored in a structure data file (SDF), which is better known by the *.sdf file extension. This file keeps the molecular structure and associated information, like compound name, score, and additional properties, for each candidate. In addition to the integrated result list as an SD file, we also keep the individual intermediate result lists and create a spreadsheet file containing the result lists and the coloured similarity matrices which can be used to examine the results in more detail.

3. Results and Discussion

In this section we discuss the results of our resubmissions and note where and why they differ from the original submissions. The challenges 2, 4, 5 and 6 from category 2 were not calibrated when initially offered to the participants, resulting in higher than stated ppm deviations. This was recognised after the contest closed, and the data of these challenges was recalibrated and made available to the participants online for the articles in the proceedings. Each participant was allowed to resubmit their findings. Additionally, our hypotheses for the neutral mass of challenges 11 and 12 were wrong in the first submission. The correct neutral mass for challenge 12 could be extracted from the available meta-data that all participants had access to. Challenge 11 did not provide $[M+H]^+$ ions, instead the $[M-H_2O]^+$ fragment was the major ion suitable for back-tracking the neutral mass by an experienced

mass spectrometrist. We used the correct neutral mass from the published CASMI solution for challenge 11.

For both MetFrag and MetFusion we report the number of candidates and the absolute rank for each challenge, and the median rank broken down to the natural compound and environmental challenges. The median is used because the distribution of ranks is heavily tailed and a few challenges with very poor ranking severely skew the mean values. In addition to the absolute rank, we also report the relative ranking position (RRP_{CASMI}), defined as $RRP_{CASMI} = \frac{1}{2} \left(1 - \frac{BC-WC}{TC-1} \right)$ where BC and WC are the number of candidates ranked better and worse than the correct solution, and TC is the number of total candidates, respectively. See [16] for more details.

3.1. MetFrag

In the initial submission, the correct solution was missing for Challenges 2, 4 and 6 because the measured mass was outside the 5 ppm margin. In addition, the simple precursor heuristics described in Section 2.1 missed the neutral mass of challenges 11 and 12. These cases were corrected with the updated information for the resubmissions.

Table 1 shows the number of candidates obtained from the PubChem snapshot with a search for the neutral mass and the absolute rank of the correct solution. For Challenges 1 to 6 we also show the ranks with the MLS score included.

Table 1. MetFrag results with neutral exact mass filter after resubmission. Shown are the number of candidates per challenge (#Cand.), the InChiKey filtered MetFrag rank and the relative ranking position (RRP). Additionally, for challenges 1–6 the InChiKey filtered MetFrag rank with the metabolite-likeness score (MLS) included is shown.

Natural Product Challenges						Environmental Challenges			
Chall.	#Cand.	Rank	RRP	MLS	RRP	Chall.	#Cand.	Rank	RRP
						10	447	260	0.441
1	994	5	0.996	4	0.997	11	465	23	0.976
2	248	3	0.992	3	0.992	12	1531	36	0.978
3	1094	12	0.990	9	0.993	13	1031	5	0.998
4	2234	547	0.757	454	0.797	14	125	27	0.810
5	2891	988	0.679	1238	0.573	15	1825	173	0.907
6	1860	1860	0.439	281	0.850	16	1948	1948	0.453
						17	475	15	0.970
Median	1477	280	0.874	145	0.921		753	32	0.939

The results achieved with the molecular formula database query are shown in Table A1. For every challenge MetFrag found the correct hit among the candidates with both types of queries, where the mass window result sets contain twice as many candidates. The absolute ranks obtained with the formula query decrease the median rank (Challenges 1–6: 280⇒270; Challenges 10–17: 32⇒22.5) compared to the

ranks of the mass query, but on the other hand the median RRP is lower (Challenges 1–6: $0.874 \Rightarrow 0.607$; Challenges 10–17: $0.939 \Rightarrow 0.917$) with the use of the molecular formula filter, because compounds within the mass search window but with the wrong molecular formula often obtain a lower MetFrag score compared to the correct solution. The molecular formula filter eliminates these worse candidates (*WC*) from the outset, which reduces the RRP.

Next, we describe the outcome if the metabolite-likeness score is considered together with the MetFrag score for the Challenges 1–6. The number of candidates remains unchanged, but natural compounds (including the correct solution) should obtain better scores and improve both the absolute rank and the RRP.

Indeed, except for Challenge 5 all ranks are better or equal with the MLS contribution in the score as shown in Table 1. The median absolute rank decreases from $280 \Rightarrow 145$ (RRP: $0.874 \Rightarrow 0.921$) and even more for the molecular formula candidate search, where the median rank improves from $270 \Rightarrow 119$ (RRP: $0.607 \Rightarrow 0.797$).

Reticuline (the correct candidate of Challenge 5) has the lowest metabolite-likeness score of 0.296 among all challenge compounds and therewith the worst rank (1209) related solely to the MLS (see Table 2), which explains why the final result for Reticuline was even worse with MLS.

Table 2. The metabolite-likeness score (MLS) of the compounds of Challenges 1 – 6 and their rankings among the retrieved candidates based on the MLS alone, while Table 1 uses the combined score.

Challenge	Trivial name	InChIKey (first block)	MLS	MLS rank
1	Kanamycin A	SBUJHOSQTJFQJX	0.508	47
2	1,2-Bis-O-sinapoyl-beta-D-glucoside	KQDOTXAUJBODDM	0.716	35
3	Glucosquerellin	ZAKICGFSIJSCSF	0.474	3
4	Escholtzine	PGINMPJZCWDQNT	0.436	439
5	Reticuline	BHLYRWXGMIUIHG	0.296	1209
6	Rhoeadine	XRBIHOLQAKITPP	0.374	132

Challenges 6 and 16 were very problematic for MetFrag, which could only assign to the given spectrum a single fragment of the correct molecule for the first case and no fragments of the correct molecule for the second case. Although the MLS improved the final rank for challenge 6, this is only based on the (second lowest among all challenges) MLS of 0.374. Figure A1 shows the rankings related to the calculated scores of all candidates of challenges 1 to 6.

The results show that MetFrag is able to rank four molecules of the total 14 challenges among the top ten hits when applying mass filtering. The number can be increased to five by including knowledge of the molecular formula of the correct compound.

The external participants Dunn *et al.* [17] and the internal participant Meringer *et al.* [19] both used MetFrag in conjunction with other methods for the identification. The combined MetFrag and manual interpretation method of Dunn *et al.* had better ranks than MetFrag alone, but missed a lot more challenges because the Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] was used for candidate retrieval, which only contains a subset of the challenge compounds.

3.2. MetFusion

The overall results for MetFusion are shown in Table 3. PubChem has grown considerably over the past two years and consequently the online query against PubChem yields more candidates: for the first six challenges, MetFrag retrieved 1477 candidates (median) from our PubChem snapshot (September 2010), whereas the corresponding online query against PubChem from January 2013 yields 3582 candidates (median)— more than twice as many, and more than three times for the environmental challenges. The same observation can be made for the remaining challenges 10–17. The rapid growth of PubChem over even short time periods becomes obvious; e.g., for Kanamycin A. In January 2013, 37 isomers with an identical first block of their InChIKey were retrieved, whereas only eight weeks later three additional isomers were found.

Table 3. MetFusion results per challenge after resubmission. Shown are number of candidates per challenge (#Cand.), the InChIKey filtered MetFusion rank as well as the maximum Tanimoto similarity (Max. TS) between the candidates and the MassBank results and finally the relative ranking position (RRP).

Natural Product Challenges					Environmental Challenges				
Chall.	#Cand.	Rank	Max. TS	RRP	Chall.	#Cand.	Rank	Max. TS	RRP
					10	1085	981	0.40	0.096
1	2229	1	1.0	1.0	11	1444	170	0.28	0.883
2	625	4	0.93	0.995	12	3772	136	0.28	0.964
3	2945	14	0.99	0.995	13	3344	1	1.0	1.0
4	4219	74	0.84	0.983	14	507	3	1.0	0.996
5	4280	1426	0.42	0.667	15	3394	1	1.0	1.0
6	6175	25	0.79	0.996	16	4427	1351	0.33	0.695
					17	1848	88	0.35	0.953
Median	3582	20	0.89	0.995		2596	112	0.38	0.959

The results for challenges 1 to 6 and challenges 10 to 17 show that more similar spectra are present in MassBank for the biological compounds than for the environmental challenges. The median Tanimoto similarity between the challenges and the most similar compound in MassBank is 0.89 for the natural compounds, compared to 0.38 for the environmental challenges where the reference spectra did not contribute significantly to the integrated MetFusion score in five cases. This can be attributed to a much larger chemical diversity of natural products in MassBank. This is also evident by the low maximum spectral similarity. The lack of reference spectra for diverse non-biological compounds is the major reason for the mediocre performance of MetFusion in these cases. We expect a considerable improvement in this area as contributions to MassBank from the environmental community have recently increased.

In addition to the ranked list of candidates, MetFusion also creates a ranked similarity matrix, where the columns correspond to the result list from MassBank (best hits on the left, ordered by the MassBank score) and the rows correspond to the MetFrag results. Each cell contains the Tanimoto similarity (TS)

of the corresponding structures from MassBank and MetFrag. Examples are shown in Figures 1 and 2. Tanimoto similarities are also visualised through a colour code ranging from red via yellow to green with increasing TS.

Figure 1. The top-left part of the reranked similarity matrix from MetFusion for Challenge 6. The correct compound rheadine is ranked 25th (CID 5318652) and is highlighted with a green border. The maximum Tanimoto similarity (TS) for rheadine has bicuculline with a similarity of 0.79, but a MassBank score of only 0.3 (data not shown). There are other alkaloids with better similarity that are thus ranked higher. Six columns were removed for better readability, altogether with a low maximum TS of 0.4.

	KOX00837	KO008812	WA001623	BML00811	CO000309		BML00783	BML00840	ZMS00126	EA280410	BML00613
44483244	0.388	0.636	0.427	0.391	0.924		0.577	0.410	0.310	0.327	1.000
11717916	0.401	0.640	0.424	0.390	0.973		0.564	0.408	0.311	0.327	0.898
5316069	0.390	0.625	0.420	0.386	0.904		0.577	0.407	0.311	0.333	0.972
68331626	0.394	0.624	0.420	0.388	0.920		0.571	0.402	0.307	0.331	0.919
7348779	0.385	0.643	0.432	0.398	0.882		0.569	0.415	0.300	0.322	0.954
11731734	0.386	0.643	0.430	0.397	0.879		0.567	0.414	0.301	0.323	0.950
18728255	0.413	0.613	0.447	0.381	0.876		0.576	0.405	0.317	0.330	0.901
59991416	0.417	0.595	0.435	0.377	0.900		0.559	0.400	0.313	0.332	0.831
371260	0.389	0.873	0.318	0.431	0.674		0.551	0.430	0.301	0.303	0.617
21763791	0.397	0.617	0.415	0.391	0.840		0.560	0.410	0.309	0.339	0.903
68152375	0.393	0.563	0.415	0.385	0.788		0.575	0.406	0.319	0.350	0.842
68131382	0.403	0.640	0.372	0.394	0.805		0.570	0.415	0.311	0.328	0.853
10905079	0.372	0.728	0.368	0.414	0.724		0.593	0.426	0.307	0.317	0.777
21589025	0.372	0.653	0.352	0.397	0.680		0.561	0.413	0.331	0.334	0.684
601054	0.384	0.693	0.352	0.381	0.651		0.538	0.395	0.319	0.329	0.623
605862	0.386	0.774	0.319	0.406	0.700		0.576	0.404	0.295	0.302	0.656
21768980	0.398	0.598	0.317	0.366	0.582		0.514	0.380	0.335	0.326	0.588
131593	0.402	0.674	0.324	0.393	0.663		0.549	0.392	0.326	0.309	0.607
68152387	0.396	0.576	0.375	0.353	0.738		0.534	0.379	0.323	0.345	0.756
44559282	0.378	0.636	0.342	0.386	0.662		0.554	0.407	0.329	0.348	0.661
57581018	0.386	0.742	0.321	0.383	0.619		0.553	0.388	0.301	0.300	0.627
11058079	0.385	0.627	0.345	0.382	0.662		0.561	0.403	0.334	0.348	0.667
5315436	0.390	0.669	0.328	0.393	0.667		0.551	0.392	0.323	0.304	0.625
13875892	0.376	0.618	0.347	0.396	0.653		0.540	0.416	0.340	0.340	0.656
5318652	0.415	0.587	0.377	0.369	0.732		0.537	0.392	0.307	0.335	0.772
337868	0.392	0.657	0.360	0.372	0.682		0.518	0.399	0.333	0.332	0.654

Six columns are left out for better readability.

Overall, MetFusion was able to rank the correct candidate in the top position for the three challenges 1, 13 and 15. Challenges 2 and 14 had the correct compound ranked at position 4 and 3, respectively.

For Challenge 6, using MetFrag alone have a very poor result because 3812 candidates had an identical score of 0.0. MassBank does not contain spectra for the correct compound rheadine, and the most similar spectrum returned is palmatine (KOX00837), with a low 0.42 TS to the correct structure (as shown in Figure 1), while the structurally most similar entry (bicuculline, TS = 0.79) in MassBank has a poor spectral score of only 0.3. The main contribution from the MassBank results are three spectra from other alkaloids (alocryptopine, noscapine, and hydrastine) with a similarity between 0.59 and 0.77.

For Challenge 14, shown in Figure 2, MassBank returned a spectrum of carbazole ranked first, an isomer of the correct 1H-Benz[g]indole, followed by three spectra of compounds with both a different molecular formula and lower TS than the MetFrag candidates. During the contest, spectra of the correct 1H-Benz[g]indole measured on the same instrument as the challenge data were submitted to MassBank by one of the MassBank consortium members. The UF011410 hit in MassBank was only ranked fifth, with an unexpectedly low MassBank score of only 0.70, most likely because we used a merged query spectrum and MassBank applies a 5% intensity cut-off. These two factors led to a greater difference between the merged spectrum and the deposited reference spectrum. The available Orbitrap spectra

would benefit from a lower cut-off threshold of 2 rather than 5, but we relied on the default cut-off. With this low spectral similarity, the MassBank contribution was unable to lift the correct compound to the first rank, but only to rank 25.

Figure 2. Excerpt of reranked similarity matrix from MetFusion for Challenge 14. The correct compound is ranked 3rd (CID 98617) and highlighted with a green border. The two better ranking candidates have slightly higher MetFrag scores that add to their corresponding MetFusion scores. Compound 6854 is carbazole, a structurally highly similar compound towards the correct 1H-Benz[g]indole. The presence of Tanimoto similarities with value of 1.0 indicate perfect structural matches according to corresponding reference spectra available in MassBank for both 1H-Benz[g]indole (UF011410) and carbazole (UF026313).

	UF026313	UF024612	UF015113	WA002682	UF011410	WA000556	UF026913	UF011312	WA001663
59832560	0.875	0.135	0.141	0.063	0.969	0.098	0.495	0.101	0.067
59832555	0.845	0.135	0.142	0.059	0.979	0.093	0.497	0.102	0.067
98617	0.863	0.137	0.144	0.059	1.000	0.094	0.495	0.098	0.068
11344211	0.854	0.136	0.143	0.059	0.968	0.089	0.484	0.102	0.062
12450009	0.844	0.137	0.144	0.059	0.958	0.090	0.478	0.102	0.062
6854	1.000	0.146	0.152	0.057	0.863	0.089	0.454	0.102	0.065
13908560	0.760	0.145	0.151	0.057	0.894	0.097	0.451	0.101	0.071
13287594	0.750	0.146	0.152	0.062	0.863	0.098	0.446	0.102	0.065
12867691	0.747	0.140	0.146	0.058	0.823	0.089	0.427	0.098	0.073
10877507	0.747	0.140	0.146	0.058	0.823	0.089	0.427	0.098	0.073
14399831	0.740	0.139	0.146	0.057	0.814	0.089	0.432	0.098	0.072
11171191	0.740	0.139	0.146	0.057	0.814	0.089	0.432	0.098	0.072
21163914	0.740	0.139	0.146	0.057	0.814	0.089	0.432	0.098	0.072
22349125	0.138	0.481	0.667	0.122	0.132	0.122	0.158	0.190	0.100
22641511	0.118	0.604	0.536	0.115	0.113	0.116	0.145	0.173	0.080
12667390	0.435	0.255	0.291	0.084	0.489	0.120	0.436	0.155	0.080
12667393	0.430	0.252	0.289	0.080	0.477	0.119	0.433	0.164	0.079

For challenges 1 to 6 MetFusion performed significantly better than MetFrag, and the median rank of the correct compound was 20, compared to 280 with MetFrag and 145 with MLS. This is even more remarkable because we used the online PubChem query, which returned 3145 candidates (median), whereas the PubChem snapshot only provided 1063 candidates (median) over all challenges.

MetFusion results for challenges 10 to 12 were significantly worse when compared to MetFrag alone. This can be attributed to the low Tanimoto similarity of the correct candidate to any of the spectral hits. For each of these challenges, the MassBank scores are between 0.31 and 0.68 for the top hit, indicating a lack of reference spectra for these compound classes. The missing spectral coverage is expressed in both mediocre spectral scores and almost no Tanimoto similarity, visualised by the red-orange coloured matrix cells with maximum Tanimoto similarity of 0.4. This indicates the case where the spectral library cannot confirm any of the *in silico* candidates, thus leaving the user with no additional information.

4. Conclusions

The IPB entered the CASMI contest unofficially, because as part of the organising team and challenge data providers we could not be considered independent. However, we entered CASMI as internal participants with MetFrag and MetFusion and did not tune the parameters to obtain optimal results for the initial submission.

The use of small, domain-specific compound databases like KEGG, focussing on natural compounds increases the risk that the correct compound is missed. While such a compound may be more likely to be found in PubChem or ChemSpider, the number of false positives will increase due to the large number of synthetic compounds. We used the metabolite-likeness score [11] as an additional term in the scoring function of MetFrag. The metabolite-likeness score penalizes synthetic compounds and improved the rankings for the natural product challenges 1–6 in all but one case. Moreover, we see potential for further improvement of these preliminary results by optimisation of the weight factor ω and the evaluation on a larger dataset than available in the CASMI contest.

MetFusion was used without additional scoring terms, such as the metabolite-likeness score. The similarity matrices provide a deeper insight into the integrated MetFusion score to (manually) assess the reliability of the MassBank spectral summary.

Both approaches were applied fully automatically to the challenge data, but the selection of the neutral mass for the candidate failed in two cases, and the scoring did not always rank the correct solution in the top positions. Although expert knowledge is still required for a reliable interpretation, our approaches can reduce the manual effort for small compound identification.

We are looking forward to participating in the next CASMI contest as external participants.

Acknowledgements

We thank Julio Peironcely for releasing the program to calculate the metabolite-likeness score as Open Source, and similarly Sebastian Wolf for his previous work on MetFrag. Christoph Ruttkies acknowledges funding from Deutsche Forschungsgesellschaft (DFG) grant NE/1396/5-1.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinforma.* **2010**, *11*, 148, doi:10.1186/1471-2105-11-148.
2. Schymanski, E.L.; Gallampois, C.M.J.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal. Chem.* **2012**, *84*, 3287–3295.
3. Bolton, Evan E.; Wang, Y.; Thiessen, Paul A.; Bryant, Stephen H.; Wheeler, Ralph A.; Spellmeyer, David C. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities *Elsevier* **2008**, *4*, 217–241.
4. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; *et al.* MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass. Spectrom.* **2010**, *45*, 703–714.
5. Gerlich, M.; Neumann, S. MetFusion: Integration of compound identification strategies. *J. Mass Spectrom.* **2013**, *48*, 291–298.

6. Smith, C.; Want, E.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal. Chem.* **2006**, *78*, 779–787.
7. Kazmi, S.; Ghosh, S.; Shin, D.; Hill, D.; Grant, D. Alignment of high resolution mass spectra: Development of a heuristic approach for metabolomics. *Metabolomics* **2006**, *2*, 75–83.
8. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI-the worldwide chemical structure identifier standard. *J. Cheminf.* **2013**, *5*, 7, doi:10.1186/1758-2946-5-7.
9. Bolton, E.E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H.; Wheeler, R.A.; Spellmeyer, D.C. Chapter 12 PubChem: Integrated platform of small molecules and biological activities. *Ann. Rep. Comput. Chem.* **2008**, *4*, 217–241.
10. Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinforma.* **2007**, *8*, 105, doi:10.1186/1471-2105-8-105.
11. Peironcelly, J.E.; Reijmers, T.; Coulier, L.; Bender, A.; Hankemeier, T. Understanding and classifying metabolite space and metabolite-likeness. *PLoS One* **2011**, *6*, e28966.
12. Liaw, A.; Wiener, M. Classification and regression by randomforest. *R News* **2002**, *2*, 18–22.
13. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; *et al.* HMDB: The human metabolome database. *Nucleic Acids Res.* **2007**, *35*, D521–D526.
14. Irwin, J.J.; Sterling, T.; Mysinger, M.M.; Bolstad, E.S.; Coleman, R.G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model* **2012**, *52*, 1757–1768.
15. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.
16. Schymanski, E.L.; Neumann, S. CASMI: And the winner is *Metabolites* **2013**, *3*, 412–439.
17. Allwood, J.W.; Weber, R.J.; Zhou, J.; He, S.; Viant, M.R.; Dunn, W.B. CASMI—The small molecule identification process from a Birmingham perspective. *Metabolites* **2013**, *3*, 397–411.
18. Gerlich, M.; Neumann, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **1999**, *1*, 29–34.
19. Meringer, M.; Schymanski, E.L. Small molecule identification with MOLGEN and mass spectrometry. *Metabolites* **2013**, *3*, 440–462.

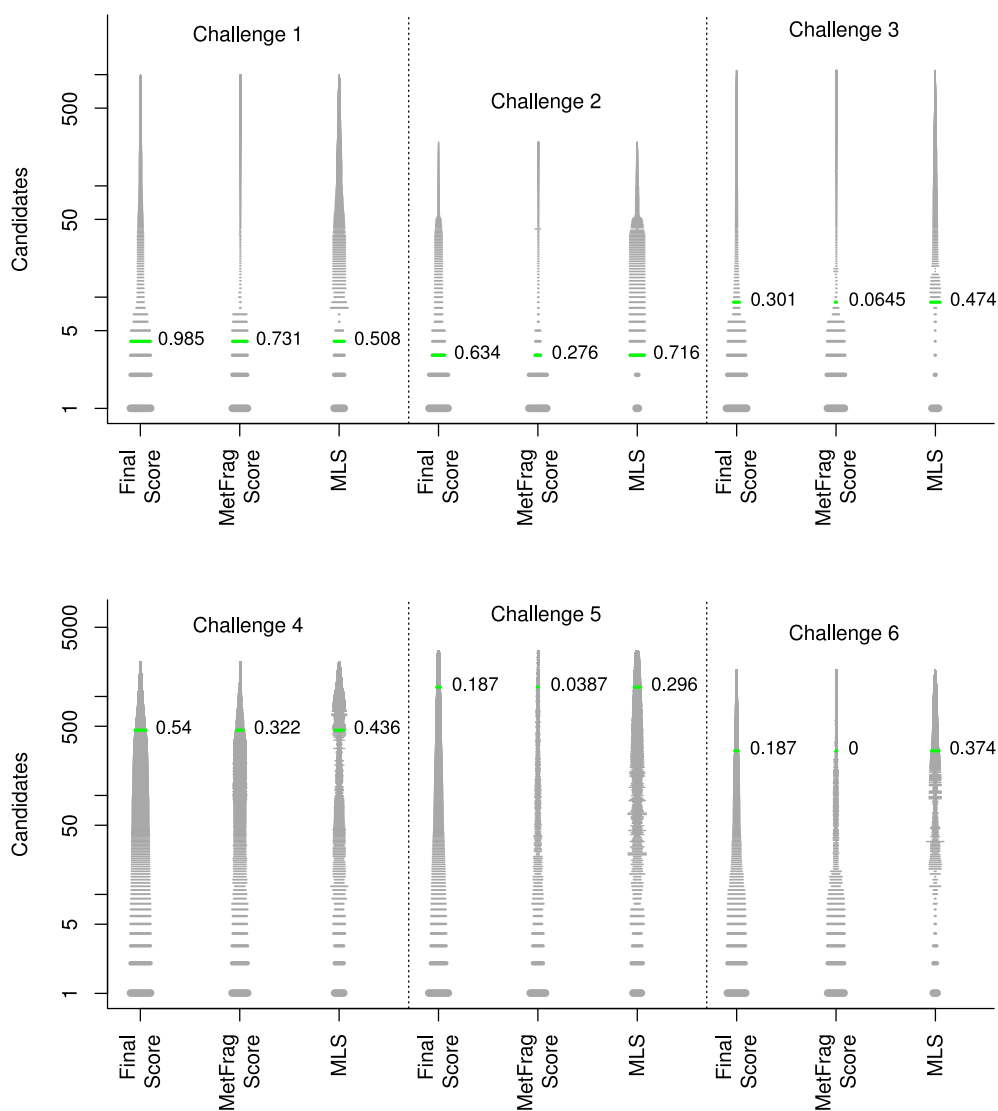
Appendix

A. Additional MetFrag Results

Table A1. MetFrag results with molecular formula filter after resubmission. Shown are the number of candidates per challenge, the InChIKey filtered MetFrag rank and the relative ranking position (RRP). Additionally, for challenges 1-6 the InChIKey filtered MetFrag rank with the metabolite-likeness score (MLS) included is shown.

Natural Product Challenges						Environmental Challenges			
Chall.	#Cand.	Rank	RRP	MLS	RRP	Chall.	#Cand.	Rank	RRP
						10	257	170	0.377
1	9	5	0.500	4	0.625	11	104	9	0.961
2	43	1	1.000	1	1.000	12	950	26	0.975
3	2	2	0.500	1	1.000	13	22	4	0.929
4	2005	534	0.735	444	0.779	14	111	19	0.859
5	2429	754	0.714	920	0.623	15	1789	172	0.905
6	1250	1250	0.416	234	0.814	16	1397	1397	0.438
						17	415	15	0.966
Median	646	270	0.607	119	0.797		336	22.5	0.917

Figure A1. Scores plot of challenges 1–6. The MetFrag and metabolite-likeness score (MLS) as well as the final scores of the candidates are shown for the challenges, respectively. The green line marks the position of the correct candidate and the given score. The width of each line correlates with the represented value of the score, respectively.



B. Spectral Merging

```
library(xcms)

## Read spectra into a list
tandemms <- lapply(c("MSMSneg10_Challenge3-A,1_01_2186-243.txt",
                    "MSMSneg20_Challenge3-A,1_01_2184-244.txt",
                    "MSMSneg30_Challenge3-A,1_01_2185-244.txt",
                    "MSMSneg40_Challenge3-A,1_01_2187-243.txt"),
                  function(x) {read.table(x,
                                          as.is=TRUE,
                                          sep="\t",
                                          header=FALSE,
                                          col.names=c("mz", "intensity"))})

## join into (redundant) peaklist
peaks <- do.call(rbind, tandemms)

## perform grouping of peaks based on m/z
g <- xcms::mzClust_hclust(peaks[, "mz"],
                        eppm=5*10e-6, eabs=0.001)

## create composite spectrum
mz <- tapply(peaks[, "mz"], as.factor(g), mean)
intensity <- tapply(peaks[, "intensity"], as.factor(g), max)
compositeSpectrum <- cbind(mz, intensity)
```

© 2013 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).

5.7 Solving CASMI 2013 with MetFrag, MetFusion and MOLGEN-MS/MS

Emma L. Schymanski, Michael Gerlich, **Christoph Ruttkies**, Steffen Neumann. Solving CASMI 2013 with MetFrag, MetFusion and MOLGEN-MS/MS. *Mass Spectrometry (Tokyo)*. 3(Spec Iss 2):S0036, 2014. 16 Citations⁷

https://www.jstage.jst.go.jp/article/massspectrometry/3/Special_Issue_2/3_S0036/_article

Contributions

The study was designed by Emma L. Schymanski who prepared the submission results for Category I using MOLGEN-MS/MS. I prepared the submission results for Category II using MetFrag. Michael Gerlich prepared submission results for Category II using MetFusion. The manuscript was prepared by Emma L. Schymanski and Steffen Neumann.

Copyright

I hereby declare that the copyright of this publication is by the Mass Spectrometry Society of Japan. The full-text article can be found under the above mentioned URL.

⁷<https://scholar.google.com> (accessed on 01/2021)

5.8 Critical Assessment of Small Molecule Identification 2016: automated methods

Emma L. Schymanski, **Christoph Ruttkies**, Martin Krauss, Céline Brouard, Tobias Kind, Kai Dührkop, Felixity Allen, Arpana Vaniya, Dries Verdegem, Sebastian Böcker, Johu Rousu, Huibin Shen, Hiroshi Tsugawa, Tanvir Sajed, Oliver Fiehn, Bart Ghesquière, Steffen Neumann. Critical Assessment of Small Molecule Identification 2016: automated methods. *Journal of Cheminformatics*, 9: 22, 2017. 128 Citations⁸
<https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0207-1>

Contributions

The paper is a result of the submissions made for the Critical Assessment of Small Molecule Identification Contest 2016 and contains contributions by over 10 participants. I produced the candidate lists together with Steffen Neumann and prepared the submission results for Category II and III using MetFrag.

⁸<https://scholar.google.com> (accessed on 01/2021)

RESEARCH ARTICLE

Open Access



Critical Assessment of Small Molecule Identification 2016: automated methods

Emma L. Schymanski^{1*}, Christoph Ruttkies², Martin Krauss³, Céline Brouard^{4,5}, Tobias Kind⁶, Kai Dührkop⁷, Felicity Allen⁸, Arpana Vaniya^{6,9}, Dries Verdegem¹⁰, Sebastian Böcker⁷, Juho Rousu^{4,5}, Huibin Shen^{4,5}, Hiroshi Tsugawa¹¹, Tanvir Sajed⁸, Oliver Fiehn^{6,12}, Bart Ghesquière¹⁰ and Steffen Neumann²

Abstract

Background: The fourth round of the Critical Assessment of Small Molecule Identification (CASMI) Contest (www.casmi-contest.org) was held in 2016, with two new categories for automated methods. This article covers the 208 challenges in Categories 2 and 3, without and with metadata, from organization, participation, results and post-contest evaluation of CASMI 2016 through to perspectives for future contests and small molecule annotation/identification.

Results: The Input Output Kernel Regression (CSI : IOKR) machine learning approach performed best in "Category 2: Best Automatic Structural Identification—*In Silico* Fragmentation Only", won by Team Brouard with 41% challenge wins. The winner of "Category 3: Best Automatic Structural Identification—Full Information" was Team Kind (MS-FINDER), with 76% challenge wins. The best methods were able to achieve over 30% Top 1 ranks in Category 2, with all methods ranking the correct candidate in the Top 10 in around 50% of challenges. This success rate rose to 70% Top 1 ranks in Category 3, with candidates in the Top 10 in over 80% of the challenges. The machine learning and chemistry-based approaches are shown to perform in complementary ways.

Conclusions: The improvement in (semi-)automated fragmentation methods for small molecule identification has been substantial. The achieved high rates of correct candidates in the Top 1 and Top 10, despite large candidate numbers, open up great possibilities for high-throughput annotation of untargeted analysis for "known unknowns". As more high quality training data becomes available, the improvements in machine learning methods will likely continue, but the alternative approaches still provide valuable complementary information. Improved integration of experimental context will also improve identification success further for "real life" annotations. The true "unknown unknowns" remain to be evaluated in future CASMI contests.

Keywords: Compound identification, *In silico* fragmentation, High resolution mass spectrometry, Metabolomics, Structure elucidation

Background

The Critical Assessment of Small Molecule Identification (CASMI) Contest [1] was founded in 2012 as an open contest for the experimental and computational mass spectrometry communities [2, 3]. Since then, CASMI contests have been held in 2013 [4], 2014 [5] and now in 2016, which is summarized in this article. The focus of

CASMI has changed slightly with each contest, reflecting differences in focus of the organizers as well as the perceived interest and challenges in structure elucidation with mass spectrometry. CASMI is purely a research activity—there is no fee for participation but likewise also no prize money for the winners.

In 2016, Category 1 was "Best Structural Identification on Natural Products", with 18 challenges available, a number achievable for both manual and automatic methods. Any methods could be used to submit entries and seven groups participated in this category. The outcomes

*Correspondence: emma.schymanski@eawag.ch

¹ Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland

Full list of author information is available at the end of the article



© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

of this category are presented separately [6] and reported here briefly for comparison purposes.

In contrast, Categories 2 and 3 were defined with 208 challenges in total. Candidate lists containing the correct solution were provided, along with training data for parameter optimization. These categories were specifically designed for automated methods, as no participant with a manual approach could be expected to invest so much time in solving all challenges. Category 2 was defined as “Best Automatic Structural Identification—*In Silico* Fragmentation Only”. The aim was to compare the different fragmentation approaches, ranging from combinatorial, to rule-based, to simulations; the use of mass spectral library searching or additional information was not allowed. In contrast, Category 3 was “Best Automatic Structural Identification—Full Information”. The same data files and candidate lists were provided as for Category 2, but any form of additional information could be used (retention time information, mass spectral libraries, patents, reference count, etc.). This was to assess the influence of additional information (hereafter termed metadata) on the results of the contest. Participants were required to detail their submissions in an abstract submitted with the results. The rules and submission formats were communicated on the CASMI rules website [7] prior to the release of the challenge data; the evaluation was automated provided the submission format passes all checks. In contrast to previous years, participants were allowed to submit up to three entries each, to evaluate the performance of different approaches. More details are given below.

This article summarizes Categories 2 and 3 of CASMI 2016, including organization, participation and additional post-contest analysis. Six external groups participated in these categories (see Graphical Abstract); 10 in total combined with the Category 1 participants, which is more than ever before.

Methods

Contest data for CASMI 2016

Mass spectra

All MS/MS spectra were obtained on a Q Exactive Plus Orbitrap (Thermo Scientific), with <5 ppm mass accuracy and nominal MS/MS resolving power of 35,000 at $m/z = 200$ using electrospray ionization (ESI) and stepped 20/35/50 nominal higher-energy collisional dissociation (HCD) energies. The spectra were obtained by measuring 22 mixes of authentic standards with the same liquid chromatography–mass spectrometry (LC–MS) method, in data-dependent acquisition mode using inclusion lists containing the $[M + H]^+$ (positive) and $[M - H]^-$ ion masses. Positive and negative mode data were acquired separately. Each mix contained between 10 and 94 compounds. A reversed phase column was

used (Kinetex C₁₈ EVO, 2.6 μ m, 2.1 \times 50 mm with a 2.1 \times 5 mm precolumn from Phenomenex). The gradient was (A/B): 95/5 at 0 min, 95/5 at 1 min, 0/100 at 13 min, 0/100 at 24 min (A = water, B = methanol, both with 0.1% formic acid) at a flow rate of 300 μ L/min.

The MS/MS peak lists were extracted with RMassBank [8] using the ion mass and a retention time window of 0.4 min around the expected retention time and reported as absolute ion intensities. To obtain high-quality spectra, the data was cleaned and recalibrated to within 5 ppm using known subformula annotation [8], all other peaks without a valid subformula within 5 ppm of the recalibrated data were removed. All substances with double chromatographic peaks, different substances with identical spectra (detected via the SPECTRAL HASH (SPLASH) [9, 10]), MS/MS containing only one peak or with a maximum intensity below 1×10^5 were excluded from the datasets. Substances that were measured multiple times (because they were present in more than one mix) in the same ionization mode were only included once, selected by higher intensity. MS/MS from positive and negative mode were included if the substance ionized in both modes. The final peak lists were saved in plain text format and Mascot Generic Format (MGF). All MS/MS spectra are now available on MassBank [11].

Candidates

The candidates were retrieved from ChemSpider via MetFrag2.3 [12] using the monoisotopic exact mass ± 5 ppm of the correct candidate on February 14th, 2016. The SMILES from the MetFrag output were converted to standard InChIs and InChIKeys with OpenBabel (version 2.3.2) [13]. Candidates were removed if the SMILES to InChI conversion failed, all other candidates were retained without any additional filtering. The presence of the correct solution in the candidate list was verified and the lists were saved as CSV files.

Training and challenge datasets

The MS/MS spectra and corresponding candidates were split into training and challenge datasets, according to the spectral similarity to MassBank spectra (as many substances were already in MassBank). Challenge spectra were those where no MassBank spectrum was above 0.85 similarity (calculated with MetFusion [14]); all spectra where there was a match in MassBank above 0.85 were included in the CASMI training set. There were two exceptions: Alizarin, similarity 0.88 to laxapur (FIO00294), and anthrone, similarity 0.86 to phosphocreatine (KO003849), to ensure a sufficient number of natural products remained as challenges for Category 1 (see below). Many of the natural products in the mixes did not ionize well with the experimental setup used.

The challenge dataset consisted of 208 peak lists from 188 substances, 127 obtained in positive mode (all $[M+H]^+$) and 81 in negative mode (all $[M-H]^-$). The retention times for each substance was provided in a summary CSV file. The training dataset consisted of 312 MS/MS peak lists (from 285 substances), of which 254 were obtained in positive mode (all $[M+H]^+$) and 58 negative mode (all $[M-H]^-$). The identities and retention times of the substances in the training dataset were provided in a summary CSV file. All files were uploaded to the CASMI website [15]. Participants were asked to contact the organizers if they required additional formats.

To allow a comparison with manual approaches, Challenges 10–19 in Category 1 were a (re-named) subset of the dataset in Categories 2 and 3. The corresponding challenge numbers are given in Table 1.

Information about the full scan (MS1) data was not originally provided for CASMI 2016, but was provided retrospectively for Challenges 10–19 in Category 1 upon request and post-contest for Categories 2 and 3 for another publication [16]. All data is now available on the CASMI website [15].

Rules and evaluation

The goal of the CASMI contest was for participants to determine the correct molecular structure for each challenge spectrum amongst the corresponding candidate set, based on the data provided by the contest organizers. A set of rules were fixed in advance to clarify how the submissions were to be evaluated and ranked, to ensure that the evaluation criteria were transparent and objective. All participants were encouraged to follow the principles of reproducible research and accurately describe how their results were achieved in an abstract submitted with the results. Submission formats were defined in advance (described below) to satisfy the R scripts used to

perform the automatic evaluation, results and web page generation. Test submissions could be submitted pre-deadline to check for issues; any post-deadline problems were resolved prior to the release of the solutions.

Participants could enter a maximum of three submissions per approach and category, provided they used these submissions to assess the influence of different strategies on the outcomes. The rationale and differences had to be detailed in the abstract. The *best overall performing submission* per participant was considered in declaring the winner(s). The submission requirements were an abstract file (per submission, see website for details) plus results files for each challenge to be considered in the contest. There was no explicit requirement to submit entries for all challenges. Valid challenge submissions were plain text, tab separated files with two columns containing the representation of the structure as the standard InChI or the SMILES code (column 1) and the score (column 2). To be evaluated properly, the score was to be non-negative with a higher score representing a better candidate.

For each challenge, the absolute rank of the correct solution (ordered by score) was determined. The average rank over all equal candidates was taken where two or more candidates had the same score. Due to inconsistencies with how participants dealt with multiple stereoisomers (and since stereoisomers amongst the candidates could not be separated with the analytical methods used), submissions were filtered post-submission to remove duplicate stereoisomers using the first block of the InChIKey. The *highest scoring isomer* was retained. The ranks were then compared across all eligible entries to declare the gold (winner), silver and bronze positions for each challenge. *Gold was awarded to the contestant(s) with the lowest rank among all contestants for that challenge*. This way, a winner could be declared even if no method ranked the correct candidate in the Top 1. Joint positions were possible in case of ties. The overall winner was determined using an Olympic medal tally scheme, i.e. the participants with the most gold medals per category won. The winners were declared on the basis of this automatic evaluation.

Table 1 Overlapping challenges between Category 1 and Categories 2 and 3

Name	Category 1	Categories 2 and 3	Mode
Creatinine	Challenge-010	Challenge-084	Positive
Anthrone	Challenge-011	Challenge-162	Positive
Flavone	Challenge-012	Challenge-166	Positive
Medroxyprogesterone	Challenge-013	Challenge-184	Positive
Abietic acid	Challenge-014	Challenge-207	Positive
Estrone-3- β -D-glucuronide)	Challenge-015	Challenge-034	Negative
Alizarin	Challenge-016	Challenge-045	Negative
Thyroxine	Challenge-017	Challenge-048	Negative
Purpurin	Challenge-018	Challenge-054	Negative
Monensin	Challenge-019	Challenge-079	Negative

Additional scores

Further scores that were used to interpret the results included the mean and median ranks, Top X rank counts, relative ranking positions (RRPs, defined in [2]) and quantiles. The *Formula 1 Score*, based on the method used in Formula 1 racing [17] since 2010, is the sum of the Top 1 to 10 ranks of the correct candidates weighted by the scores 25, 18, 15, 12, 10, 8, 6, 4, 2 and 1. The *Medal Score* (as opposed to the per-challenge Gold Medal count used in CASMI to declare the winner) is the sum of

weighted Top 1 ranks with 5 points (gold medal), Top 2 ranks with 3 points (silver) and Top 3 ranks (bronze) with 1. Non-integer ranks (due to equally-scoring candidates) were rounded up to the higher rank for calculating Top X, Formula 1 and medal scores (e.g. rank 1.5 was counted as 2).

Participant methods

Team Allen (Felicity Allen, Tanvir Sajed, Russ Greiner and David Wishart) processed the provided candidates for Category 2 using CFM-ID [18]. CFM-ID uses a probabilistic generative model to produce an *in silico* predicted spectrum for each candidate compound. It then uses standard spectral similarity measures to rank those candidates according to how well their predicted spectrum matches the challenge spectrum. The original Competitive Fragmentation Model (CFM) positive and negative models were used, which were trained on data from the METLIN database [19]. Mass tolerances of 10 ppm were used, the Jaccard score was applied for spectral comparisons and the input spectrum was repeated for low, medium and high energies to form the *CFM_orig* entry. The *CFM_retrain* entry consisted of a CFM model trained on data from METLIN and the NIST MS/MS library [20] for the positive mode spectra. This new model also incorporated altered chemical features and a neural network within the transition function. Mass tolerances of 10 ppm were used, and the DotProduct score was applied for spectral comparisons. This model combined the spectra across energies before training, so only one energy exists in the output. The negative mode entries were the same as for *CFM_orig*.

CFM-ID was also used to submit entries for Category 3, by combining the above CFM-based score with a database score (DB_SCORE). For each hit in the databases HMDB [21], ChEBI [22], FooDB [23], DrugBank [24] and a local database of plant-derived compounds, 10 was added to DB_SCORE. The *CFM_retrain+DB* and *CFM_orig+DB* submissions were formed by adding the DB_SCORE for each candidate to the *CFM_retrain* and *CFM_orig* entries from Category 2, respectively.

Team Brouard (Céline Brouard, Huibin Shen, Kai Dührkop, Sebastian Böcker and Juho Rousu) participated in Category 2 using CSI:FingerID [25] with an Input Output Kernel Regression (IOKR) machine learning approach to predict the candidate scores [26]. Fragmentation trees were computed with SIRIUS version 3.1.4 [27] for all the molecular formulas present in the candidate set. Only the tree associated with the best score was considered. SIRIUS uses fragment intensities to distinguish noise and signal peaks, while the intensities were weighted lowly during learning (see [25, 26]). Different kernel functions were computed for measuring the similarities between

either MS/MS spectra or fragmentation trees. Multiple kernel learning (MKL, see [28]) was used to combine the kernels as input for IOKR. In the *CSI:IOKR_U* submission, the same weight was associated with each kernel (uniform multiple kernel learning or “Uni-MKL”). In the *CSI:IOKR_A* submission the kernel weights were learned with the Alignf algorithm [29] so that the combined input kernel was maximally aligned to an ideal target kernel between molecules. In both submissions, IOKR was then used for learning a kernel function measuring the similarity between pairs of molecules. The values of this kernel on the training set were defined based on molecular fingerprints, using approximately 6000 molecular fingerprints from CDK [30, 31]. Separate models were trained for the MS/MS spectra in positive and negative mode. The method was trained using the CASMI training spectra, along with additional merged spectra from GNPS [32] and MassBank [33]. For the negative ion mode spectra, 102 spectra from GNPS and 714 spectra from MassBank were used. For the positive ion mode spectra, 3868 training spectra from GNPS were used. These training sets were prepared following a procedure similar to that described in [25].

The additional post-competition submission *CSI:IOKR_AR* used the same approach as *CSI:IOKR_A*, but the positive model was learned using a larger training set containing 7352 positive mode spectra from GNPS and MassBank. This training set was effectively the same as that used by Team Dührkop, with minor differences due to the pre-selection criteria of the spectra. The negative mode training set was not modified.

Team Dührkop (Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu and Sebastian Böcker) entered Category 2 with a command line version of CSI:FingerID version 1.0.1 [25], based on the original support vector machine (SVM) machine learning method. The peaklists were processed in MGF format and fragmentation trees were computed with SIRIUS version 3.1.4 [27] using the Q-TOF instrument settings. Trees were computed for all candidate formulas in the given structure candidate list; trees with a score <80% of the optimal tree score were discarded. The remaining trees were processed with CSI:FingerID. SIRIUS uses fragment intensities to distinguish noise and signal peaks, while the intensities are weighted lowly in CSI:FingerID (see [25]). Molecular fingerprints were predicted for each tree (with Platt probability estimates [34]) and compared against the fingerprints of all structure candidates (computed with CDK [30, 31]) with the same molecular formula. The resulting hits were merged together in one list and were sorted by score. A constant value of 10,000 was added to all scores to make them positive (as required in the CASMI rules). Ties of compounds with same score (and

sometimes also with same 2D structure) were ordered randomly. The machine learning method was trained on 7352 spectra (4564 compounds) downloaded from GNPS [32] and MassBank [33]. All negative ion mode challenges were omitted due to a lack of training data; i.e. entries were only submitted for positive challenges. This formed the CSI:FID entry.

Team Dührkop submitted a second “leave out” entry, CSI:FID_leaveout, during the contest. Before the correct answer was known, the team observed that the top-scoring candidate matched a compound from the CSI:FID training set in 67 challenges, which could indicate that the method had memorized the training spectra. To assess the generalization of their method, the classifiers were retrained on the same training set, plus CASMI training spectra, but with these top scoring candidates removed. As this entry was “guesswork” and did not affect the contest outcomes, upon request Team Dührkop resubmitted a true “leave out” entry post-contest where all CASMI challenge compounds were removed from their training set (not just their “guess” based on top scoring candidates) prior to retraining and calculating the CSI:FID_leaveout results. For the sake of interpretation, only these updated “leave out” results are presented in this manuscript.

Team Kind (Tobias Kind, Hiroshi Tsugawa, Masanori Arita and Oliver Fiehn) submitted entries to Category 3 using a developer version (1.60) of the freely available MS-FINDER software [35, 36] combined with MS/MS searching and structure database lookup for confirmation (entry MS-FINDER+MD). MS-FINDER was originally developed to theoretically assign fragment substructures to MS/MS spectra using hydrogen rearrangement (HR) rules, and was subsequently developed into a structure elucidation program consisting of formula prediction, structure searching and structure ranking methods. For CASMI, an internal database was used to prioritize existing formulas from large chemical databases over less common formulas and the top 5 molecular formulas were regarded for structure queries. Each formula was then queried in the CASMI candidate lists as well as an internal MS-FINDER structure database. A tree-depth of 2 and relative abundance cutoff of 1% as well as up to 100 possible structures were reported with MS-FINDER. The final score was calculated by the integration of mass accuracy, isotopic ratio, product ion assignment, neutral loss assignment, bond dissociation energy, penalty of fragment linkage, penalty of hydrogen rearrangement rules, and existence of the compound in the internal MS-FINDER structure databases (see Additional file 1 for full details). MS-FINDER uses ion intensities in the relative abundance cutoff and isotopic ratio calculations, but not in candidate scoring.

Secondly, MS/MS search was used for further confirmation via the NIST MS Search GUI [37] together with major MS/MS databases such as NIST [20], MassBank of North America (MoNA) [38], ReSpec [39] and MassBank [33]. The precursor was set to 5 ppm and product ion search tolerance to 200 ppm. Around 100 out of the 208 candidates had no MS/MS information. For these searches, a simple similarity search without precursor information was also used, or the precursor window was extended to 100 ppm. Finally, those results that gave overall low hit scores were also cross-referenced with the STOFF-IDENT database of environmentally-relevant substances [40, 41] to obtain information on potential hit candidates. This step was taken because the training set consisted of mostly environmentally relevant compounds.

Team Vaniya (Arpana Vaniya, Stephanie N. Samra, Sajjan S. Mehta, Diego Pedrosa, Hiroshi Tsugawa and Oliver Fiehn) participated in Category 2 using MS-FINDER [35, 36] version 1.62 (entry MS-FINDER). MS-FINDER uses hydrogen rearrangement rules for structure elucidation using MS and MS/MS spectra of unknown compounds. The default settings were used; precursor m/z , ion mode, mass accuracy of instrument, and precursor type (given in CASMI) were used to populate the respective fields in MS-FINDER. Further parameter settings were: tree depth of 2, relative abundance cutoff of 1, and maximum report number of 100. Although relative abundance cutoffs were used to filter out noisy data, ion abundances were not used by MS-FINDER for calculation of either the score or rank of candidate structures. The default formula finder settings were used, except the mass tolerance, which was set to ± 5 ppm mass accuracy as given by the CASMI organizers.

MS-FINDER typically retrieves candidates from an Existing Structure Database (ESD) file compiled from 13 databases, but this was disabled as candidates were provided. Instead, one ESD was created for each of the 208 challenges, containing the information from the candidate lists provided by the CASMI organizers. A batch search of the challenge MS/MS against the challenge candidate list (in the ESD) was performed on the top 500 candidates, to avoid long computational run times. Up to 500 top candidates structures were exported as a text file from MS-FINDER. Scores for automatically matching experimental to virtual spectra were ranked based on mass error, bond dissociation energy, penalties for linkage discrepancies, or violating hydrogen rearrangement rules. Final scores and multiple candidate SMILES were reported for 199 challenges for submission to CASMI 2016. Nine challenges could not be processed due to time constraints (Challenges 13, 61, 72, 78, 80, 106, 120, 133, 203). Full details on this entry, MS-FINDER and file

modifications required are given in Additional files 1 and 2.

Team Verdegem (Dries Verdegem and Bart Ghesquière) participated in Category 2 with MAGMa+ [42], which is a wrapper script for the identification engine MAGMa [43]. For any given challenge, MAGMa+ runs MAGMa twice with two different parameter sets. A total of four optimized parameter sets exist (two for positive and two for negative ionization mode), which all differ from the original MAGMa parameters. Within one ionization mode, both corresponding parameter sets were each optimized for a unique latent molecular class. Following the outcome of both MAGMa runs, MAGMa+ determines the molecular class of the top ranked candidates returned by each run using a trained two-class random forest classifier. Depending on the most prevalent molecular class, one outcome (the one from the run with the parameters corresponding to the most prevalent class) is returned to the user. The candidate lists provided were used as a structure database without any prefiltering. MAGMa determines the score by adding an intensity-weighted term for each experimental peak. If a peak is explained by the *in silico* fragmentation process, the added term reflects the difficulty with which the corresponding fragment was generated. Otherwise, an “unexplained peak penalty” is added. Consequently, MAGMa returns smaller scores for better matches, and therefore the reciprocal of the scoring values was submitted to the contest. MAGMa was run with a relative m/z precision of 10 ppm and an absolute m/z precision of 0.002 Da. Default values were taken for all other options. MAGMa+ is available from [44].

To enable a comparison between MAGMa+ (entry MAGMa+) and MAGMa, entries based on MAGMa were submitted post-contest (entry MAGMa). MAGMa was run as is, without customization of its working parameters (bond break or missing substructure penalties). Identical mass window values as for MAGMa+ were applied (see above). Default values were used for all other settings. Again, the reciprocal of the scoring values was submitted to obtain higher scores for better matches.

Additional results

Additional results were calculated using MetFrag2.3 [12] to compare these results with the other methods outside the actual contest and to investigate the influence of metadata on the competition results. MetFrag command line version 2.3 (available from [45]) was used to process the challenges, using the MS/MS peak lists and the ChemSpider IDs (CSIDs) of the candidates provided. MetFrag assigns fragment structures generated *in silico* to experimental MS/MS spectra using a defined mass difference. The candidate score considers the mass and

intensity of the explained peaks, as well as the energy required to break the bond(s) to generate the fragment. Higher masses and intensities will increase the score, while higher bond energies will decrease the score. The MetFrag submission consisted of the MetFrag fragmentation approach only. In the MetFrag+CFM entry the MetFrag and CFM-ID (version 2) [18] scores were combined. The CFM scores were calculated independently from Team Allen. Additionally, a Combined_MS/MS entry was prepared, combining six different fragmenters with equal weighting: CFM_orig, CSI:FID, CSI:IOKR_A, MAGMa+, MetFrag and MS-FINDER.

Several individual metadata scores were also prepared. A retention time prediction score was based on a correlation formed from the CASMI training set (submission Retention_time; +RT, see Additional file 1: Figure S1. The reference score (submission Refs) was the ChemSpiderReferenceCount, retrieved from ChemSpider [46] using the CSIDs given in the CASMI data. The MoNA submission ranked the candidates with the MetFusion-like [14] score built into MetFrag2.3, using the MoNA LC-MS/MS spectral library downloaded January 2016 [38]. The Lowest_CSID entry had candidates scored according to their identifier, where the lowest ChemSpider ID was considered the best entry.

The combined submissions to test the influence of different metadata on the results were as follows: MetFrag+RT+Refs, MetFrag+CFM+RT+Refs, MetFrag+CFM+RT+Refs+MoNA, Combined_MS/MS+RT+Refs and finally Combined_MS/MS+RT+Refs+MoNA. Full details of how all these submissions were prepared are given in Additional file 1.

Results

CASMI 2016 overall results

The sections below are broken up into the official results of the two categories during the contest, shown in Table 2, followed by the post-contest evaluation and a comparison with all approaches from Category 1.

Category 2: *In silico* fragmentation only

The results from Category 2 are summarized in Table 2. The participant with the highest number of wins over all challenges (i.e. gold medals) was **Team Brouard** with 86 wins over 208 challenges (41%) for CSI:IOKR_A. **Team Dührkop** with CSI:FID (82 gold, 39%) and **Team Vaniya** with MS-FINDER (70 gold, 34%) were in second and third place, respectively. This clearly shows that the recent machine-learning developments have greatly improved the performance relative to the bond-breaking approaches and even CFM. The third place for MS-FINDER shows that it performs in quite a complementary way to the CSI methods. The performance of

Table 2 Results summary for Categories 2 and 3: medal tally and other statistics

	Category 2					Category 3	
	Allen CFM orig	Brouard CSI: IOKR.A	Dührkop CSI:FID	Vaniya MS- FINDER	Verdegem MAGMa+	Allen CFM retrain +DB	Kind MS- FINDER +MD
Gold	63	86	82	70	44	156	159
Silver	71	50	21	26	53	52	38
Bronze	40	31	11	35	65	0	0
Gold (neg)	26	20	0	33	24	61	64
Gold (pos)	37	66	82	37	20	95	95
Top 1 (neg)	12	9	0	14	8	47	59
Top 1 (pos)	27	53	70	32	16	73	47
Top 1	39	62	70	46	24	120	146
Top 3	77	93	90	79	59	160	162
Top 10	123	118	100	101	105	182	174
Mean rank	47.98	127.34	25.17	19.75	70.79	13.72	6.4
Median rank	6	5.2	1	3	9.8	1	1
Mean RRP	0.906	0.874	0.945	0.804	0.88	0.971	0.904
Median RRP	0.987	0.988	1	0.922	0.972	1	1
Formula 1	1957	2276	2156	1867	1524	3861	4011
Medal Score	275	375	396	305	195	700	766

The first, second and third place by "Gold medals" (used to declare CASMI winners) are highlighted in red, orange and yellow, respectively. The best value per statistic is marked in bold

Team Dührkop is especially surprising considering that they did not submit any challenges in negative mode (due to a lack of training data).

Table 2 also includes the Top 1 (correct candidate ranked in first place), Top 3 (correct candidate amongst the top 3 scoring entries) and Top 10 entries per participant as well as the Formula 1 and Medal scores. The CSI:FID entry from Team Dührkop had the best Top 1 result (70, or 34%), followed by Team Brouard and Team Vaniya with 62 and 46 Top 1 candidates. This is an amazing improvement on previous contests and consistent with recent results [25], despite their use of larger candidate sets (PubChem instead of ChemSpider) and a slightly different ranking system. Very interesting to note is that all methods have the correct candidate in the Top 10 in $\geq 49\%$ of cases, which is likewise a dramatic improvement for automatic annotation. CFM_orig had the most the correct candidates in the Top 10 (123 or 59%) and this is reflected in the Formula 1 Score, which weighted the CFM_orig performance ahead of MS-FINDER, despite their lower Top 1 ranks.

Separating the challenges into positive and negative modes revealed that Team Dührkop clearly led the positive mode predictions (82 wins/gold medals and 70 Top 1 candidates, versus 66 wins and 53 Top 1 candidates for Team Brouard). Both MS-FINDER (14 Top 1) and CFM_orig (12 Top 1) outperformed Team Brouard for negative mode (9 Top 1), showing that a greater amount of training data for negative spectra would likely improve the CSI methods in the future. The training set used by

Team Brouard contained 7300 spectra for positive mode and only 816 negative mode spectra. The difference between positive and negative mode was less dramatic for the other approaches.

The results of Category 2 were dominated by the methods that use machine learning on large spectral databases (GNPS [32], MassBank [33], METLIN [19] and NIST [20]), namely Teams Brouard and Dührkop (CSI) and Allen (CFM). The great increase in data available for training these methods has led to the dramatic improvements in *in silico* methods seen in this contest—increasing the availability of open data will only improve this situation further! The performance of MS-FINDER, which does not use machine learning but instead chemical interpretation, is also particularly encouraging and below is shown to perform quite complementary to the machine learning methods. The influence of the training data was investigated during the contest by Teams Dührkop (CSI:FID_leaveout) and Allen (CFM_retrain); see Table 3. This was investigated for all approaches post-contest, discussed in "Machine learning approaches and training data" section.

Category 3: Full information

The results of Category 3, also summarized in Table 2, were extremely close considering the freedom given to the use of metadata in this Category. Team Kind was the winner with 159 gold (64 positive, 95 negative), closely followed by Team Allen on 156 gold (61 positive, 95 negative). Interestingly, the number of Top 1 ranks were

Table 3 Results summary for additional Category 2 entries

	Allen		Brouard		Dührkop		Ruttkies		Vaniya		Verdegem	
	CFM_orig	CFM_retrain	CSI:OKR_A	CSI:OKR_AR*	CSI:OKR_U	CSI:FID	CSI:FID_leaveout*	MetFrag*	MetFrag+CFM*	MS-FINDER	MAGMa+	MAGMa*
Top 1 Neg.	12	12	9	9	8	0	0	9	20	14	8	7
Top 1 Pos.	27	28	53	69	50	70	36	15	21	32	16	14
Top 1	39	40	62	78	58	70	36	24	41	46	24	21
Top 3	77	73	93	102	95	90	70	60	84	79	59	51
Top 10	123	116	118	131	118	100	88	108	127	101	105	106
Mean rank	47.98	44.53	127.3	95.09	123.3	25.17	52.02	51.92	33.97	19.75	70.79	70.24
Med. rank	6	7	5.25	4	5	1	3	8.75	6	3	9.8	9.8
Mean RRP	0.906	0.917	0.874	0.887	0.857	0.945	0.931	0.905	0.915	0.804	0.88	0.88
Med. RRP	0.987	0.985	0.988	0.993	0.98	1	0.995	0.98	0.991	0.922	0.972	0.969
Gold	53	52	73	91	70	74	41	32	51	61	35	31
Formula 1	1957	1900	2276	2500	2237	2156	1596	1593	2058	1867	1524	1463
Medal Sc.	275	269	375	442	371	396	252	198	292	305	195	175
Q ₁₀	1	1	1	1	1	1	1	1	1	1	1	1.4
Q ₂₅	2	2	1	1	1	1	1	3	2	1	3	3.5
Q ₅₀	6	7	5.25	4	5	1	3	8.75	6	3	9.8	9.8
Q ₇₅	36.25	27.63	55.5	36	78.75	6	17	37.88	25	17	66.1	64.5
Q ₉₀	121.8	104.6	192.9	134.9	288.9	37.5	72.4	120.9	87.65	68.75	187.1	148.5

The column header of entries used in Table 2 are given in italics. The best value per statistic is marked in bold. * indicates internal and post-competition submissions. Med. = median. Q_X indicates Xth quantile

very different, 146 (Team Kind) versus 120 (Team Allen); consistent with Category 2 *CFM_orig* had more Top 10 entries but fewer Top 1 and 3 entries than *MS-FINDER*. In this category the *CFM_retrained* model from Team Allen outperformed *CFM_orig*, which performed better in Category 2.

While very different approaches were used to obtain the “metadata”, the results of Category 3 clearly demonstrate the value of using metadata when identifying “known unknowns” as was the case in this contest where candidates were provided. This decision to provide candidates was taken deliberately to remove the influence of the candidate source on the CASMI results. The role of this “metadata” is discussed further below (Category 3: Additional Results). For true unknown identification the benefit of this style of metadata could be considerably reduced depending on the context, however this would have to be the subject of an alternative category in a future contest.

Post-contest evaluation

While the best overall results per participant were used to declare the winners, each participant was able to submit up to three entries to the contest if they chose to assess the influence of different strategies on their outcome. This has revealed many interesting aspects that would otherwise have gone undetected with only one entry per participant, as in previous contests. To explore these further and take advantage of the automatic evaluation procedure offered in CASMI, several internal and

post-contest entries were also evaluated, as described in the Methods section. The results of all these entries, including those run in the contest, are given in Table 3 for Category 2 and in Table 4 for Category 3.

Category 2: Additional results

The additional results for Category 2 (see Table 3) show that the retrained *CSI:IOKR_AR* entry from Team Brouard (using the more extensive *CSI:FID* training data plus negative mode results) would have outperformed their winning *CSI:IOKR_A* entry as well as the *CSI:FID* entry from Team Dührkop. The improvement with additional training data was dramatic for some challenges, e.g. Challenge 178 went from Rank 3101 with *CSI:IOKR_A* to rank 1 with *CSI:IOKR_AR*. Separating the Top 1 ranks into positive and negative mode (see Table 3) shows indeed that the performance for *CSI:IOKR_AR* and *CSI:FID* in positive mode was quite similar (69 vs. 70 wins, respectively), whereas all *CSI* methods are outperformed by *MS-FINDER* and *CFM_orig* in negative mode.

The *MetFrag* entry performed quite similarly to Team Verdegem (*MAGMa+*); as both are combinatorial fragmentation approaches this is not surprising. While the *MetFrag+CFM* entry improved these results dramatically, it was only slightly improved compared with the individual *CFM* entries of Team Allen. However, the improvement by combining the two fragmenters in negative mode was marked, increasing the Top 1 ranks from 9 (*MetFrag*) and 12 (*CFM*) to 20 (*MetFrag+CFM*).

Table 4 Results summary for additional Category 3 entries

	Allen		Kind	Ruttkies		
	<i>CFM orig +DB</i>	<i>CFMretrain+DB</i>	<i>MS-FINDER+MD</i>	<i>MetFrag+RT+Refs*</i>	<i>MetFrag+CFM+RT+Refs*</i>	<i>MetFrag+CFM+RT+Refs+MoNA*</i>
Top 1	117	120	146	162	163	155
Top 3	159	160	162	183	180	182
Top 10	182	182	174	191	199	194
Mean rank	14	13.62	6.4	7.04	5.39	4.25
Median rank	1	1	1	1	1	1
Mean RRP	0.969	0.971	0.904	0.987	0.989	0.990
Median RRP	1	1	1	1	1	1
Gold	124	128	148	168	174	167
Formula 1	3798	3861	4011	4469	4509	4437
Medal score	687	700	766	855	856	840
Q_10	1	1	1	1	1	1
Q_25	1	1	1	1	1	1
Q_50	1	1	1	1	1	1
Q_75	3	3	2	1	1	2
Q_90	13.7	14.0	15.0	5.0	5.0	4.3

The column header of entries used in Table 2 are given in italics. The best value per statistic is marked in bold. * Indicates internal and post-competition submissions. Q_X indicates Xth quantile

MS-FINDER still performed the best in negative mode of all the individual entries. MAGMa+ outperformed MAGMa in Top 1 and Top 3 entries.

Category 3: Additional results

The additional results for Category 3 (see Table 4) show that MetFrag+CFM+RT+Refs outperformed the other approaches both in terms of wins and the number of Top 1 ranks. Although adding MoNA to the mix resulted in a poorer performance, this was because spectral similarity was used to separate the training and challenge sets and the resulting MoNA weight was too optimistic for the challenges.

As these results are driven more by the metadata used than the fragmenter behind, a variety of entries were created to assess the contribution of the individual metadata aspects, as well as a “Combined Fragmenter” entry (Combined MS/MS) to remove the influence of the fragmentation method (see “Methods” for details). These results are given in Table 5. The Combined MS/MS entry outperformed all of the individual Category 2 entries, showing the complementarity of the different approaches. These also outperformed the MS library (MoNA) entry. The retention time prediction alone performed poorly, because this does not contain sufficient structural information to distinguish candidates, as demonstrated in Additional file 1: Figure S2. The lowest identifier strategy, which was used as a “gut feeling” decision criteria commonly in environmental studies before retrieval of reference information could be automated, takes advantage of the fact that well known substances were added to ChemSpider earlier and thus have lower identifiers. Surprisingly this still outperformed the combined fragmenters—but again this is highly dependent on the dataset. The references outperformed all individual metadata categories and even the combined fragmenters clearly. The influence of the metadata is discussed further in “Metadata and consensus identification” section.

Comparison with results from Category 1

Challenges 10–19 in Category 1 were also present among the Category 2 and 3 challenges, as given in Table 1. The results for these challenges, separated by category, are summarized in Table 6 and visualized in Figure S3 and S4 in Additional file 1. Interestingly, this shows that the results of Categories 1 and 3 were remarkably comparable, while the ranks of Category 2, using only MS/MS data, were generally worse. Again, this shows that the incorporation of metadata in automated methods is essential to guide users to the identification for known substances—but misleading when assessing the performance of computational methods. As metadata cannot assist in the identification of true unknowns for which no data exists, more work is still needed to bring the performance of the *in silico* MS/MS identification methods (Category 2) closer to that of Categories 1 and 3. However, it is clear from this 2016 contest that much progress has been made with the new machine learning methods and—as observed above—continuing to improve the availability of training data will improve these further.

Interestingly, Challenge 14 (Abietic acid) was challenging for all participants in all categories; this was the only challenge in Category 1 where no participant had the correct answer in first place despite the fact that the challenge spectrum was very informative and the candidate numbers were relatively low (see Additional file 1: Figure S7).

Discussion

Visualization of CASMI results: clustering

To visualize the CASMI 2016 results together, a hierarchical clustering was performed. The heat map of the negative mode challenges (1–81, excluding Team Dührkop) can be seen in Fig. 1, while the heat map of the positive mode challenges (82–208) is given in Fig. 2. These are discussed below; in addition interactive plots are provided

Table 5 Contribution of Metadata to the results

	RT	MoNA	Lowest CSID	Refs	Combined MS/MS	Combined MS/MS+RT+Refs	Combined MS/MS+RT+Refs+MoNA
Top 1	1	70	113	143	82	164	164
Top 3	5	87	158	177	126	183	187
Top 10	20	104	177	196	166	194	195
Mean rank	504.5	238.3	37.7	3.0	13.4	3.9	3.7
Median rank	135	10.25	1	1	2	1	1
Mean RRP	0.576	0.780	0.959	0.995	0.955	0.990	0.991
Median RRP	0.630	0.977	1	1	0.998	1	1

The first four columns contain submissions formed using just one type of metadata, the “Combined MS/MS” column was formed by equally weighting all Category 2 entries from Table 2, while the last two columns combined this with retention time and references without and with MoNA, respectively. The best value per statistic is marked in bold.

Table 6 Comparison of Categories 1, 2 and 3 results for the overlapping challenges in Category 1

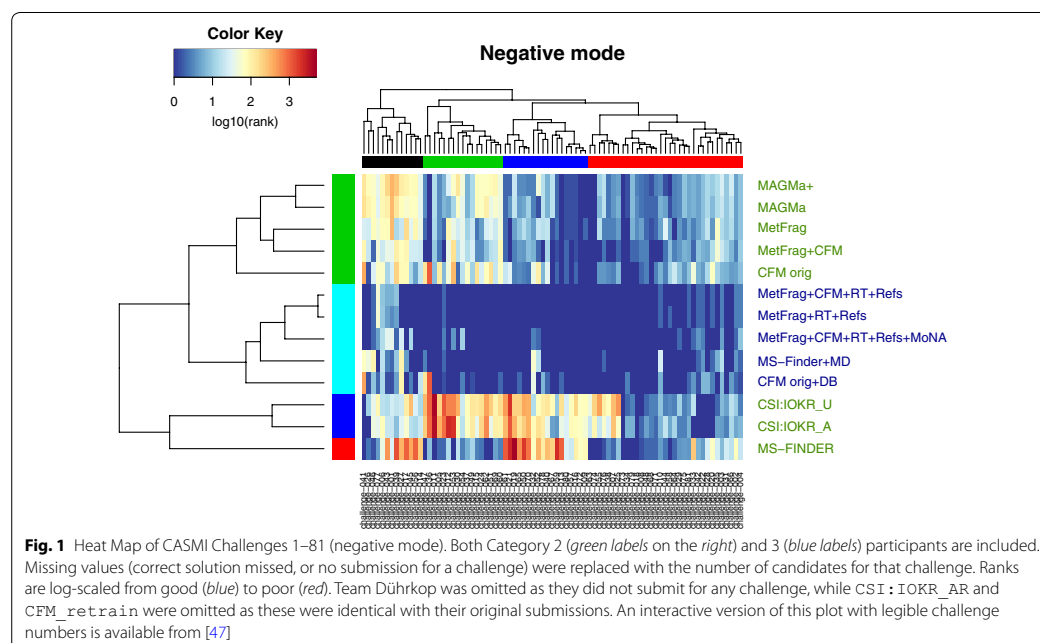
Chal.	Median rank of correct candidate per Category				Number of valid entries per category			Minimum and maximum rank of correct candidate per category (min, max)		
	All	1	2	3	1	2	3	1	2	3
10	1	1	19.5	1	14	12	6	(1, 15)	(11, 63)	(1, 1)
11	9	2	21	2	11	12	6	(1, 175)	(2, 208)	(1, 9)
12	1.5	1	16	1.5	15	11	6	(1, 88)	(1, 299.5)	(1, 8)
13	3	2	20	3.5	8	12	6	(1, 146)	(1, 270)	(1, 87)
14	25	23	26.5	20	11	12	6	(2, 292)	(17, 164.5)	(12, 144)
15	1	1	1.25	1	12	10	6	(1, 4)	(1, 6)	(1, 3)
16	2.5	2	25	2	12	9	6	(1, 25)	(14, 288)	(1, 14)
17	1	1	2.5	1	10	10	6	(1, 3)	(2, 5)	(1, 1)
18	11	4	19.5	2	9	10	6	(1, 34.5)	(3, 50)	(1, 11)
19	1	1	4.5	1	12	10	6	(1, 3)	(1, 7.5)	(1, 1)

The median ranks of Categories 1 and 3 (highlighted) are remarkably similar

(see reference links provided in the captions) for readers to investigate these clusters in more detail. Corresponding clusters excluding challenges in the training sets are available in Additional file 1: Figures S5 and S6.

The dark blue areas in Fig. 1 indicate very good ranking results. It is clear for the negative spectra that the meta-data (Category 3) really improved performance, with very few yellow or red entries for the Category 3 participants, which all grouped together in the cyan cluster (middle

left), indicated by the dark blue participant names (middle right). What is also clear is that all methods were very good for most of the compounds in the red challenge cluster (shown at the top, right-most cluster). The combinatorial fragmenters and CFM also performed well on the dark blue challenge cluster (second cluster from right)—in contrast both MS-FINDER and the CSI:IOKR methods struggled for these challenges, shown with the yellow to red coloring in the heat map.

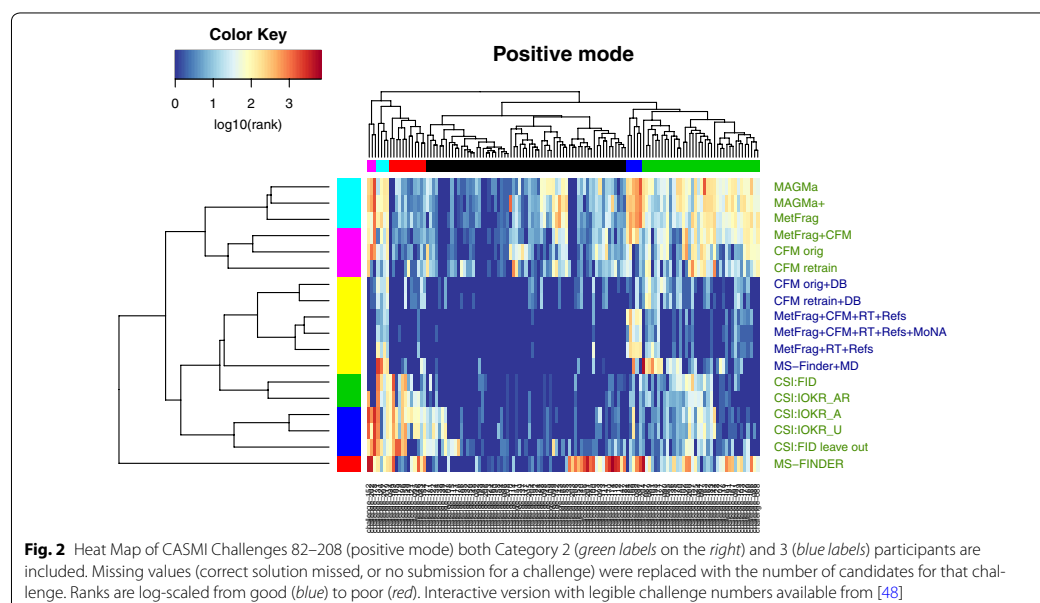


MS-FINDER outperformed other Category 2 approaches in the green challenge cluster (second from left)—showing the complementarity of the different approaches. This is reinforced by the fact that MS-FINDER was split into a participant cluster on its own and also explains partially why the Combined MS/MS entry performed better than all individual participant entries. For the clusters of challenges (top), the mean candidate numbers per cluster were (left to right): black (611), green (1603), blue (1019) and red (380), compared with a mean overall of 816. Both the red (“good” overall performance) and black (“poor”) clusters have mean candidates below the overall mean, whereas the poorly performing green cluster had mean candidates well above the overall mean. Thus, candidate numbers are not the only driver of performance.

Looking at individual challenges, all machine learning approaches performed poorly for Challenge 36, which was a 3 peak spectrum of a substance typically measured in positive mode (see Additional file 1: Figure S8). The combinatorial approaches performed poorly for Challenge 41 (see Additional file 1: Figure S9), monobenzyl phthalate, where the main peak is a well-known rearrangement that is not covered by these approaches. For this challenge, both CSI:IOKR and MS-FINDER performed well, indicating that this substance is in the training data domain (many phthalate spectra are in the open domain) and that MS-FINDER interprets the spectrum beyond combinatorial methods. The compounds in the

dark blue and green challenge clusters are likely not to be covered too well in the training data for CSI:IOKR. While it appears that MS-FINDER performs very poorly for some challenges, this is in fact an artifact of their submissions; for all the red entries in the heatmap, either the correct answer was absent from their submission (as they took only the top 500 candidates—this applied for 15 challenges) or no answer was submitted (5 challenges). In these cases the total number of candidates was used for the clustering. Removing the challenges where no submission was made from the clustering did not drastically alter any of the outcomes discussed above.

The positive mode cluster (Fig. 2) revealed an even darker blue picture (and thus generally very good results) than the negative mode cluster. The large dark blue patch in the middle of the heatmap indicates that for the majority of challenges, largely those in the black challenge cluster (top, middle), both the metadata but also the more extensive training data in positive mode for the machine learning approaches ensured that many Top 1 ranks were achieved. This is also shown well in the green challenge cluster, where the improvements that the metadata and machine learning add beyond the combinatorial approaches can be seen moving down and getting darker from the generally yellow top right corner. As for negative mode, the mean candidate numbers per challenge cluster were calculated (left to right): magenta (5297), cyan (1029), red (886), black (1534), blue (978), green



(722), with an overall mean of 1281. The performance for the magenta, cyan and blue clusters were all relatively “poor”, yet only the magenta cluster contained mean candidate numbers far above the overall mean. The combinatorial fragmenters performed poorly for the green cluster, which had mean candidate numbers below the overall mean. As mentioned above, candidate numbers are again not the only driver of performance. Investigations into other parameters that may influence the challenge clusters, such as number of peaks in the spectra, revealed similarly inconclusive results.

In contrast to negative mode, several participant clusters were formed in positive mode. The top two clusters contained the combinatorial fragmenters *MAGMa*, *MAGMa+* and *MetFrag*, which clustered apart from the *CFM-ID* entries, either alone or in combination with *MetFrag*. Below this was one very large cluster with all Category 3 entries (metadata, yellow). This is followed by three smaller clusters, one in green with the two best *CSI* entries (*CSI:FID* and *CSI:IOKR_AR*), one blue cluster with the remaining *CSI* entries, followed by *MS-FINDER* by itself. Note that *MS-FINDER* still clustered by itself in both positive and negative mode, even when compensating for the challenges with no submission, as mentioned above. This is due in part to their strategy to only select the top 500—again for the vast majority of the red *MS-FINDER* entries in the heat map either the correct candidate was missing in the submission (29 challenges in positive mode), or no submission was made (4 challenges). However, their location in a separate cluster is also possibly due to the fact that *MS-FINDER* does indeed use a different approach to fragmentation than either the combinatorial fragmenters or the machine learning approaches.

The challenge clusters revealed some interesting patterns: four small clusters contained challenges that were problematic for different approaches. Most metadata-free methods performed poorly for the pink cluster (challenges 152, 202, 178); all approaches performed relatively poorly for the cyan cluster adjacent (challenges 131, 126, 207 and 119). The challenges in the red cluster were likely reasonably dissimilar to the other substances in the machine learning training sets, as the combinatorial fragmenters outperformed the *CSI* approaches clearly in this cluster. The machine learners performed well on the dark blue cluster (challenges 184, 168, 199, 92, 197), where surprisingly the metadata even failed the combinatorial fragmenters. Three of these (92, 168, 199) involve breaking an amide bond, which may be something for these approaches to investigate further. Challenge 197 is a fused N heterocycle with one fragment. Spectra of these challenges, with additional comments, are available in Additional file 1: Figures S7–S20.

Visualization of CASMI results: candidate numbers and raw scores

Additional plots have been included in Additional file 1 to provide further visualization of the results. Additional file 1: Figure S21 shows the number of candidates for each challenge, ordered by the number of candidates versus the results for all CASMI entries (during and post-contest). Interestingly, fewer Top 1 entries and higher median/mean ranks were observed for the challenges with moderate candidate numbers (200–1000 candidates); lower median ranks and more Top 1 entries were observed for lower and higher candidate numbers. Additional file 1: Figures S22–S30 show the raw scores for selected submissions per participant and category, in order: *MAGMa+*, *CSI:IOKR_A*, *CSI:FID*, *CFM_orig*, *CFM_retrain+DB*, *MS-FINDER*, *MS-FINDER+MD*, *MetFrag* and *MetFrag+CFM+RT+Refs+MoNA*. These reveal interesting differences in the raw data behind each submission, including for instance the influence of training data availability on the positive and negative challenge results for *CSI:IOKR_A*, the metadata step function in *CFM_retrain+DB* as well as the effect of score scaling on *MetFrag*.

Machine learning approaches and training data

The CASMI2016 results show very clearly how the training data influences the performance of different approaches. The difference in Top 1 positive mode ranks between *CSI:IOKR_A*, 62 and *CSI:FID*, 70 (see Table 2) were due to the different training sets used, the *CSI:IOKR_AR* results (retrained on the same data as *CSI:FID*) had 69 Top 1 ranks. The results for *CSI:IOKR* in negative mode were also generally worse than all other approaches, which shows that the decision of Team Dührkop not to submit entries due to a lack of training data was quite well justified (even though it likely cost them the overall contest “win” for Category 2).

Team Dührkop noted that there was a large overlap between the challenges and their training set and investigated this with the *CSI:FID_leaveout* entry (described in the methods). For the sake of interpretation in this manuscript, this entry was updated post-contest once the exact solutions were known to make it a true “leave out” analysis. Although the performance was reduced compared with *CSI:FID* (36 vs. 70 Top 1 ranks in positive mode), the *CSI:FID_leaveout* entry still had more Top 1 ranks than any other non-*CSI* method in the contest (for positive mode only).

Following the idea of Team Dührkop, the CASMI results were evaluated for all participants on only those challenges where no contestant had the correct candidate in their training sets. Teams Dührkop, Allen and Brouard provided comprehensive lists of their training sets. These

were used to determine the overlap between all training sets and the CASMI challenges. The results over those challenges that were not in *any* training set (44 positive and 43 negative challenges) are given in Table 7.

The general observations made on the full contest data are supported by this reduced dataset as well, despite the unsurprising fact that the results on this reduced dataset were generally worse than the official contest results (see Table 2). This demonstrates that, as expected, machine learning methods do better on compounds from within their training sets (for example, the percentage of maximum Top 1 ranks dropped from 34 to 18%). Although the median ranks were worse, the Top 10 ranks still remained around 40–50% for most methods. Cluster plots on this reduced dataset for negative and positive mode, given in the supporting information (Additional file 1: Figures S5, S6), show similar patterns to the cluster plots on the full dataset.

Interestingly, these results show that the CSI:FID_leaveout entry outperformed CSI:FID, while CSI:IOKR_A also outperformed CSI:IOKR_AR, the retrained dataset, also for some different scores—similar observations could be made for CFM_orig versus CFM_retrain. While this could be a potential sign for overfitting, this is a small dataset and some or all of these observations could be due to fluctuations in the data. Overfitting is a potential problem that developers, especially of non-standard machine learning methods should test for, e.g. by checking if their performance decreases significantly for compounds which are structural dissimilar to compounds in the training data. These results highlight just one means by which the choice of training set can influence the performance of automated methods. The training set can also impact challenge results in a range of other ways that are harder to disambiguate. One training set may be more or less compatible with the challenge set, even after common compounds are removed. This suggests the importance of assessing automated methods using the same training set, where at all possible.

Metadata and consensus identification

The dataset for CASMI 2016 was predominantly well-known anthropogenic substances and as a result there are many distinct and highly referenced substances in the candidate lists. This is shown in the huge improvement that the metadata made to the ranking performance (Tables 4, 5). Figure 3 shows clearly that the vast majority of substances were either ranked first or second based purely on the reference count, with most other candidates having much lower counts. Figure 4 gives an overview of the contribution the metadata made to each approach based on the CASMI 2016 entries,

merging team results in the case of MS-FINDER. In the environmental context, it is quite common to search an exact mass or formula in databases such as ChemSpider, where e.g. the highest reference count as well as the substance with the “lowest CSID” are often picked as the most promising hit in many cases, discussed e.g. in [49]. The success with these strategies would have been quite considerable with this dataset. However, for new (emerging) anthropogenic substances and transformation products of known chemicals, these strategies would not work so well as they would have neither a high reference count nor a low database identifier. This situation is also likely to be drastically different for natural products and metabolites, where many more closely-related substances or even isomers could be expected.

The metadata results in Category 3 show that the importance of the sample context cannot be ignored during identification, especially for studies looking to find well-known substances. This is also highlighted by the comparison with the approaches used in Category 1, where also manual and semi-automatic approaches were considered. The current reality is that most automated approaches still depend on retrieving candidates from compound databases containing known structures—i.e. the situation replicated in this CASMI contest. Compound databases such as the Metabolic *In Silico* Network Expansion Databases (MINEs) [50] could be used as alternative sources of candidates for predicted metabolites in the metabolomics context, but would have had limited relevance in this contest.

While metadata, the way it was used here, will not help in the case of true unknowns, there are two cases to consider for automated approaches at this stage. For “unknowns” that happen to be in a database almost accidentally (e.g. a to-date unknown transformation product), the automated fragmentation approaches are very useful, because these structures can be retrieved from substance databases. However for true “unknown unknowns” that are not in any database, fragmenters could only be used in combination with structure generation, which is still impractical with the quality of data and methods at this stage unless candidate numbers can be restrained sufficiently. These cases are often extremely difficult to elucidate using MSⁿ alone and the information from additional analysis such as NMR will usually be necessary.

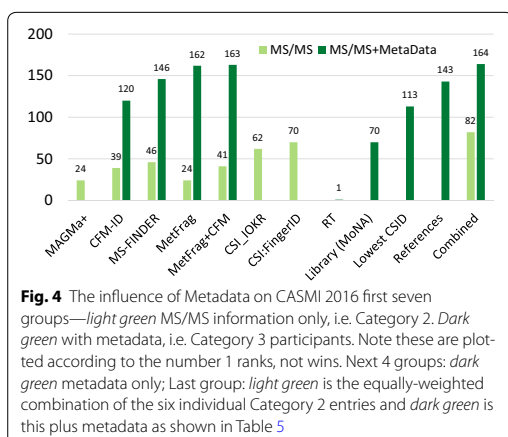
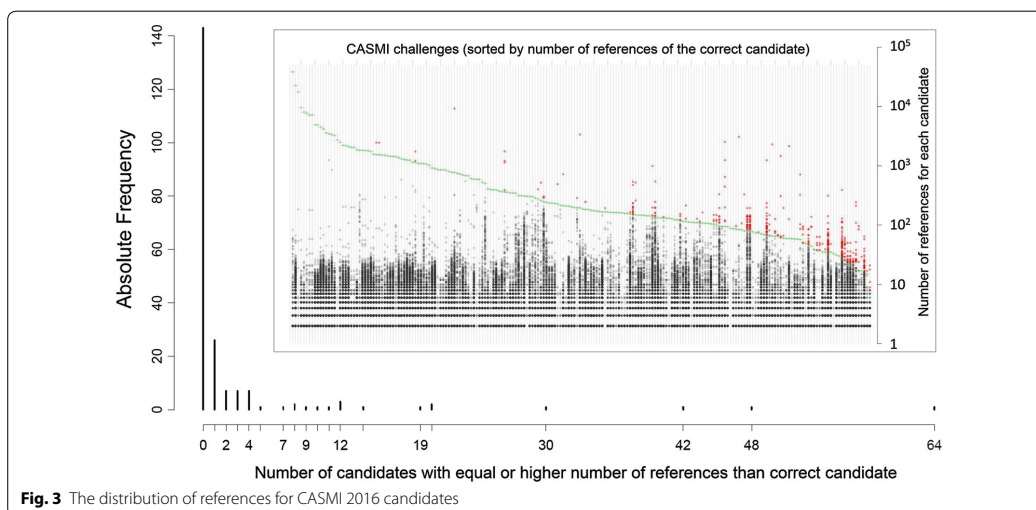
Stereoisomerism is another aspect of identification that was not covered in this contest. None of the current approaches are able to distinguish stereoisomers (even cis/trans isomers) using only MS/MS information for known unknowns. The evaluation of this contest addressed this by taking the best scoring stereoisomer and eliminating others (see “Methods”) to reduce the

Table 7 Global leaveout analysis for additional Category 2 entries—including only challenges where the correct answer was not in any training set

	Allen			Brouard			Dührkop			Ruttkies			Vaniya			Verdegem	
	CFM_orig	CFM_re-train	CSI_IOKR_A	CSI_IOKR_AR*	CSI_IOKR_U	CSI:RID_orig	CSI:FID_leaveout*	MetFrag*	MetFrag+CFM*	MS-FINDER	MAGMa+	MAGMa*					
Top 1 Neg.	6	6	6	6	4	0	0	4	10	7	4	3					
Top 1 Pos.	4	9	10	7	9	13	13	1	3	3	2	2					
Top 1	10	15	16	11	9	13	13	5	13	10	6	5					
Top 3	23	24	26	27	17	23	23	16	27	25	16	14					
Top 10	46	40	46	40	25	32	32	39	47	38	35	35					
Mean rank	52.57	64.05	106.5	97.84	99.92	52.81	41.48	68.38	37.16	28.7	76.75	100.4					
Med. rank	10	12.5	8	10	12	7	3	14.5	8	7.5	23.5	20.5					
Mean RRP	0.863	0.872	0.849	0.856	0.837	0.891	0.91	0.863	0.878	0.738	0.832	0.811					
Med. RRP	0.966	0.961	0.963	0.967	0.956	0.981	0.993	0.942	0.972	0.806	0.924	0.902					
Gold	18	21	26	26	19	11	17	7	18	18	10	9					
F1 score	628	654	735	691	632	403	557	484	707	594	462	434					
Medal Sc.	79	94	105	98	91	59	87	50	95	85	46	46					

n = 43 (negative) and n = 44 (positive)

The best value for selected statistics is marked in bold



influence of stereoisomers on the ranking results. However, for electron ionization (EI) MS it is already possible to distinguish stereoisomers in some cases using ion abundances. This is an aspect that should be developed in the future for MS/MS once the spectrum generation is sufficiently reproducible to allow this. Coupling with suitable chromatography will potentially enhance the ability to distinguish between stereoisomers further.

Evaluating methods and winner declaration

Contests such as CASMI always generate much discussion about how the winner was evaluated and declared;

this year's contest was no exception. A “contest” setting is different to the way individual methods compare their performance with others and this is the role of CASMI—to look at the approaches in different ways, relative to one another. One change in CASMI 2016 was to use the “average rank” instead of the “worst-case” rank to account for equal candidate scores, as participants pointed out that for previous contests one could add small random values to break tied scores and improve results in the contest. There will be several cases where candidates are indistinguishable according to the MS and it is important to capture this aspect in CASMI. While equal scores may make most chemical sense in these cases, computational methods deal with this differently; some report equal scores, others generate slightly different scores for effectively equal candidates. The average rank deals with this better than the “worst-case” rank, but can now disadvantage methods that report equal scores compared with others, as the chances are that at least one other method will beat it each time.

The criteria for declaring the winner in this contest was that the best performing participant(s), i.e. the winner, was defined per challenge and then the wins were added to determine the overall winner. This allows the declaration of a winner per challenge, irrespective of the actual performance (i.e. the winner could have rank 100, if all other participants were worse). The drawback of this approach is that it creates cross-dependencies between participants, i.e. the removal (or addition) of one participant completely changed the rank of the other participants. CFM likely suffered from

this, as a machine-learning approach with similar training set coverage to CSI, which allowed the complementary approach of MS-FINDER to claim third place ahead of CFM. An alternative approach could be to look at this in terms of overall success and say that if a team had the correct structure as the 20th hit and other teams were even worse, none of the approaches were really sufficient to the task and nobody should then earn a 'win'. This may reflect real structure elucidation cases better, where investigators would likely also consider the Top 3, Top 5, or maybe even Top 10 structures, but is perhaps not so good to declare a winner in a contest as some (difficult) challenges would have no "winner" and the performance of methods on difficult challenges is also an important aspect of the contest. This idea was investigated in this publication by also providing the Top 1, Top 3, Top 10 ranks per participant, as well as the Formula 1 Score (scaled Top 1–10 results) and Medal Score, where the medal count is based on Top 1, 2 and 3 ranks. The results of these metrics confirm the overall pattern observed in the contest: the two CSI teams outperformed all others in Category 2, followed by either MS-FINDER or CFM depending on exactly which score was used. In other words, the approaches have made fantastic progress, are complementary to one another but actually quite difficult to tell apart. Although 208 challenges is an order of magnitude in terms of challenge numbers above previous CASMIs, these numbers are still quite small and almost random differences between the methods resulted sometimes in large changes in the various scores, as shown with the different CSI entries.

Participant perspectives

Team Allen submitted two alternative versions of CFM, the main difference being that for CFM_retrain version, additional training data was added from the 2014 NIST MS/MS database. While the addition of extra training data may have been expected to improve the results, this appears not to have been the case for this competition. One possible reason for this is that the additional data were generally of poorer (often integer) mass accuracy as compared to that used to train the original CFM model. This required a wider mass tolerance (0.5 Da) to be used during the retraining (compared to 0.01 Da previously), which may have hindered the training algorithm from accurately assigning explanations to peaks, and so modeling their likelihoods. This highlights that while the production of larger, more comprehensive data sets is likely crucial for better training of automated methods, the quality of these data sets is also very important. Most automated methods would likely benefit from training on cleaner data with better mass accuracies.

Team Dührkop investigated how CSI:FingerId compared with a direct spectral library search. A spectral library containing all structures and spectra used to train CSI:FingerId was created and searched with a 10 ppm precursor mass deviation. The resulting spectra were sorted via cosine similarity (normalized dot product), again with 10 ppm mass accuracy. Candidates were returned for 91 of the 127 (positive mode) challenges; the correct answer was contained in the library for 69 of these. The spectral library search correctly identified 63 of the 69 structures in total, 40 of these were "trivial" (the correct answer was the only candidate). On average, candidate lists for the spectral library search contained only 2.4 candidates, which was almost three orders of magnitude below the average CASMI candidate list of 1114 candidates. The cosine product between the challenge spectrum and the corresponding training spectrum of the same compound was only 0.76 on average; for one challenge it was below 0.01. For example, the cosine similarity between the spectrum for Challenge 202 (Pendimethalin) and the training spectrum was only 0.137, but it was still "correctly identified" as it was the only candidate with this precursor mass. This compound was correctly identified in the original CSI:FID submission, and ranked 569 for the CSI:FID_leaveout submission. This indicates that CSI:FingerId and other machine-learning approaches are capable of learning inherent properties from the mass spectra, beyond simple spectral similarity.

Team Vaniya The CASMI Category 2 contest was a reshuffling contest: potential structures were given to all participants, listing one to over 8000 potential structures for each challenge. These structures were within 5 ppm mass accuracy and often included different elemental formulas. Therefore, Category 2 was a 'structure dereplication' contest, finding the best structure within a pre-defined list of structures, not a completely open *in silico* test on all exhaustive structures in the chemosphere. In practical terms, it is important to note that an *in silico* software does not eliminate the time consuming aspects of data preparation, formatting, and interpretation. Counting the computing power and manual effort between two people, it took about 24 h to complete the 208 challenges for the MS-FINDER submission.

From Table 2, one could say that MS-FINDER was best based on the mean rank (19.75), but ranks lower than 10 are less relevant in reality. While MS-FINDER had almost 50% of the challenges within the top 10 ranks, so did every other software (or team). In reality, no chemist would use a software without any database or mass spectral library behind it. The importance of using *a priori* knowledge is seen by Team Allen's submission that improved the Top 1 correct structure hits from 39

to 120 challenges in Category 3, a bit more than 50% of the challenges. Hence, we conclude that the glass is half full: if only *in silico* methods are used, some 50% of the challenges are within the top 10 hits within the structures given by the CASMI organizers. However, many challenges would score much higher if other metadata are used, e.g. constraining the search database to particular classes of compounds that can be expected for a specific study. Which parameters need to be optimized, and which *a priori* metadata should be used? Those questions may be answered in a more tailored future CASMI contest.

Team Verdegem participated in Category 2 of the CASMI 2016 contest with *MAGMa+*, which is a fast, plug-and-play method relying on combinatorial fragmentation without requiring a preliminary training phase for improved performance. The entire submission, including scripting for automation and single core calculations, took less than 1 day. *MAGMa+* outperformed *MAGMa*, showing the use of the parameter optimization performed to improve several second and third ranked candidates to first place. *MAGMa+* shared the best ranking for 44 of 208 challenges (see Table 2) and performed considerably better than other contestants for nine of those challenges (21, 32, 36, 40, 52, 61, 121, 157 and 189), indicating the relevance of the underlying algorithm.

Since *MAGMa+* outperformed *MAGMa* according to some (e.g. number of gold medals, Top 1 and 3 ranks) but not all metrics, further more advanced parameter optimizations are planned to achieve a more global performance improvement. However, further improvements to the performance of *MAGMa/MAGMa+* will require interventions of a different kind. The performance of *MAGMa+* decreases with increasing candidate numbers (in this contest 1116 on average after the removal of duplicate stereoisomers), however, in case of smaller numbers, it starts to outperform some of the other methods [25, 42]. For untargeted metabolite identification in biological/biomedical setups, it is arguably more suitable to restrict the candidate structure database to those metabolites known to exist in the organism under study, e.g. using only the $\approx 42,000$ metabolites currently present in the HMDB [21] for samples of human origin. This was noted also in previous CASMI contests [2]. Many candidate structures had identical scores with *MAGMa+*, resulting in the correct matches being given lower ranks according to the evaluation rules. Whereas on average 1098 structures were retained from the structure database based on the parent mass match, only 616 different score values were observed (on average). *Team Verdegem* will investigate more discriminative scoring options for *MAGMa+* in the future.

Conclusions

This was the first CASMI contest to use a large set of challenges, targeted especially at the automated methods. This decision was taken on the basis of feedback from several representatives at the 2015 Dagstuhl seminar in Computational Metabolomics [51], to allow a statistically more robust comparison of the methods. The decision to provide candidates this year was also on the basis of Dagstuhl discussions, to eliminate the data source as an influence on the contest outcomes and thus focus more on the role of the *in silico* fragmentation approaches themselves.

From the perspective of the organizers, it was a great success to have participants contribute from each of the major different approaches; *MetFrag* was added internally for the sake of completion as this was not otherwise represented and allows this paper to complement the work in [25] on a different dataset. Very interesting and constructive discussions have resulted from choosing to prepare this article with “all on board” and the post-contest analysis has been instrumental in teasing apart some of the differences between the actual contest results.

The contest winners, **Team Brouard** with *CSI:IOKR_A* in Category 2 and **Team Kind** with *MS-FINDER+MD* in Category 3 prove that the latest developments in this field have indeed resulted in great progress in automated structure annotation. Despite the very large candidate sets, the majority of methods achieved around 50% in the Top 10, which is very positive for real-life annotation, especially with an outlook to higher-throughput untargeted analysis. The combination of the Category 2 submissions resulted in even better overall performance than each individual method, indicating the complementarity of the approaches and supporting the potential use of *consensus* fragmentation results as has been shown earlier for fragmenters [12, 52] and also recently for toxicity modeling using a more sophisticated weighting than that attempted here [53]. The role of the metadata and comparison with Category 1 shows that sample context cannot be ignored during identification.

In this contest, few participants used the CASMI training set provided, which was also a suggestion from Dagstuhl. In the end this was too “big” for pure parameter optimization (where a few spectra may suffice), but too small for serious method training. *Team Brouard* added it to their other training data in their original submissions, while it was used to determine the score weights in the *MetFrag* entries. *Team Vaniya* did not use this for *MS-FINDER* to avoid over-training; *Team Allen* due to a lack of time. One conclusion from the post-contest evaluation is that future CASMIs could consider providing an extensive, open training dataset (e.g. the GNPS/MassBank collection used by *CSI:FID*) and ensure all CASMI challenges are absent from this set. This

would, however, force all machine-learning approaches to retrain their methods prior to submission. Another option is that the organizers would have to ensure that all challenges are outside all available datasets—which is possible but also difficult with the number of private and closed collections available. A compromise could be to ensure that a sufficient majority of the candidates are outside the “major” mass spectral resources, with some overlap to ensure sufficient challenges are available (finding data sources for CASMI is a challenging task!) and require participants to submit InChIKey lists of their training sets with their submissions; as done with Teams Allen, Brouard and Dührkop post-contest here.

Challenges for future contests remain true unknowns, i.e. substances that are not present in compound databases. This would currently be feasible for manual approaches and was attempted already once in CASMI 2014, Challenges 43–48 [54], albeit with limited success. Automated approaches would need either a metabolite database such as MINEs [50] or structure generation [55], but finding sufficient appropriate data for an automated category will also be a challenge for the contest organizers, let alone the participants! The ability to distinguish stereoisomers using MS/MS alone also remains a challenge for the future that is not yet ripe enough for a CASMI contest; distinguishing (positional) isomers is likely sufficient challenge for the next few years.

The huge improvements in machine learning approaches will continue as more training data becomes available—the more *high quality* data with likewise *high quality* annotations that becomes available in the open data domain will ensure that the best computational people can work on the best identification methods. The complementarity of the chemistry behind MS-FINDER and the machine learning behind CSI shows that developments in both directions will carry the field forward.

The “take home” messages of CASMI 2016 are:

- The latest developments in the field, CSI:IOKR and MS-FINDER were well-deserved winners of Categories 2 and 3, respectively.
- The complementarity of different approaches is clear; combining several *in silico* fragmentation approaches will improve annotation results further.
- The best methods are able to achieve over 30% Top 1 ranks and most methods have the correct candidate in the Top 10 for around 50% of cases using fragmentation information alone, such that the outlook for higher-throughput untargeted annotation for “known unknowns” is very positive.
- This success rate rises to 70% Top 1 ranks (MS-FINDER) and 87% Top 10 ranks (CFM) when including metadata.
- The machine learning approaches clearly improve with larger training data sets—the more high quality annotated, open data that is available, the better they will get.
- Developments that focus on the chemistry such as MS-FINDER are also essential, especially to cover the cases where no training data is available.
- Despite the above, several challenges remain where the simple combinatorial approach of MetFrag and MAGMa still performs best.
- Improved incorporation of experimental “metadata” will increase annotation successes further, especially for large candidate sets.
- Challenges for future contests remain true unknowns, assessing the ability of methods to distinguish positional isomers and eventually also stereoisomers.

Finally, a big thank you to all those who participated in CASMI 2016 in any way, shape or form and keep an eye on the CASMI website [1] for future editions.

Availability and requirements

- Project name: CASMI
- Project home page: <http://www.casmi-contest.org/>
- Operating system(s): Platform independent
- Programming language: Various
- License: N/A
- Any restrictions to use by non-academics: none.

Additional files

Additional file 1. This file contains additional content (methods, results and selected spectra) to complement the manuscript. See PDF for details.

Additional file 2. Table A1 ESD file used in MS-FINDER version 1.62 for a total of 220,212 compounds. Additional columns for InChIKey, short InChIKey, exact mass, formula, SMILES are not shown here. The use of N/A and a database identifier represents the presence or absence of a compound in a given database. For example, 1,3-butadiyne is only present in ChEBI database (CHEBI:37820). This ESD file was replaced by a dummy file where all HMDB identifiers were modified to dummy identifiers AV001... AV00n and all other identifiers replaced by -1 or N/A. Table A2: Formatted ESD file for CASMI 2016 Category 2 Challenge-001. The first 10 compounds from the candidates list for Challenge-001 are listed above. Columns for InChIKey, short InChIKey, PubChem CID, exact mass, formula, SMILES are shown in this table. Databases from BMDB through PubChem are replaced by dummy information.

Abbreviations

CASMI: Critical Assessment of Small Molecule Identification; CSI:IOKR: Compound Structure Identification:Input Output Kernel Regression; MS/MS: tandem mass spectrum; ESI: electrospray ionization; HCD: higher-energy collisional dissociation; LC-MS: liquid chromatography coupled to mass spectrometry; [M+H]⁺, [M-H]⁻: protonated and deprotonated molecular ions; SPLASH: SPectral hASH; MGF: Mascot Generic Format; SMILES: Simplified Molecular Input Line Entry System; InChI, InChIKey: IUPAC International Chemical Identifier and (hash) key; CSV: comma-separated values; MS1: full scan mass spectrum; RRP: relative ranking position; CFM-ID: Competitive

Fragmentation Modeling for Metabolite Identification; NIST: National Institute of Science and Technology (USA); HMDB: human metabolome database; ChEBI: Chemical Entity of Biological Interest; CSI:FID: Compound Structure Identification; FingerID; IOKR: Input Output Kernel Regression; (Uni-)MKL: (Uniform) Multiple Kernel Learning; CDK: Chemistry Development Kit; GNPS: Global Natural Products Social Networking; SVM: support vector machine; Q-TOF: Quadrupole Time of Flight; HR: hydrogen rearrangement; GUI: graphical user interface; MoNA: MassBank of North America; ESD: Existing Structure Database; CSIDs: ChemSpider Identifiers; RT: retention time; MINEs: Metabolic *In Silico* Network Expansion Databases; EI-MS: electron ionization mass spectrometry.

Authors' contributions

ES and SN jointly organized Categories 2 and 3 of CASMI 2016; MK selected the challenge compounds and recorded the spectra; ES wrote the majority of the manuscript, SN performed the majority of the automatic evaluation. CR prepared the additional post-contest results. All participants (CB, TK, KD, FA, AV, DV, SB, JR, HS, HT, TS, OF, BG) contributed via their submissions and comments/contributions to the manuscript. All authors read and approved the final manuscript.

Author details

¹ Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland. ² Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, 06120 Halle, Germany. ³ Department of Effect-Directed Analysis, UFZ: Helmholtz Centre for Environmental Research, Permoserstrasse 15, 04318 Leipzig, Germany. ⁴ Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland. ⁵ Helsinki Institute for Information Technology, Tekniikantie 14, 02150 Espoo, Finland. ⁶ West Coast Metabolomics Center and Genome Center, University of California Davis, 451 Health Sciences Drive, Davis, CA 95616, USA. ⁷ Chair of Bioinformatics, Friedrich-Schiller-University, Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany. ⁸ Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E9, Canada. ⁹ Department of Chemistry, University of California Davis, One Shields Avenue, Davis, CA 95616, USA. ¹⁰ Metabolomics Expertise Center, Vesalius Research Center (VRC), VIB, KU Leuven – University of Leuven, 3000 Louvain, Belgium. ¹¹ RIKEN Center for Sustainable Resource Science (CSRS), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. ¹² Department of Biochemistry, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi Arabia.

Acknowledgements

CASMI is a joint effort and we wish to thank all organizers and participants of previous CASMIs as well as other interested parties for valuable discussions. For this manuscript, AV wishes to acknowledge the contributions of Stephanie N. Samra, Sajjan S. Mehta, Diego Pedrosa and Hiroshi Tsugawa. KD and team acknowledge the contributions of Marvin Meusel. FA and TS acknowledge the contributions of Russ Greiner and David Wishart. MK acknowledges the help of Janek P. Dann with spectral acquisition. We thank the reviewers for their comments and suggestions.

Competing interests

The authors declare that they have no competing interests.

Funding

CR, ES and MK acknowledge funding from the European Commission for the FP7 project SOLUTIONS under Grant Agreement No. 603437. CB, JR and HS acknowledge funding from the Academy of Finland (Grant 268874/MIDAS) and computational resources provided by the Aalto Science-IT project. SB acknowledges funding from the Deutsche Forschungsgemeinschaft (BO 1910/16). SN acknowledges basic institutional funding by the Leibniz Association. FA and team were funded by NSERC, AICML, AIHS, Genome Alberta, CIHR, The Metabolomics Innovation Centre (TMIC) and their work was carried out using the Compute Canada Westgrid facility.

Received: 14 December 2016 Accepted: 13 March 2017

Published online: 27 March 2017

References

1. Neumann S, Schymanski EL (2016) CASMI contest webpage. <http://www.casmi-contest.org>. Accessed 8 Dec 2016
2. Schymanski EL, Neumann S (2013) CASMI: and the winner is. *Metabolites* 3(2):412–439
3. Schymanski EL, Neumann S (2013) The Critical Assessment of Small Molecule Identification (CASMI): challenges and solutions. *Metabolites* 3(3):517–538
4. Nishioka T, Kasama T, Kinumi T, Makabe H, Matsuda F, Miura D, Miyashita M, Nakamura T, Tanaka K, Yamamoto A (2014) The winner of CASMI 2013 is... *Mass Spectrom* 3(Special Issue 2), 0039
5. Nikolic D, Jones M, Sumner L, Dunn W (2017) CASMI2014: challenges, solutions and results. *Current Metab*. doi:10.2174/2213235X04666160617113437
6. Genta-Jouve G, Thomas OP, Touboul D, Schymanski EL, Neumann S (2016) CASMI 2016: Category 1: Natural products. <http://www.casmi-contest.org/2016/results-cat1.shtml>. Accessed 20 Mar 2017
7. Neumann S, Schymanski EL (2016) CASMI contest rules and evaluation. <http://www.casmi-contest.org/2016/rules.shtml>. Accessed 8 Dec 2016
8. Stravs MA, Schymanski EL, Singer HP, Hollender J (2013) Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom* 48(1):89–99
9. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O (2016) SPLASH: The SPectral HaSH Identifier. <http://splash.fiehnlab.ucdavis.edu/>. Accessed 8 Dec 2016
10. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, EL Tomáš Schymanski, Willighagen EL, Wilson M, Wishart DS, Arita M, Dorrestein PC, Bandeira N, Wang M, Schulze T, Salek RM, Steinbeck C, Nainala VC, Mistrik R, Nishioka T, Fiehn O (2016) SPLASH, a hashed identifier for mass spectra. *Nat Biotechnol* 34(11):1099–1101
11. CASMI2016 Mass Spectra. <http://massbank.eu/MassBank/jsp/Result.jsp?ty=pe=&rcid=&idtype=site&srchkey=36>. Accessed 12 Dec 2016
12. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* 8(1):1
13. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. *J Cheminf* 3:33
14. Gerlich M, Neumann S (2013) MetFusion: integration of compound identification strategies. *J Mass Spectrom* 48(3):291–298. doi:10.1002/jms.3123
15. Neumann S, Schymanski EL (2016) CASMI contest challenges. <http://www.casmi-contest.org/2016/challenges-cat2+3.shtml>. Accessed 8 Dec 2016
16. Meusel M, Hufsky F, Panter F, Krug D, Möller R, Böcker S (2016) Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal Chem* 88(15):7556–7566. doi:10.1021/acs.analchem.6b01015
17. Formula One Scoring Systems. https://en.wikipedia.org/wiki/List_of_Formula_One_World_Championship_points_scoring_systems. Accessed 8 Dec 2016
18. Allen F, Greiner R, Wishart D (2014) Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*. doi:10.1007/s11306-014-0676-4
19. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27:747–751
20. NIST, EPA, NIH: NIST Mass Spectral Library 2014 Edition. U.S. Secretary of Commerce, USA
21. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorn Dahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A (2013) HMDB 3.0—the Human Metabolome Database in 2013. *Nucleic Acids Res* 41(D1):D801–D807
22. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, McNaught A, Alcántara R, Darsow M, Guedj M, Ashburner M (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 36(Suppl 1):344–350

23. Wishart DS (2016) FooDB. <http://foodb.ca/>. Accessed 8 Dec 2016
24. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36(Suppl 1):901–906
25. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci*. doi:10.1073/pnas.1509788112. <http://www.pnas.org/content/early/2015/09/16/1509788112.full.pdf>
26. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J (2016) Fast metabolite identification with input output kernel regression. *Bioinformatics* 32(12):28–36. doi:10.1093/bioinformatics/btw246. <http://bioinformatics.oxfordjournals.org/content/32/12/28.full.pdf+html>
27. Böcker S, Dührkop K (2016) Fragmentation trees reloaded. *J Cheminform* 8:5. doi:10.1186/s13321-016-0116-8
28. Shen H, Dührkop K, Böcker S, Rousu J (2014) Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics* 30(12):157–164
29. Cortes C, Mehryar M, Rostamizadeh A (2012) Algorithms for learning kernels based on centered alignments. *J Mach Learn Res* 13:795–828
30. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL (2006) Recent developments of the Chemistry Development Kit (CDK)—an open-source java library for chemo- and bioinformatics. *Curr Pharmaceut Des* 12(17):2111–2120
31. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E (2003) The Chemistry Development Kit (CDK): an open-source java library for chemo- and bioinformatics. *J Chem Inf Comput Sci* 43(2):493–500
32. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T et al (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 34(8):828–837
33. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, Oda Y, Kakazu Y, Kusano M, Tohge T, Matsuda F, Sawada Y, Hirai MY, Nakanishi H, Ikeda K, Akimoto N, Maoka T, Takahashi H, Ara T, Sakurai N, Suzuki H, Shibata D, Neumann S, Iida T, Tanaka K, Funatsu K, Matsuura F, Soga T, Taguchi R, Saito K, Nishioka T (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45:703–714
34. Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola AJ, Schölkopf B (eds) *Advances in large margin classifiers*, vol 5. MIT Press, Cambridge
35. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, Saito K, Fiehn O, Arita M (2016) Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal Chem* 88(16):7946–7958
36. Tsugawa H et al (2016) MS-FINDER. http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/index.html. Accessed 8 Dec 2016
37. NIST MS Search GUI. <http://chemdata.nist.gov/>. Accessed 8 Dec 2016
38. MassBank of North America. <http://mona.fiehnlab.ucdavis.edu/>. Accessed 8 Dec 2016
39. Sawada Y, Nakabayashi R, Yamada Y, Suzuki M, Sato M, Sakata A, Akiyama K, Sakurai T, Matsuda F, Aoki T et al (2012) RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* 82:38–45
40. LfU: Bayerisches Landesamt für Umwelt: STOFF-IDENT (login Required). <http://bb-x-stoffident.hswt.de/stoffidentjpa/app>. Accessed 13 June 2016
41. NORMAN Association: NORMAN Suspect List Exchange. <http://www.norman-network.com/?q=node/236>. Accessed 8 Dec 2016
42. Verdegem D, Lambrechts D, Carmeliet P, Ghesquière B (2016) Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. *Metabolomics* 12(6):1–16. doi:10.1007/s11306-016-1036-3
43. Ridder L, van der Hoof JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J (2012) Substructure-based annotation of high-resolution multistage MSn spectral trees. *Rapid Commun Mass Spectrom* 26(20):2461–2471. doi:10.1002/rcm.6364
44. MAGMa+. <https://github.com/savantes/MAGMa-plus>. Accessed 8 Dec 2016
45. MetFrag Command Line. <http://c-rukkies.github.io/MetFrag/projects/metfragcl/>. Accessed 8 Dec 2016
46. Royal Society of Chemistry: ChemSpider. <http://www.chemspider.com/>
47. Interactive Heat Map of CASMI 2016 Challenges Negative Mode. <http://www.casmi-contest.org/2016/heatmapNegCat2.html>. Accessed 8 Dec 2016
48. Interactive Heat Map of CASMI 2016 Challenges Positive Mode. <http://www.casmi-contest.org/2016/heatmapPosCat2.html>. Accessed 8 Dec 2016
49. McEachran AD, Sobus JR, Williams AJ (2016) Identifying “known unknowns” using the US EPA’s CompTox Chemistry Dashboard. submitted
50. Jeffries JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KE, Henry CS (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminform* 7(1):1
51. Böcker S, Rousu J, Schymanski E (2016) Computational metabolomics (Dagstuhl Seminar 15492). *Dagstuhl Rep* 5(11):180–192. doi:10.4230/DagRep.5.11.180
52. Schymanski EL, Gallampois CMJ, Krauss M, Meringer M, Neumann S, Schulze T, Wolf S, Brack W (2012) Consensus structure elucidation combining GC/EI-MS, structure generation, and calculated properties. *Anal Chem* 84:3287–3295
53. Mansouri K, Abdelaziz A, Rybacka A, Roncaglioni A, Tropsha A, Varnek A, Zakharov A, Worth A, Richard AM, Grulke CM, Trisciuzzi D, Fourches D, Horvath D, Benfenati E, Muratov E, Wedebye EB, Grisoni F, Mangiatordi GF, Incisivo GM, Hong H, Ng HW, Tetko IV, Balabin I, Kancherla J, Shen J, Burton J, Nicklaus M, Cassotti M, Nikolov NG, Nicolotti O, Andersson PL, Zang Q, Politi R, Beger RD, Todeschini R, Huang R, Farag S, Rosenberg SA, Slavov S, Hu X, Judson RS (2016) CERAPP: Collaborative estrogen receptor activity prediction project. *J Environ Health Perspect* 124(7):1023–1033
54. CASMI 2014 challenges. <http://www.casmi-contest.org/2014/results-cat2.shtml>. Accessed 8 Dec 2016
55. Kerber A, Laue R, Meringer M, Rücker C, Schymanski E (2014) Mathematical chemistry and chemoinformatics: structure generation, elucidation and quantitative structure–property relationships. Walter de Gruyter, Berlin

Eidesstattliche Erklärung / Declaration under Oath

This work was conducted from 05/2012 to 06/2019 under the supervision of Dr. Steffen Neumann at the Leibniz Institute of Plant Biochemistry (IPB) in Halle.

I declare under penalty of perjury that this thesis is my own work entirely and has been written without any help from other people. I used only the sources mentioned and included all the citations correctly both in word or content. This thesis has not previously been presented in identical or similar form to any other German or foreign examination board.

Halle (Saale), 14/12/2022

Christoph Ruttkies

Christoph Ruttkies

Date of birth 27/12/1985
Place of birth Merseburg

Education

04/2012 **Diploma in Bioinformatics**, *Martin Luther University Halle-Wittenberg*, Halle (Saale).
06/2005 **Abitur**, *Novalis Gymnasium*, Bad Dürrenberg.

Diploma thesis

Title *Developing computational methods for the identification of poly-functional crosslinked peptides*
Supervisor Prof. Dr. Dr. h. c. Reinhard Neubert
Description With the help of tandem mass spectrometry data, statistical methods were used for the characterization and identification of poly-functional crosslinked peptides.

Vocational experience

since 8/2018 **Bioinformatician**, *OntoChem GmbH*, Halle (Saale)
- Planning and implementing cloud infrastructure solutions
- Implementating Java backend applications
- Designing and implementing cheminformatics solutions for molecular compound and reaction search in large databases
5/2012 – 8/2018 **Research associate**, *Leibniz Institute of Plant Biochemistry*, Halle (Saale)
- Developing computational methods to assist the identification of small molecules using tandem mass spectrometry data
- Enhancing developed computational methods by combining data from different analytical and statistical methods
- Integrating scientific software applications in a cloud-based e-infrastructure
3/2008 – 3/2012 **Research Assistant**, *Institute of Computer Science, Martin Luther University Halle-Wittenberg*, Halle (Saale).
- Administration of inhouse high performance computing clusters
- Administration of inhouse linux servers and infrastructure

Selected publications

2019 Ruttkies C., Neumann S., Posch S.: *Improving MetFrag with statistical learning of fragment annotations*. BMC Bioinformatics 20:376, (2019).
2016 Ruttkies C., Schymanski E. L., Wolf S., Hollender J. & Neumann S.: *MetFrag relaunched: incorporating strategies beyond in silico fragmentation*. J Cheminform 8:3, (2016).

Halle (Saale), 02/05/2022