Scheubert *et al. Journal of Cheminformatics* 2013, **5**:12
http://www.jcheminf.com/content/5/1/12

Journal of Cheminformatics

REVIEW Open Access

Computational mass spectrometry for small molecules

Kerstin Scheubert[1*], Franziska Hufsky[1,2] and Sebastian Böcker[1]

Abstract

The identification of small molecules from mass spectrometry (MS) data remains a major challenge in the interpretation of MS data. This review covers the *computational* aspects of identifying small molecules, from the identification of a compound searching a reference spectral library, to the structural elucidation of unknowns. In detail, we describe the basic principles and pitfalls of searching mass spectral reference libraries. Determining the molecular formula of the compound can serve as a basis for subsequent structural elucidation; consequently, we cover different methods for molecular formula identification, focussing on isotope pattern analysis. We then discuss automated methods to deal with mass spectra of compounds that are not present in spectral libraries, and provide an insight into *de novo* analysis of fragmentation spectra using fragmentation trees. In addition, this review shortly covers the reconstruction of metabolic networks using MS data. Finally, we list available software for different steps of the analysis pipeline.

Keywords: Mass spectrometry, Metabolomics, Spectral library, Molecular formula identification, Structure elucidation, Fragmentation trees, Networks

Introduction

Mass spectrometry (MS) is a key analytical technology for detecting and identifying small biomolecules such as metabolites [1-3]. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Several analytical techniques have been developed, most notably gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). Both analytical setups have their advantages and disadvantages, see Section "Experimental setups" for details.

In recent years, it has been recognized that one of the most important aspects of small molecule MS is the automated processing of the resulting data. In this review, we will cover the development of computational methods for small molecule mass spectrometry during the last decades. Here, the term "small molecule" refers to all small biomolecules excluding peptides. Obviously, our review cannot be complete: In particular, we will not cover the "early years" of computational mass spectrometry of small molecules. First rule-based approaches for predicting fragmentation patterns, as well as explaining experimental mass spectra with the help of a molecular structure, were developed as part of the *DENDRAL* project that started back in 1965 [4-7]; see also Chapter 7 of [8]. Citing Gasteiger *et al.* [9]: "However, it is sad to say that, in the end, the *DENDRAL* project failed in its major objective of automatic structure elucidation by mass spectral data, and research was discontinued."

We will not cover methods that deal with processing the raw data, such as de-noising and peak picking, as this is beyond the scope of our review; see Section "Software packages" for a list of available software packages for this task. Furthermore, we do not cover the problem of aligning two or more LC-MS or GC-MS runs [10-13]. Finally, we will not cover computational methods that deal with the chromatography part of the analysis, such as predicting retention indices [14,15].

Structure confirmation of an unknown organic compound is always performed with a set of independent methods, in particular NMR. The term "structure elucidation" usually refers to full *de novo* structure identification of a compound, including stereochemical assignments. It is commonly believed that structure elucidation is impossible using MS techniques alone, at least without using strong background information. We will not cover this

*Correspondence: kerstin.scheubert@uni-jena.de
[1]Chair of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, Jena, Germany
Full list of author information is available at the end of the article

aspect, but concentrate on the information that MS experiments *can* give.

"Computational mass spectrometry" deals with the development of computational methods for the automated analysis of MS data. Over the last two decades, much research has been focused on methods for analyzing proteomics MS data, with literally hundreds of articles being published in scientific journals [16-21]. The proteomics field has benefited tremendously from this development; often only the use of these automated methods enables high-throughput proteomics experiments. Computational methods for the analysis of proteins and peptides, as well as DNA and RNA [22,23], glycans [24-26], or synthetic polymers [27,28] are also part of computational mass spectrometry, but outside the scope of this review. Finally, disclosing methods is important for reproducible science. Thus, we will also not cover "anecdotal" computational MS where an automated method is mentioned in a paper, but no details of the method are provided.

Review of reviews

Existing reviews on computational MS for small molecules, usually focus on a much more narrow area of the field such as raw data processing [29], metabolomics databases and laboratory information management systems [30], or metabolite identification through reference libraries [31]. Other reviews simply list available tools for processing the data without discussing the individual approaches [32].

A broad overview on experimental as well as theoretical structure elucidation techniques for small molecules using mass spectrometry is given in [33]. Methods specific for qualitative and quantitative metabolomics using LC-MS/MS are covered in [34]. Methods specific for metabolite profiling by GC-MS are covered in [35]. An overview of isotope pattern simulation is given in [36]. Annotation and identification of small molecules from fragmentation spectra using database search as well as *de novo* interpretation techniques is covered in [37].

For a general introduction to metabolomics and metabolomic profiling see [2,3,38]; for recent work in the field see [39].

Experimental setups

Analysis of small molecules by GC-MS is usually performed using Electron Ionization (EI). Historically seen, EI is the oldest ionization technique for small-molecule investigations. Because of the selected constant ionization energy at 70 eV, resulting fragment-rich mass spectra are, in general, consistent across instruments, and specific for each compound. A major disadvantage of mass spectra obtained under EI conditions is the low abundant or missing molecular ion peak; to this end, the mass of the compound is often unknown. GC-MS requires that an analyte is volatile and thermally stable. For non-volatile analytes such as polar compounds, chemical derivatization has to be performed.

Recently, LC-MS has been increasingly used for the analysis of small molecules. Here, compounds are fragmented using tandem MS, for example by Collision Induced Dissociation (CID). This has the advantage that the mass of all molecular ions is known, which is particularly beneficial for *de novo* approaches discussed below. Unfortunately, tandem mass spectra are not as reproducible as EI spectra, in particular across different instruments or even instrument types [40]. Furthermore, using different collision energies can make tandem mass spectra hard to compare. Comparing spectra from different instrument types, only 64–89% of the spectra pairs match with more than 60% identity, depending on the instrument pair [41]. Finally, tandem mass spectra usually contain much less fragments than EI fragmentation spectra. Chemical derivatization can dramatically increase the sensitivity and specificity of LC-MS for less polar compounds [42].

Several methods have been proposed to create more reproducible and informative tandem MS spectra. For example, to increase the number of fragments, tandem MS spectra are often recorded at more than one fragmentation energy. Alternatively, "CID voltage ramping" continuously increases the fragmentation energy during a single acquisition [43]. Also, some progress has been made to normalize fragmentation energies across instruments and instrument types [40,44,45].

Besides the two "standard" experimental setups described above, many other setups have been developed: This includes "alternative" ionization techniques such as Matrix-Assisted Laser Desorption/Ionization [46], Atmospheric Pressure Chemical Ionization [47], Atmospheric Pressure Photoionization [48], and Desorption Electrospray Ionization [49]. Also several chromatographic methods such as High Performance LC [50] and Ultra High Performance LC (UHPLC) [51] have been developed. In particular, a sensitive capillary UHPLC shows good results in lipid identification [52]. Covering the details of these modified setups is far beyond the scope of this review. From the computational side, we can usually classify these modified setups with regards to the two "standard" setups: For example, is the mass of the molecular ion known (LC-MS/MS) or unknown (GC-EI-MS)? Is the fragmentation spectrum rich (GC-EI-MS) or sparse (LC-MS)? What is the mass accuracy of the measurement (see below)? Given that new MS technologies and experimental setups are constantly being developed, we see it as a prerequisite for a "good" method from computational MS that it is not targeted at one particular experimental setup. Note, though, that the

effort required for adapting a method can differ significantly: For example, methods for identifying molecular formulas from isotope patterns (see Section "Molecular formula identification") can be applied to any experimental setup where isotope patterns are recorded. In contrast, rule-based prediction of fragmentation spectra (see Section "In silico fragmentation spectrum prediction") requires expert-curated "learning" of fragmentation rules.

Many methods for the computational analysis of small molecule MS, that go beyond the straightforward library search, require that masses in the mass spectra are measured with an appropriate mass accuracy. It appears that this mass accuracy is much more important for the computational analysis than the often-reported resolving power of MS instruments. Historically, GC-MS is often performed on instruments with relatively bad mass accuracy (worse than 100 ppm, parts per million). In contrast, LC-MS and tandem MS are often performed on instrumental platforms (such as Orbitrap or orthogonal Quadrupole Time-of-Flight MS) that result in a much better mass accuracy, often below 10 ppm or better. This refers to the mass accuracy that we can expect in everyday use of the instrument, not to the "anecdotal mass accuracy" of a single measurement [53]. It must be understood, though, that this is not a fundamental problem of GC-MS; in fact, GC-MS measurements of high mass accuracy are increasingly reported in the literature [54-56].

Reporting standards for metabolomics analysis

For the maturation of metabolomics the lack of standards for presenting and exchanging data needs to be filled. MIAMET (Minimum Information About a METabolomics experiment) [57] suggests reporting standards regarding experimental design, sample preparation, metabolic profiling design and measurements. ArMet [58] is a data model that allows formal description to specify the full experimental context. The Metabolomics Standards Initiative (MSI) [59] develops guidelines and standards for sharing high-quality, structured data following the work of the proteomics community. The Data Analysis Working Group (DAWG) [60] as part of the MSI proposed reporting standards for metabolomics studies that include a reporting vocabulary and will help reproducing these studies and drawing conclusions from the resulting data. The Chemical Analysis Working Group (CAWG) established confidence levels for the identification of non-novel chemical compounds [61], ranging from level 1 for a rigorous identification based on independent measurements of authentic standards, to unidentified signals at level 4. The NIH Metabolomics Fund recently supported an initiative to create a repository that enforces the submission of metadata.

Data storage and spectral libraries

To allow data-driven development of algorithms for small molecule identification, mass spectrometric reference datasets must be made *publicly available* via reference databases. Examples of such databases include MassBank [62,63], METLIN [64,65], Madison Metabolomics Consortium Database (MMCD) [1], Golm Metabolome Database (GMD) [66], the Platform for RIKEN Metabolomics (PRiMe) [67], or MeltDB [68]. Unfortunately, making available experimental data is much less pronounced in the metabolomics and small-molecule research community, than it is in proteomics or genomics. For example, several of the above-mentioned databases do not allow for the batch download of the database. Citing [69], "to make full use of research data, the bioscience community needs to adopt technologies and reward mechanisms that support interoperability and promote the growth of an open 'data commoning' culture." Possibly, the MetaboLights database that is part of the ISA (Investigation, Study, Assay) commons framework can fill this gap. Note that the PubChem database allows free access to more than 35 million molecular structures, and this includes batch download of the data.

Besides the open (or partly open) libraries mentioned above, there exist two important commercial libraries: The National Institute of Standards and Technology (NIST) mass spectral library (version 11) contains EI spectra of more than 200 000 compounds; the Wiley Registry (9th edition) contains EI spectra of almost 600 000 unique compounds. For comparison, the GMD [66] contains EI fragmentation mass spectra of about 1 600 compounds; and the FiehnLib library contains EI spectra for more than 1 000 metabolites [70].

The size of tandem MS libraries is still small, compared to EI libraries (see Figure 1). The NIST 11 contains collision cell spectra for about 4 000 compounds. The Wiley Registry of Tandem Mass Spectral Data [71,72] comprises positive and negative mode spectra of more than 1 200 compounds. As for EI spectra, both databases are commercially available.

As even the commercial libraries are small, there have been several attempts to make tandem mass spectra publicly available. METLIN [64] contains high resolution tandem mass spectra for more than 10 000 metabolites for diagnostics and pharmaceutical biomarker discovery and allows to build a personalized metabolite database from its content [73]. MassBank [62,63] is a public repository with more than 30 000 spectra of about 4 000 compounds collected from different consortium members. The MMCD [1] is a hub for NMR and MS spectral data containing about 2 000 mass spectra from the literature collected under defined conditions. Some databases address specific research interests. The Human Metabolome DB [74,75] comprises reference MS-MS spectra for more than

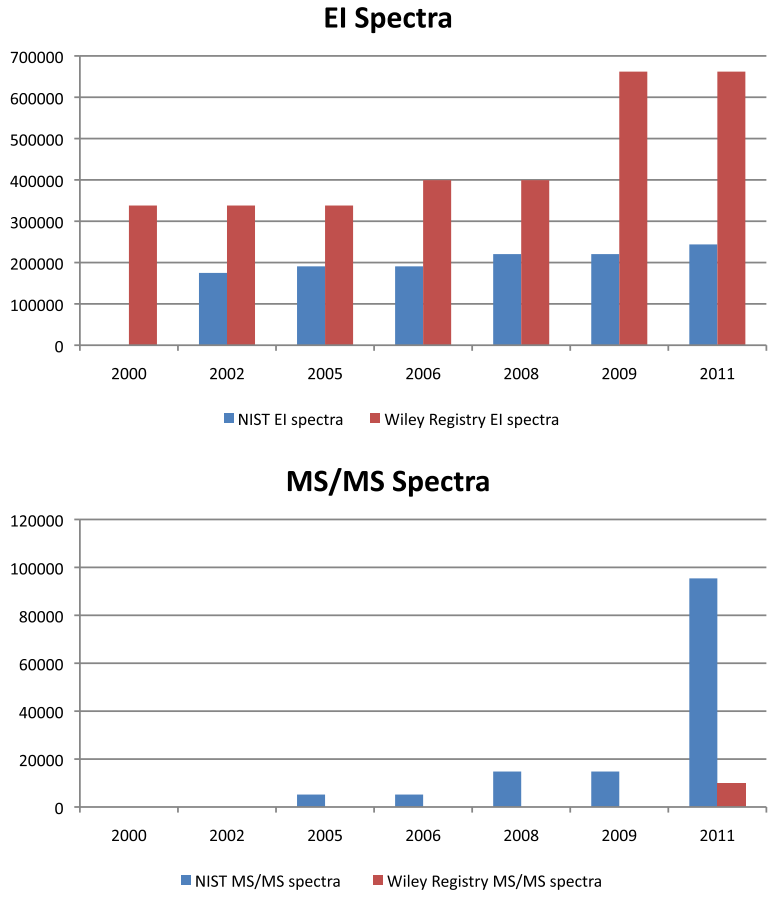

Figure 1 Number of EI spectra (top) and tandem mass spectra (bottom) in NIST and Wiley Registry from 2000 until 2011.

2 500 metabolites found in the human body. The Platform for RIKEN Metabolomics (PriMe) [67,76] collects MS^n spectra for research on plant metabolomics.

Searching spectral libraries

The usual approach for identification of a metabolite is looking it up in a spectral library. Database search requires a similarity or distance function for spectrum matching. The most fundamental scorings are the "peak count" family of measures that basically count the number of matching peaks. A slightly more complex variant is taking the dot product of the two spectra, taking into account peak intensities.

Establishing the confidence is the more difficult part of compound identification using library search [31]. False negative identifications occur if the spectrum of the query compound differs from the spectrum in the library, for example due to contaminations, noise (especially in low signal spectra), or different collision energies (CID). A reliable identification of a compound depends on the uniqueness of its spectrum, but the presence and intensity of peaks across spectra is highly correlated, as these depend on the non-random distribution of molecular (sub-)structures. Therefore, structurally related compounds generally have similar mass spectra. Hence, false positive hits may hint at correct "class identifications", see Section "Searching for similar compounds" below. Different from proteomics, False Discovery Rates (FDR) cannot be estimated as no appropriate decoy databases can be constructed. Usually, confidence in search results must be manually assessed by the user, based on the used search algorithm and the quality of spectrum and library [77]. Another method that overcomes this limitation is the calculation of fragmentation trees from fragmentation spectra, see Section "Fragmentation trees" below. For a review on using spectral libraries for compound identification, see [31].

Electron ionization fragmentation spectra

To compare EI mass spectra, a huge number of scorings (or similarity measures) have been developed over the years. In 1971, the Hertz similarity index was introduced [78], representing the weighted average ratio of the two spectra. The Probability Based Matching (PBM) [79,80]

takes into account that some peaks are more informative than others. Atwater *et al.* [81] statistically evaluated the effects of several parameters on the PBM system, to provide a quantitative measure of the predicted reliability of the match. *SISCOM* [82] encodes spectra by selecting the most informative peaks within homologous ion series. Computing the dot product cosine of two mass spectra (that is, the inverse cosine of the dot product of the normalized spectra) was used in the *INCOS* data system [83]. Stein and Scott [84] evaluated normalized Euclidean distances [85], PBM, Hertz similarity index, and dot product for searching EI databases. Among these, they found the dot product to perform best. They proposed a composite search algorithm that optimizes the cosine score by varying the scaling and mass weighting of the peak intensities. Koo *et al.* [86] introduced novel composite similarity measures that integrate wavelet and Fourier transform coefficients, but found only a slight improvement over cosine correlation or the composite similarity measure. Kim *et al.* [87] showed how to find optimal weight factors for fragment masses using a reference library.

Regarding the differentiation between true and bogus hits in the database, not much progress has been made: Probabilistic indicators of correct identifications using "match factors" were introduced in [88]. Jeong *et al.* [89] used an empirical Bayes model to improve the accuracy of identifications and gave a false positive estimate. For this purpose, a competition score was added to the similarity score, based on the similarity score to other spectra in the library.

Tandem mass spectra

We noted above that LC-MS/MS is much less reproducible than fragmentation by GC-MS (see Figure 2). Reliable library identifications can be achieved when a spectrum is acquired under the same conditions as the reference spectrum [90]. For each compound, libraries must contain tandem mass spectra at different collision energies and replicates on different instruments, to allow for an effective identification [91]. For example, Oberacher and coworkers [71,72,92] presented an inter-instrument and inter-laboratory tandem mass spectral reference library obtained using multiple fragmentation energy settings.

For searching in tandem mass spectral libraries it is possible to start with a precursor ion mass filtering with a specific m/z or mDa range. In case the actual compound is not in the database, it can be beneficial to omit this filtering step. This may reveal valuable information about structurally similar compounds [92]. Subsequently, similar approaches as for EI mass spectra can been applied, such as PBM [79,80] or dot product cosine [84,93]. Again, intensities can be weighted using peak masses [62,63]. The scoring in [92] extends the common peak count. Zhou

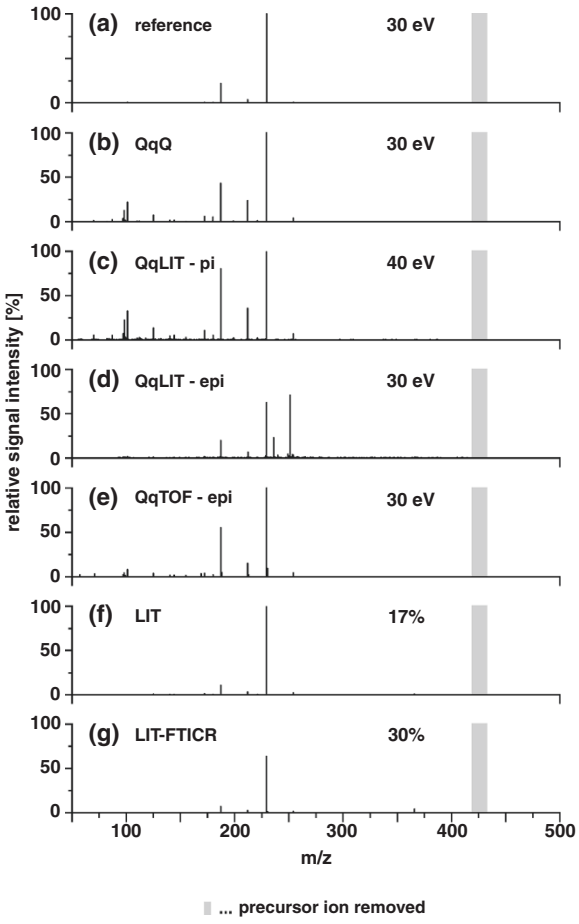

Figure 2 Inter-instrument comparability of dixyrazine-specific tandem mass spectra collected on different instrumental platforms. Figure provided by Herbert Oberacher, compare to Figure one in Oberacher *et al* [71].

et al [94] proposed a support vector machine (SVM)-based spectral matching algorithm to combine multiple similarity measures. Hansen and Smedsgaard [95] used the Jeffrey-Matusitas distance [96] to find a unique correspondence between the peaks in the two spectra.

X-Rank replaces peak intensities by their rank, then estimates the probability that a peak in the query spectrum matches a peak in the reference spectrum based on these ranks [97]. Oberacher *et al* [71,72] tackled the problem of low reproducibility of metabolite CID fragmentation using a dynamic intensity cut-off, counting neutral losses, and optimizing the scoring formula. To improve running times, the database can be filtered using the most intense peaks and user-defined constraints [98].

Molecular formula identification

One of the most basic — but nevertheless highly important — steps when analyzing an unknown compound, is to determine its molecular formula, often referred to as

the "elemental composition" of the compound. Common approaches first compute candidate molecular formulas using a set of potential elements. The six elements most abundant in metabolites are carbon (C), hydrogen (H), nitrogen (N), oxygen (O), phosphorus (P), and sulfur (S) [99]. For each candidate molecular formula, an isotope pattern is simulated and compared to the measured one, to determine the best matching molecular formula. For this purpose, high mass accuracy is required and is nowadays available from a multitude of MS platforms. The molecular formula of the compound can serve as a basis for subsequent structure elucidation. Some software packages for molecular formula identification using isotope patterns are summarized in Table 1.

Table 1 Software for the three basic steps of molecular formula identification using isotope patterns

Decomposing monoisotopic peaks	
Decomp [100,101]	for arbitrary alphabets of elements
	requires only little memory
	swift in practice
SIRIUS [102,103]*	implementing *Decomp* approach for MS
	decomposing real-valued masses
"Seven Golden Rules" [104]	to filter molecular formulas
Simulating isotope patterns	
IsoPro [105]	multinomial expansion to predict "center masses"
	memory- and time-consuming
Mercury [106]	pruning by probability thresholds and/or mass range
	reduced memory and time consumption
	reduced accuracy of the predictions
Emass [107]* & *SIRIUS* [102]*	iterative (stepwise) computation of isotope pattern
	probability-weighted center masses
	probabilities and masses are updated as atoms are added
IsoDalton [108]	models the folding procedure as a Markov process
BRAIN [109]*	Newton-Girard theorem and Vietes formulae to calculate intensities and masses
Fourier [110]*	2D Fast Fourier Transform that splits up the calculation in a coarse and a fine structure
	running time improvement for large compounds
Scoring candidate compounds	
SigmaFit	commercial software by Bruker Daltonics
SIRIUS [102]*	Bayesian statistics for scoring intensities and masses of the isotope pattern
MZmine [111]	simple scoring based only on intensities

*Recommended tools.

Different from the above, some authors propose to use molecular structure databases to determine the candidate molecular formulas [112]. This "simplifies" the problem as the search space is severely restricted; but only those molecular formulas can be determined where a compound is available in the structure database. To this end, we will ignore this somewhat arbitrary restriction of the search space.

In the following, we assume that elements are unlabeled or only partially labeled. If certain elements are (almost) completely labeled by heavy isotopes such as ^{13}C, and both the unlabeled and the labeled compound are present, this allows us to directly "read" the number of atoms from the spectrum using the mass difference. We will come back to this particular type of data in Section "Isotope labeling".

Decomposing monoisotopic peaks

Here, "decomposing a peak" refers to finding all molecular formulas (over the fixed alphabet of elements) that are sufficiently close to the measured peak mass. Robertson and Hamming [113] and Dromey and Foyster [114] proposed a naïve search tree algorithm for this purpose. One can show that the running time of this algorithm linearly depends on m^{k-1} where m is the mass of the peak we want to decompose, and k is the number of elements [102]. This means that doubling the peak mass we want to decompose, will increase the running time of the algorithm 32-fold for the alphabet of elements CHNOPS. Hence, running time can easily get prohibitive, in particular if we consider larger alphabets of elements, or have to perform many decompositions. In 1989, Fürst *et al* [115] proposed a faster decomposition algorithm which, unfortunately, is limited to the four elements CHNO. In 2005, Böcker and Lipták [100,101] presented an algorithm that works for arbitrary alphabets of elements, requires only little memory, and is swift in practice. Initially developed for decomposing integer masses, this algorithm was later adapted to real-valued masses [102,103,116].

Decomposing alone is not sufficient to exclude enough possible molecular formulas in higher mass regions even with very high mass accuracy [117]. Kind and Fiehn [104] proposed "Seven Golden Rules" to filter molecular formulas based on chemical considerations. However, for larger masses, many molecular formulas pass these rules.

As the monoisotopic mass of a compound is insufficient to determine its molecular formula, we can use the measured isotope pattern of the compound to rank all remaining molecular formula candidates. Kind and Fiehn [117] estimated that mass spectrometers capable of 3 ppm accuracy and 2% error for isotopic abundances, can outperform mass spectrometers with hypothetical mass accuracy of 0.1 ppm that do not include isotopic

information. To this end, we now consider the problems of simulating and matching isotope patterns.

Simulating isotope patterns

Due to limited resolution of most MS instruments the isotopic variants are not fully separated in the spectra but pooled in mass bins of approximately 1 Da length. This is called the aggregated isotopic distribution [36] and in the following we will refer to it as "isotope pattern".

Most elements have several naturally occurring isotopes. Combining elements into a molecular formula also means to combine their isotope distributions into an isotope distribution of the entire compound. Masses of all isotopes are known with very high precision [118,119]. This is, to a much lesser extend and with certain exceptions, also true for the natural abundances of these isotopes on earth [120]. (For example, the abundances of boron isotopes vary strongly.) To this end, we can simulate the theoretical isotope pattern of a molecular formula, and compare the simulated distribution to the measured pattern of a compound. See Valkenborg *et al* [36] for an introduction.

The intensity of a peak in an isotope pattern is the superposition of all isotope variants' abundances that have identical nominal mass (nucleon number) [36]. In the early 1960's, mass accuracy of MS instruments was relatively low. Thus, first approaches for simulating isotope patterns ignored the exact mass of the isotope peaks, and concentrate solely on isotope peak intensities, that is, the isotope distribution [121]. In 1991, Kubinyi [122] suggested a very efficient algorithm for this problem, based on convoluting isotope distributions of "hyperatoms".

As instruments with improved mass accuracy became commercially available, focus shifted towards also predicting masses of isotope peaks, named "center masses" by Roussis and Proulx [123]. For this purpose, methods based on polynomial [124] and multinomial expansion [105,125] were developed. *IsoPro* is an implementation of [105] by M.W. Senko. Unfortunately, these expansion approaches are very memory- and time-consuming. Pruning by probability thresholds or mass range or both was introduced to reduce memory and time consumption; but this comes at the price of reduced accuracy of the predictions [106,126-128]. The approach of [106] was implemented in the software package *Mercury*.

Starting in 2004, methods that use an iterative (stepwise) computation of isotope pattern were developed [107,116,123]. These algorithms are similar in spirit to the early algorithms for computing peak intensities [121,122]. But for the new algorithms, probabilities *and masses* of isotope peaks are updated as atoms are added. This results in probability-weighted center masses. Two implementations are *Emass* [107] and *SIRIUS* [102]. To speed up computations, both approaches combine this with a smart Russian multiplication scheme, similar to Kubinyi [122].

Later approaches model the folding procedure as a Markov process [108,129,130]. *IsoDalton* implements the approach of Snider [108]. All approaches have in common that a truncation mechanism must be applied due to the exponential growth of states.

In 2012, Claesen *et al* [109] applied the Newton-Girard theorem and Vietes formulae to calculate the intensities and masses of an isotope pattern. This method is implemented in the software tool *BRAIN*. They compared their method against five other software tools: *IsoPro*, *Mercury*, *Emass*, *NeutronCluster* [131], and *IsoDalton*. In this evaluation, *BRAIN* outperformed all other software tools but *Emass* in mass accuracy of the isotope peaks. Running times were comparable for *BRAIN*, *Emass*, *Mercury*, and *NeutronCluster*, whereas *IsoPro* and *IsoDalton* required much higher computation times. Later, Böcker [132] showed that *SIRIUS* and *BRAIN* have practically identical quality of results and running times for simulating isotope patterns.

The currently fastest algorithm was presented by Fernandez-de-Cossio Diaz and Fernandez-de-Cossio [110]. This algorithm improves on earlier work were a 2D Fast Fourier Transform is applied that splits up the calculation in a coarse and a fine structure [133]. *Fourier* [110] shows a significantly better performance than *BRAIN* and, hence, *Emass* and *SIRIUS*. It must be noted, though, that this running time improvement is only relevant for large compounds: The smallest compound considered in [109,110,132] has mass above 1000 Da, and significant running time differences for *Fourier* are observed only for compounds with mass above 10 kDa. For compounds of mass above, say, 50 kDa the problem of simulating isotope patterns becomes somewhat meaningless: The abundances of isotope species are known with limited precision, and vary depending on where a sample is taken. These small deviations in the isotopic distribution of elements cause huge deviations in the aggregated distribution, if the compound is sufficiently large [134].

For the efficient and accurate simulation of isotope patterns of small compound, it is recommended to use one of the approaches behind *Fourier* [110], *BRAIN* [109], *Emass* [107], or *SIRIUS* [102].

Scoring candidate compounds by comparing isotope patterns

Decomposing the monoisotopic peak can result in a large number of candidate molecular formulas that are within the measured mass [117]. We can rank these candidates based on evaluating their simulated isotope patterns. For each candidate molecular formula, the isotope distribution is simulated and compared with the measured one.

The best matching formula is considered to be the correct molecular formula of the compound. See Figure 3.

Initially, mass spectrometers were limited in mass accuracy and resolution. To this end, first attempts of scoring isotope patterns only considered the intensity of the isotopic peaks but not their masses. Kind and Fiehn [117] calculated a root mean square error for the differences between measured and theoretical isotopic intensities. Stoll *et al* [135] filtered candidates using double-bond equivalents and number of valences, then rank candidates based on correlating the isotope distributions [136]. Commercial software for the same purpose was also provided by instrument vendors, such as *SigmaFit* by Bruker Daltonics. *Tal-Aviv* [137] targets GC-MS EI data using a supersonic molecular beam, which results in highly abundant molecular ions.

Böcker *et al* [102] introduced *SIRIUS*, first suggested in [116]. Here, both the intensities and masses of the isotope pattern are used to score candidate molecular formulas using Bayesian statistics: The authors estimate the likelihood of a particular molecular formula to produce the observed data. For a dataset of 86 compounds measured on an oa-TOF MS instrument, the correct formula was identified in more than 91% of the cases. Ipsen *et al* [138] developed a method to determine confidence regions for isotope patterns, tailored towards TOF MS data. They employ that the rate of ion arrivals at the detector plate is governed by the Poisson distribution. A test on three

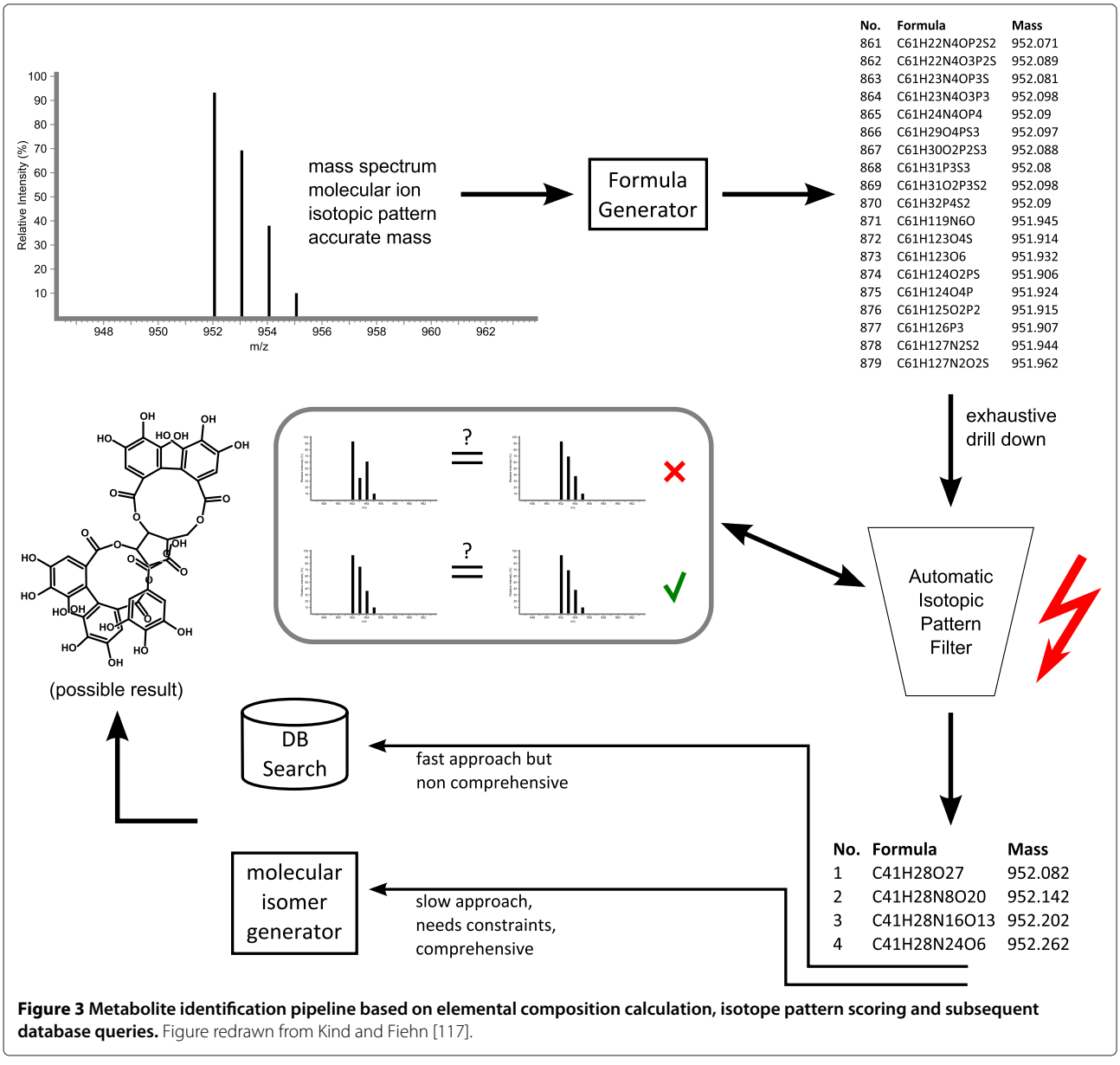

Figure 3 Metabolite identification pipeline based on elemental composition calculation, isotope pattern scoring and subsequent database queries. Figure redrawn from Kind and Fiehn [117].

compounds showed that the method rejects about 70% of the candidate formulas (for pooled data) but keeps the true formula, at the 5% significance level.

Isotope labeling

Labeling compounds by isotope-enriched elements such as ^{13}C or ^{15}N, helps to identify the correct molecular formula. The shift in the mass spectrum between the unlabeled compound and the labeled compound indicates the number of atoms in the compounds. Once the number of atoms for the labeled elements is known, the number of possible molecular formula is significantly reduced. Rodgers *et al* [139] showed that enrichment with 99% ^{13}C isotopes reduces the number of possible molecular formulas for a 851 Da phospholipid from 394 to one. Hegeman *et al* [140] used isotopic labeling for metabolite identification. They improved the discriminating power by labeling with ^{13}C and ^{15}N isotopes. Giavalisco *et al* [141] additionally labeled compounds with ^{34}S isotopes. By this, the number of carbon, nitrogen as well as sulfur atoms can be determined upfront, and the number of potential molecular formula that we have to consider, is reduced considerably. *Baran* et al [142] applied this approach to untargeted metabolite profiling and showed its potential to uniquely identify molecular formulas.

Other approaches for molecular formula identification

Tandem or multiple-stage MS can give additional information about the molecular formula of the intact compound: We can exclude all molecular formulas of the compound if, for one of the fragment (product ion) peaks, we cannot find a sub-formula that explains this peak [143-146]. Unfortunately, such approaches are susceptible to noisy data. To this end, Konishi and coworkers [143,144] suggested to use only product ions below a certain threshold, e.g., 200 Da, that have a unique decomposition.

Pluskal *et al* [111] combined matching isotope patterns with filtering based on the molecular formulas of product ions. For 79% of the 48 compounds considered, they identified the correct molecular formula. There exist commercial tools that follow the same line of thought: For example, *SmartFormula3D* [146] (commercial, Bruker Daltonics) appears to implement a similar approach. Pluskal *et al* [111] also evaluated their new, simple scoring of isotope patterns against *SIRIUS* [102], and reported that it performs better.

A generalization of this concept are fragmentation trees which were initially introduced to compute molecular formulas [147]. For each potential molecular formula of the intact compound, a fragmentation tree and its score are computed. Potential molecular formulas of the compound are then sorted with respect to this score. Rasche *et al* [148] combined this with isotope pattern analysis [102], and for the 79 considered compounds measured on two instruments, they could identify the correct molecular formula in all cases. For more details on fragmentation trees, see Section "Fragmentation trees" below.

All of the above approaches assume that only the monoisotopic peak is selected for dissociation. Selecting a non-monoisotopic peak can reveal valuable information about the molecular formulas of the product ions. Singleton *et al* [149] developed an approach to predict the expected isotope pattern for tandem mass spectra for precursor ions that contain only one element with one heavy isotope. Rockwood *et al* [150] generalized this and developed an algorithm that can be applied to arbitrary precursor ions. It is based on the convolution of isotope distributions of the product ion and the loss. Again, comparing theoretical and experimental isotope patterns shed light on the correct product ion formula. Ramaley and Herrera [151] modified the algorithm from [149] to apply it to arbitrary precursor ions; results are comparable to [150].

Rogers *et al* [152] used the information of potential metabolic pathways to identify the correct molecular formula. If there is a putative chemical transformation between two molecular formulas, these formulas get a better score than other explanations of the peak. This does not only improve molecular formula identification, but can potentially be used to reconstruct biochemical networks. See Section "Network reconstruction" for details.

Identifying the unknowns

To yield information beyond the compound mass and molecular formula, the analyte is usually fragmented, and fragmentation mass spectra are recorded. Using spectral comparison one can identify huge numbers of metabolites that are cataloged in libraries. However, where the compound is unknown, comparing the spectrum obtained to a spectral library will result in imprecise or incorrect hits, or no hits at all [33,35,99]. The limited capability for metabolite identification has been named one of the major difficulties in metabolomics [117]. Manual analysis of unidentified spectra is cumbersome and requires expert knowledge. Therefore, automated methods to deal with mass spectra of *unknown unknowns* (that is, "unexpected" compounds that are not present in spectral libraries [31]) are required. Some approaches for analyzing fragmentation mass spectra of *unknown unknowns* are summarized in Table 2.

Searching for similar compounds

In case a database does not contain the sample compound an obvious approach is to search for similar spectra, assuming that spectral similarity is based on structural similarity of the compounds. Back in 1978, Damen *et al* [82], already suggested that *SISCOM* can also be

Table 2 Approaches for analyzing fragmentation mass spectra of *unknown unknowns* that is, "unexpected" compounds that are not present in spectral libraries [31]

		In silico fragmentation		
Searching for similar compounds	**Mass spectral classifiers**	**Rule-based spectrum prediction**	**Combinatorial fragmentation**	**Fragmentation trees**
searching for similar spectra in a library, assuming that spectral similarity is based on structural similarity	predicting substructures or compound classes by learning spectral classifiers	predicting spectra by applying fragmentation rules to known molecular structures	mapping the fragmentation spectrum to the compound structure to explain the peaks	computing a fragmentation tree that explains the peaks; aligning fragmentation trees to find similar compounds
NIST MS Interpreter [153]	*FingerID* [169]	*Mass Frontier*, *ACD/MS Fragmenter*, *MOLGEN-MS* [196]	*MetFrag* [179]	*SIRIUS* [147,221]

used to detect structural similarities such as common substructures.

The NIST *MS Interpreter* [153] for EI spectra uses a nearest-neighbor approach to generate substructure information. A library search provides a list of similar spectra. Structural features of the unknown compound, such as aromatic rings or carbonyl groups, are deduced from common structural features of the hits. Demuth *et al* [154] proposed a similar approach, and evaluated whether spectral similarity is correlated with structural similarity of a compound. Based on this evaluation, they proposed a threshold for spectral similarity that supposedly yields hit lists with significantly similar structures. For multiple MS data, Sheldon *et al* [155] used precursor ion fingerprints (PIF) and spectral trees for finding similar compounds and utilized previously characterized ion structures for the structural elucidation of the unknown compounds.

Mass spectral classifiers

Another natural approach to deal with mass spectra of compounds that cannot be found in a spectral library, is to find patterns in the fragmentation spectra of reference compounds, and to use the detected patterns for the automated interpretation of the unidentified spectrum. Initially, this was accompanied by knowledge about the fragmentation processes; but this applies only for fragmentation by EI, whereas fragmentation by CID is less reproducible and not completely understood [156].

To characterize an unknown compound, we have to come up with "classifiers" that assign the unknown to a certain class: such classes can be based on the presence or absence of certain substructures, or more general structural properties of the compound. As EI fragmentation is already well understood, many mass spectral classifiers have been provided to date. Already in 1969, Venkataraghavan *et al* [157] presented an automated approach "to identify the general nature of the compound and its functional groups." The Self-Training Interpretive and Retrieval System (*STIRS*) [158] mixes a rule-based approach with some early machine learning techniques to obtain structural information from related EI spectra. Further, *STIRS* can predict the nominal molecular mass of an unknown compound, even if the molecular ion peak is missing from the EI spectrum. Scott and coworkers [159-161] proposed an improved method for estimating the nominal molecular mass of a compound. Using pattern recognition the compound is classified, and class-specific rules are applied to estimate the molecular mass.

Structural descriptors (that is, fragments of a certain integral mass) have been used to retrieve compound classes for many decades [162]. The Varmuza feature-based classification approach for EI spectra [163] uses a set of mass spectral classifiers to recognize the presence/absence of 70 substructures and structural properties in the compound. This approach is integrated to *MOLGEN-MS* and *AMDIS*. For example, Schymanski *et al* [164] combined mass spectral classifiers with methods for structure generation (see Section "Molecular isomer generators") to interpret EI spectra classifiers from *MOLGEN-MS* and the NIST05 software. Further MS classifiers for substructures are provided in [165,166]. Hummel *et al* [167] used structural features to subdivide the Golm Metabolome Database into several classes. They proposed a decision tree-based prediction of the most frequent substructures, based on mass spectral features and retention index information, for classification of unknown metabolites into different compound classes. In 2011, Tsugawa *et al* [168] used Soft Independent Modeling of Class Analogy (SIMCA) to build multiple class models. However, back in 1996, Varmuza and Werther [163] observed that SIMCA (which is based on the Principle Component Analysis) performed worst among all investigated methods.

Whereas all of the above methods are targeted towards GC-MS and EI fragmentation, few methods target LC-MS and CID fragmentation. A novel approach by Heinonen *et al* [169] predicts molecular properties of the unknown metabolite from the mass spectrum using a support vector machine, then uses these predicted properties for matching against molecular structure databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and PubChem (see Figure 4). To this end, we can replace the small

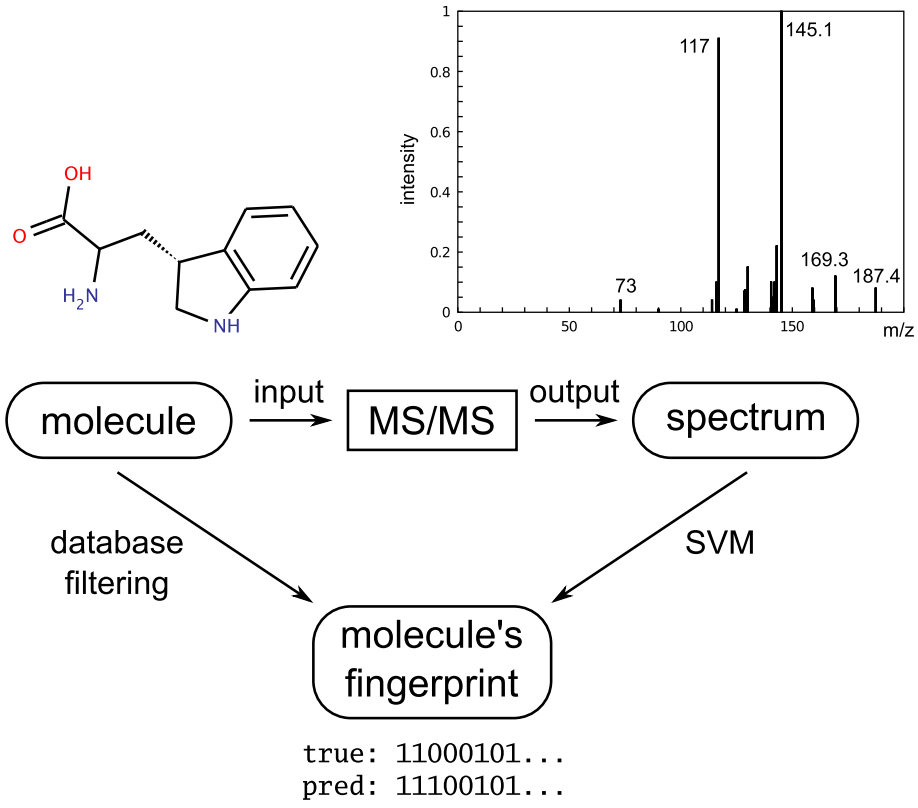

Figure 4 Predicting chemical properties (molecule fingerprints) from tandem MS data using a support vector machine (SVM) as done by Heinonen *et al* [169]. The predicted fingerprints are used to search a molecular structure database for metabolite identification. Figure redrawn from Heinonen *et al* [169].

spectra libraries by the much larger structure databases. Using QqQ MS data and searching the smaller KEGG database, they could identify the correct molecular structure in about 65% of the cases, from an average of 25 candidates.

Molecular isomer generators

Molecular isomer generators such as *MOLGEN* [170-172], *SMOG* [173], and *Assemble* [174] have helped with the structural elucidation of unknowns for many years [175,176]. Recently, the open source software OMG was introduced [177]. Molecular isomer generators enumerate all molecular structures that are chemically sound, for a given molecular formula or mass. In addition, the space of generated structures can be constrained by the presence or absence of certain substructures, see Section "Mass spectral classifiers". An overview on generating structural formulas is given by Kerber *et al* [172]. Enumerating all possible isomers allows us to overcome the boundaries of database searching: Simply generate all molecular structures corresponding to the parent mass or molecular formula, and use the output of the structure generator as a "private database". Unfortunately, this approach is only valid for relatively small compounds (say, up to 100 Da): For molecular formula *C8H6N2O* with mass 146 Da there exist 109 240 025 different molecular structures [172].

In silico fragmentation spectrum prediction

In silico fragmentation aims to explain "what you see" in a fragmentation spectrum of a metabolite. Initially, this was targeted at a manual interpretation of fragmentation spectra; but recently, this approach has been increasingly used for an automated analysis [178,179]. Here, searching in spectral libraries is replaced by searching in molecular structure databases. We mentioned above that spectral libraries are (and will be) several orders of magnitude smaller than molecular structure databases: For example, the CAS Registry of the American Chemical Society and PubChem currently contain about 25 million compounds each. We can also use molecular structure generators (see "Molecular isomer generators") to create a "private database". However, whereas structure generators can enumerate millions of structures in a matter of seconds, it is already a hard problem to rank the tens or hundreds of molecular structures found in molecular structure databases for a particular parent mass [178,179].

In silico fragmentation has been successfully applied to compounds with consistent fragmentation pattern, such

as lipids [180], oligosaccharides [181], glycans [182], peptides [183-185] or non-cyclic alkanes and alkenes [186]. However, general fragmentation prediction of arbitrary small molecule remains an active field of research, due to the structural diversity of metabolites and the complexity of their fragmentation patterns.

Basically there are two types of *in silico* fragmentation methods. *Rule-based fragmenters* are based on fragmentation rules that were extracted from the MS literature over the years. *Combinatorial fragmenters* use a bond disconnection approach to dissect a compound into hypothetical fragments.

Rule-based fragmenters

Although much is known about EI fragmentation, it is a hard ionization technique that can result in very complex rearrangements and fragmentation events [187] which are hard to predict. For tandem MS, the fragmentation behavior of small molecules under varying fragmentation energies is not completely understood [156], and has been investigated in many studies to find general fragmentation rules [188,189]. *Mass Frontier* (see below) currently contains the largest fragmentation library, manually curated from several thousand publications [33].

The first rule-based approaches for predicting fragmentation patterns and explaining experimental mass spectra with the help of a molecular structure were developed as part of the *DENDRAL* project. For example, Gray *et al* [190] introduced *CONGEN* that predicts mass spectra of given molecular structures using general models of fragmentation, as well as class-specific fragmentation rules. Intensities for EI spectra were modeled with equations found by multiple linear regression analysis of experimental spectra and molecular descriptors [191].

Gasteiger et al [9] introduced *MASSIMO* (MAss Spectra SIMulatOr) to automatically derive knowledge about mass spectral reaction types directly from experimental mass spectra. Part of *MASSIMO* is the Fragmentation and Rearrangement ANalyZer (*FRANZ*) that requires a set of structure-spectrum-pairs as input. The MAss Spectrum SImulation System (*MASSIS*) [192-194] combines cleavage knowledge (McLafferty rearrangement, retro-Diels-Alder reaction, neutral losses, oxygen migration), functional groups, small fragments (end-point and pseudo end-point fragments) and fragment-intensity relationships for simulating electron ionization spectra. Unfortunately, these three software packages were neither sufficiently validated nor made publicly available. As a consequence, they were never used or applied by the broad community and should be considered with caution.

Mass Frontier (HighChem, Ltd. Bratislava, Slovakia; versions after 5.0 available from Thermo Scientific, Waltham, USA) contains fragmentation reactions collected from mass spectrometry literature. Besides predicting a spectrum from a molecular structure, it can also explain a measured fragmentation spectrum. The *ACD/MS Fragmenter* (Advanced Chemistry Labs, Toronto, Canada) can only interpret a given fragmentation spectrum using a known molecular structure [195]. Initially, these programs were designed for the prediction and interpretation of fragmentation by EI, but recently, there has been a tendency to interpret tandem MS data with theses programs, too. Both programs are commercial, and no algorithmic details have been published. A third commercial tool is *MOLGEN-MS* [196,197] that uses general mass spectral fragmentation rules but can also accept additional fragmentation mechanisms.

For the interpretation of tandem mass spectra, Hill *et al* [178] proposed a "rule-based identification pipeline". First, they retrieved candidate molecular structures from PubChem using exact mass. Next, *Mass Frontier* 4 was used to predict the tandem mass spectra of the candidates, which were matched to the measured spectrum, counting the number of common peaks. In this way, a rule-based fragmenter can be used to search in a molecular structure database. Pelander *et al* [198] used *ACD/MS Fragmenter* for drug metabolite screening by tandem MS. For the simulation of EI fragmentation spectra, Schymanski *et al* [195] compared the three commercial programs, and indicated that at the time of evaluation, mass spectral fragment prediction for structure elucidation was still far from daily practical usability. The authors also noted that *ACD Fragmenter* "should be used with caution to assess proposed structures [...] as the ranking results are very close to that of a random number generator." Later, Kumari *et al* [199] implemented a pipeline for EI spectra integrating *Mass Frontier* that is similar to the one for tandem MS data [178], but integrates retention time prediction. They retrieved candidate structures from PubChem using molecular formulas predicted from the isotope pattern [104]. They filtered molecular structures using Kováts retention index prediction [15]. Using *Mass Frontier* 6 for spectrum prediction, the correct structure was reported in 73% within the TOP 5 hits.

It is worth mentioning that rule-based systems did not have much success in proteomics: There, it is apparent from the very beginning that, in view of the huge search space, only optimization- and combinatorics-based methods can be successful.

Combinatorial Fragmenters

The problem with rule-based fragmenters is that even the best commercial systems cover only a tiny part of the rules that should be known. Constantly, new rules are discovered that have to be added to the fragmentation rule databases. However, all of these rules do not necessarily apply to a newly discovered compound.

Sweeney [200] observed that many compounds can be described in a modular format, that is, substructures which account for most of the fragments observed in the fragmentation spectrum (see Figure 5). Combinatorial fragmenters use bond disconnection to explain the peaks in the observed fragmentation spectrum. Fragments resulting from structural rearrangements are initially not covered by this approach. Usually, such rearrangements have to be individually "woven" into the combinatorial optimization; this is often complicated and done only for a few, particularly important rearrangements. Note that handling rearrangement reactions is problematic for both combinatorial and rule-based methods [200-202].

EPIC (elucidation of product ion connectivity) [201] was the first software using systematic bond disconnection and ranking of the resulting substructures. It was tested only against two hand annotated spectra from the literature and is not publicly available. The Fragment iDentificator (*FiD*) [202,203] enumerates all possible fragment candidates using a Mixed Integer Linear Programming approach, and ranks the candidates according the cost of cleaving a fragment. Due to the computational complexity of the underlying problem [204], running times can be prohibitive even for medium-size compounds.

The most recent approach is *MetFrag* [179], a somewhat greedy heuristic to match molecular structures to measured spectra that makes no attempt to create a mechanistically correct prediction of the fragmentation processes. It is therefore fast enough to screen dozens to thousands of candidates retrieved from compound databases, and to subsequently rank them by the agreement between measured and *in silico* fragments (see Figure 6). Hill *et al* On the same test set that was used by [178], *MetFrag* performed better than the commercial *Mass Frontier* 4. *MetFrag* predictions were included in the recent METLIN database release [65]. *MetFrag* has also been extended to analyze EI fragmentation [205]. Recently, Gerlich and Neumann [206] introduced *MetFusion* that combines the *MetFrag* approach with a similarity fingerprint to re-rank the molecular structures.

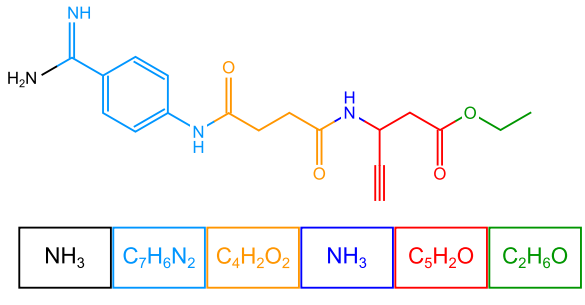

Figure 5 Modular structure of xemilofiban. Figure redrawn from Sweeney [200].

Other experimental measures such as retention indices or drift time, can be used for candidate filtering [205,207]. *Ridder et al* [208] presented a closely related approach for substructure prediction using multistage MS data.

One problem of combinatorial fragmenters is how to choose the costs for cleaving edges (bonds) in the molecular structure graph. For this, *MetFrag* uses bond dissociation energies whereas "unit weights" are used in [208]. Kangas *et al* [180] used machine learning to find bond cleavage rates. Their *In silico* identification software (*ISIS*) currently works only for lipids and is not modeling rearrangements of atoms and bonds. Different from the other approaches, *ISIS* simulates the spectrum of a given lipid, and does not require experimental data to do so.

Consensus structure approaches

Many of the above mentioned techniques are rather complementary yielding different information on the unknown compound. Combining the different results will therefore greatly improve the identification rates. For EI fragmentation data, [205] used a consensus scoring to selected candidates. These structural candidates are generated using molecular formula and substructure information retrieved from *MOLGEN-MS* and *MetFrag*, and further characteristics (e.g., retention behavior). Ludwig *et al* [209] proposed a greedy heuristic to find the characteristic substructure that is "embodied" in a list of database search results; see also Section "Fragmentation trees".

Nonribosomal peptides

Usually the structure of small molecules cannot be deduced from the genomic sequence. However, for particular molecules such as nonribosomal peptides (NRPs) a certain predictability has been established [210]. NRPs are excellent lead compounds for the development of novel pharmaceutical agents such as antibiotics, immunosuppressors, or antiviral and antitumor agents [211]. They differ from ribosomal peptides in that they can have a nonlinear structures (for example, cyclic or tree-like) and may contain non-standard amino acids [211]. This increases the number of possible building blocks from 20 to several hundreds, and certain amino acid masses not even known in advance. To this end, common approaches for sequencing ribosomal peptides using tandem mass spectrometry are not applicable to NRPs. For cyclic peptides, fragmentation steps beyond tandem MS are required, as tandem MS simply results in the linearization of the cyclic peptide. Nevertheless, NRPs are structurally much more restricted than the vast variety of metabolites known from plants or microbes. Computational methods for *de novo* sequencing and dereplication of NRPs have been established [17,211-214]. Unfortunately, these computational

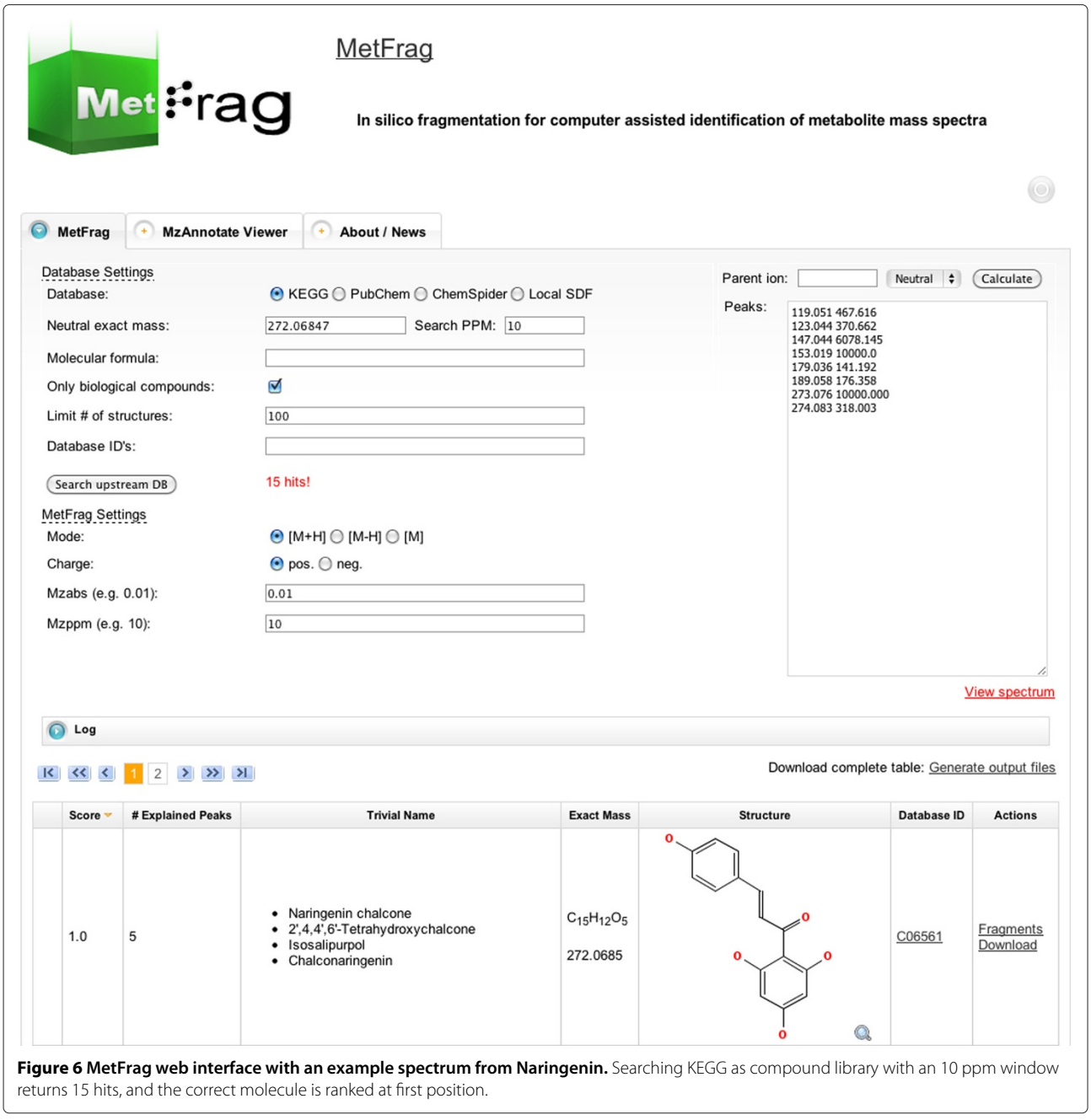

Figure 6 MetFrag web interface with an example spectrum from Naringenin. Searching KEGG as compound library with an 10 ppm window returns 15 hits, and the correct molecule is ranked at first position.

methods rely on the "polymeric character" of NRPs and, hence, cannot be generalized for analyzing other classes of metabolites.

Fragmentation trees

If we want to assign molecular formulas to the precursor and product ions, we may use the formula of the precursor to filter bogus explanations of the product ions, and *vice versa*. This fact has been exploited repeatedly, see for example [111,146] and Section "Molecular formula identification" above. This is only the most simplistic description of the fragmentation process: It is obvious that all product ions must be fragments of the precursor; but what is the dependency between the fragments? In fact, MS experts have drawn fragmentation diagrams for decades. For this task, the MS expert usually has to know the molecular structure of the compound and its tandem MS fragmentation spectrum.

Fragmentation trees must not be confused with *spectral trees* for multiple stage mass spectrometry [155], or the closely related *multistage mass spectral trees* of Rojas-Cherto *et al* [145] (referred to as "fragmentation trees"

in [145,215,216]). Spectral trees are a formal representation of the MS setup and describe the relationship between the MS^n spectra, but do not contain any additional information. We stress that all computational approaches described below target *tandem MS*, unless explicitly stated otherwise. To compute a fragmentation tree, we need neither spectral libraries nor molecular structure databases; this implies that this approach can target "true unknowns" that are not contained in *any* molecular structure database.

Böcker and Rasche [147] introduced fragmentation trees (see Figure 7) to find the molecular formula of an unknown, without using databases: Here, the highest-scoring fragmentation tree for each molecular formula candidate is used as the score of the molecular formula itself. Only later, fragmentation trees were conceived as a means of structural elucidation [148]. Algorithmic aspects of computing fragmentation trees were considered in [217]. Hufsky *et al* [56] computed fragmentation trees from EI fragmentation spectra with high mass accuracy, and used this to identify the molecular ion peak and the molecular formula of compounds. Fragmentation trees computed from both tandem MS [148] and EI fragmentation data [218] were found to be of good "structural quality" by expert evaluation. Finally, Scheubert *et al* [219,220] computed fragmentation trees from multiple MS data.

To further process fragmentation trees, Rasche *et al* [221] introduced fragmentation tree alignments to cluster unknown compounds, to predict chemical similarity, and to find structurally similar compounds in a spectral library using *FT-BLAST* (Fragmentation Tree Basic Local Alignment Search Tool). *FT-BLAST* also offers the possibility to identify bogus hits using a decoy database, allowing the user to report results for a pre-defined False Discovery Rate. Faster algorithms for the computationally demanding alignment of fragmentation trees were presented in [222]. *FT-BLAST* results were parsed for "characteristic substructures" in [209]. Rojas-Chertó *et al* [215] presented a related approach for the comparison of multistage mass spectral trees, based on transforming the trees into binary fingerprints and =comparing these fingerprints using the Tanimoto score (Jaccard index). This was applied for metabolite identification in [216].

Aligning fragmentation trees is similar in spirit to the feature tree comparison of Rarey and Dixon [223]. Feature trees were computed from the molecular structure of a *known* compound, and represent hydrophobic fragments and functional groups of the compound, and the way these groups are linked together.

Network reconstruction

Network elucidation based on mass spectrometry data is a wide field. On the one hand, detailed information like quantitative fluxes of the network is achieved by metabolic flux analysis. Here, based on isotope labeled compounds, the flux proceeding from these compounds can be tracked. On the other hand, measured metabolites can be mapped on a known network. This can

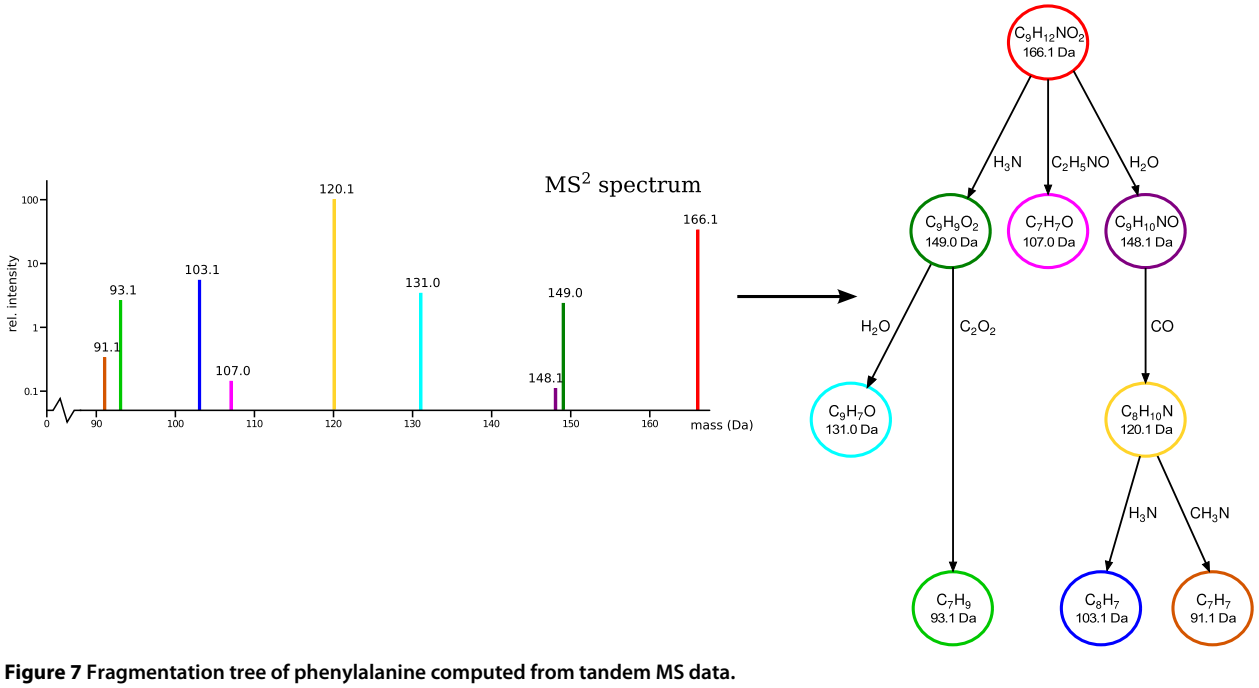

Figure 7 Fragmentation tree of phenylalanine computed from tandem MS data.

elucidate distinct metabolic pathways that are differentially "used" dependent on environmental conditions. Both of these variants require previous known metabolic network graphs. In this section, we will only cover the pure *de novo* reconstruction of networks from metabolite mass spectrometry data.

The reconstruction of networks solely from metabolic mass spectrometry data is a very young field of research. It can be subdivided into two main approaches: either the network reconstruction is based on metabolite level correlation of multiple mutant and wild type samples, or on data from only one sample by using information of common reactions or similarity between metabolites.

A first approach that used metabolite mass spectrometry data of multiple expressed samples was introduced by Fiehn *et al* [224]. Their method clusters metabolic phenotypes for example by principle component analysis (PCA). In contrast Arkin *et al* [225] and Kose *et al* [226] developed a method that does not group samples but metabolites with correlating intensity regarding all samples. Metabolites of a group have a similar concentration behavior in all samples. This leads to the assumption that the metabolites of a group are probably somehow connected in a metabolic network. As the concentration of metabolites taken from plants with identical genotype and grown under uniform conditions still show variability, this approach can also be used if no multiple mutant genotypes are available [227]. The disadvantage of this simple approach is, that it results in very dense networks that do not only cover direct reactions but also indirect ones. Krumsiek *et al*. 2011 [228] suggested to apply Gaussian graphical models to such data. Gaussian graphical networks have the ability to calculate only direct correlations while indirect correlations are not taken into account.

In 2006, Breitling *et al* [229] reconstructed networks based on high-resolution mass spectrometry data of only one dataset. They inferred accurate mass differences between all measured metabolites. These mass differences give evidences of biochemical transformations between the metabolites and allow the reconstruction of a network. Rogers *et al* [152] used a similar approach on molecular formula level to assign better molecular formulas to metabolites (see Section "Other approaches for molecular formula identification").

Watrous *et al* [230] used additional information from spectral alignments of tandem MS data to determine a structural similarity between the metabolites. Two structurally similar metabolites are supposed to be connected in the network (see Figure 8). They found the compound thanamycin in *Pseudomonas sp. SH-C52* that has an antifungal effect and protects sugar beet plants from infections by specific soil-borne fungi.

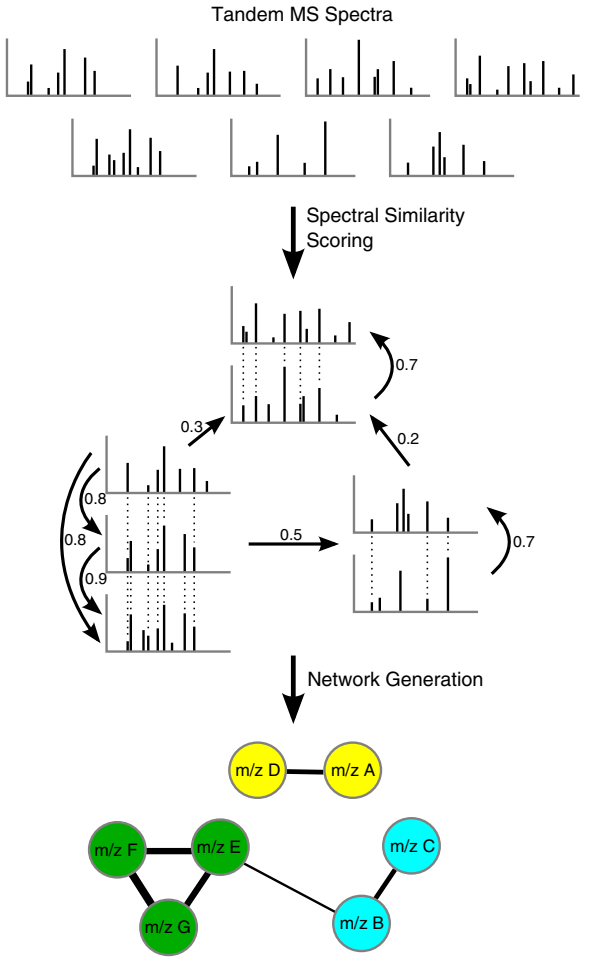

Figure 8 Using spectral alignment of tandem MS data to generate a molecular network. The thickness of the edges indicates the similarity between the spectra. Figure redrawn from Watrous *et al* [230].

Software packages

Several open source, or at least freely available, software packages assist with processing and analyzing GC-MS metabolomics data. The freely available *AMDIS* [231] is the most widely used method for extracting individual component spectra (mass spectral deconvolution) from GC-MS data. *MathDAMP* [232] helps with the identification and visualization of differences between complex metabolite profiles. *TagFinder* [233,234] supports the quantitative analysis of GC-MS-based metabolite profiling experiments. The *MetaboliteDetector* [235] detects and subsequently identifies metabolites and allows for the analysis of high-throughput data. *TargetSearch* [236] iteratively corrects and updates retention time indices for searching and identifying metabolites. *Metab* [237] is an R package that automates the pipeline for analysis of metabolomics GC-MS datasets processed by *AMDIS*.

PyMS [238] comprises several functions for processing raw GC-MS data, such as noise smoothing, baseline correction, peak detection, peak deconvolution, peak integration, and peak alignment. *ADAP-GC* 2.0 [239] helps with the deconvolution of coeluting metabolites, aligns components across samples and exports their qualitative and quantitative information. Castillo *et al*. 2011 [240] developed a tool to process GC×GC-TOF-MS data.

For LC-MS data, *XCMS* [13] enables retention time alignment, peak detection and peak matching. *XCMS*^{2} [241] additionally searches LC-MS/MS data against METLIN and also provides structural information for unknown metabolites. It also allows for the correction of mass calibration gaps [242] caused by regular switches between the analyte and a standard reference compound. *XCMS Online* [243] is the web-based version of the software. *AStream* [244] enables the detection of outliers and redundant peaks by intensity correlation and retention time, as well as isotope detection. *MetSign* [245] provides several bioinformatics tools for raw data deconvolution, metabolite putative assignment, peak list alignment, normalization, statistical significance tests, unsupervised pattern recognition, and time course analysis. *CAMERA* [246] is designed to post-process *XCMS* feature lists and integrates algorithms to extract compound spectra, annotate peaks, and propose compound masses in complex data. *MetExtract* [247] detects peaks corresponding to metabolites by chromatographic characteristics and isotope labeling. *IDEOM* [248] filters and detects peaks based on *XCMS* [13] and *mzMatch.R* [249], enables noise filtering based on [249,250] and allows for database matching and further statistics. Brodsky *et al* [251] presented a method for evaluating individual peaks in a LC-MS spectrum, based on replicate samples.

For both, GC-MS and LC-MS data, *MZmine* [252] and *MZmine2* [253] allow for data visualization, peak identification and peak list alignment. *MET-IDEA* [254] proceeds from complex raw data files to a complete data matrix. *MetAlign* [255] is capable of baseline correction, peak picking, as well as spectral alignment.

To compare the power of these software packages, an independent validation would be desirable. But up to now, there exists no such comparison. One reason is the limited amount of freely available mass spectra, see Section "Conclusion". Another reason is that some of the packages are developed for special experimental setups or instruments, and have to be adapted for other data, what makes an independent validation difficult.

Conclusion

No computational *de novo* method is able to elucidate the structure of a metabolite solely from mass spectral data. They can only reduce the search space or give hint to the structure or class of the compound. Computational mass spectrometry of small molecules is, at least compared to proteomics, still very much in a developmental state. This may be surprising, as methods development started out many years before computational mass spectrometry for proteins and peptides came into the focus of bioinformatics and cheminformatics research [183-185]. But since then, methods development in computational proteomics has proliferated [16-21] and long surpassed that in metabolomics and small molecule research. To a great extend, this can be attributed to the fact that freely sharing data and benchmark test sets has become a tradition in proteomics, providing developers of novel computational methods with the required input for training and evaluation of their methods.

In metabolomics, a comparative evaluation of methods is very limited due to restricted data sharing. Recently, a first benchmark test for small molecules was provided as part of the CASMI challenge[a]. CASMI is a contest in which GC-MS and LC-MS data is released to the public, and the computational mass spectrometry community is invited to identify the compounds. Results and methods will be published in a special issue of the Open Access MDPI journal *Metabolites*. This is a first step towards reliable evaluation of different computational methods for the identification of small molecules. Lately, the importance of computational methods has gained more attention in small molecule research: Citing Kind and Fiehn [33], "the ultimate success of structure elucidation of small molecules lies in better software programs and the development of sophisticated tools for data evaluation."

With the advent of novel computational approaches [169,206,207], searching spectral libraries may be replaced by searching molecular structure databases within in the next five to ten years. Beyond molecular databases, only few approaches aim at overcoming the limits of the "known universe of organic chemistry" [256], one example being fragmentation trees [56,148,221].

Endnote

[a]Critical Assesment of Small Molecule Identification, http://casmi-contest.org/.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KS wrote the Sections "Molecular formula identification" and "Network reconstruction". FH wrote the Sections "Searching spectral libraries" and "Identifying the unknowns". SB wrote the Section "Fragmentation trees". All authors read and approved the final manuscript.

Acknowledgements

K Scheubert funded by Deutsche Forschungsgemeinschaft (BO 1910/10). F Hufsky supported by the International Max Planck Research School Jena.

Author details
[1]Chair of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, Jena, Germany. [2]Max Planck Institute for Chemical Ecology, Beutenberg Campus, Jena, Germany.

Received: 23 November 2012 Accepted: 1 February 2013
Published: 1 March 2013

References

1. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL: **Metabolite identification via the Madison Metabolomics Consortium Database.** *Nat Biotechnol* 2008, **26**(2):162–164.
2. Last RL, Jones AD, Shachar-Hill Y: **Towards the plant metabolome and beyond.** *Nat Rev Mol Cell Biol* 2007, **8:**167–174.
3. Patti GJ, Yanes O, Siuzdak G: **Innovation: Metabolomics: The apogee of the omics trilogy.** *Nat Rev Mol Cell Biol* 2012, **13**(4):263–269.
4. Lederberg J: **Topological mapping of organic molecules.** *Proc Natl Acad Sci USA* 1965, **53**(1):134–139.
5. Lederberg J: **How DENDRAL was conceived and born.** In *ACM Conf. on the History of Medical Informatics, History of Medical Informatics archive*; 1987:5–19.
6. Mun IK, Mclafferty FW: **Computer methods of molecular structure elucidation from unknown mass spectra.** In *Supercomputers in Chemistry, ACS Symposium Series, chapter 9*; 1981:117–124.
7. Smith DH, Gray NA, Nourse JG, Crandell CW: **The DENDRAL project: Recent advances in computer-assisted structure elucidation.** *Anal Chim Acta* 1981, **133**(4):471–497.
8. November JA: *Digitizing Life: The Introduction of Computers to Biology and Medicine*. PhD thesis. Princeton, USA: Princeton University; 2006.
9. Gasteiger J, Hanebeck W, Schulz KP: **Prediction of mass spectra from structural information.** *J Chem Inf Comput Sci* 1992, **32**(4):264–271.
10. Bylund D, Danielsson R, Malmquist G, Markides KE: **Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data.** *J Chromatography A* 2002, **961:**237–244.
11. Jeong J, Shi X, Zhang X, Kim S, Shen C: **Model-based peak alignment of metabolomic profiling from comprehensive two-dimensional gas chromatography mass spectrometry.** *BMC Bioinformatics* 2012, **13:**27.
12. Lommen A, Kools HJ: **MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware.** *Metabolomics* 2012, **8**(4):719–726.
13. Smith CA, Want EJ, 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**(3):779–787.
14. Garkani-Nejad Z, Karlovits M, Demuth W, Stimpfl T, Vycudilik W, Jalali-Heravi M, Varmuza K: **Prediction of gas chromatographic retention indices of a diverse set of toxicologically relevant compounds.** *J Chromatogr A* 2004, **1028**(2):287–295.
15. Stein SE, Babushok VI, Brown RL, Linstrom PJ: **Estimation of Kováts retention indices using group contributions.** *J Chem Inf Model* 2007, **47**(3):975–980.
16. Cox J, Mann M: **MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.** *Nat Biotechnol* 2008, **26**(12):1367–1372.
17. Bandeira N, Pham V, Pevzner P, Arnott D, Lill JR: **Automated de novo protein sequencing of monoclonal antibodies.** *Nat Biotechnol* 2008, **26**(12):1336–1338.
18. Liu G, Zhang J, Larsen B, Stark C, Breitkreutz A, Lin ZY, Breitkreutz BJ, Ding Y, Colwill K, Pasculescu A, Pawson T, Wrana JL, Nesvizhskii AI, Raught B, Tyers M, Gingras AC: **ProHits: Integrated software for mass spectrometry-based interaction proteomics.** *Nat Biotechnol* 2010, **28**(10):1015–1017.
19. Fusaro VA, Mani DR, Mesirov JP, Carr SA: **Prediction of high-responding peptides for targeted protein assays by mass spectrometry.** *Nat Biotechnol* 2009, **27**(2):190–198.
20. Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP: **Intensity-based protein identification by machine learning from a library of tandem mass spectra.** *Nat Biotechnol* 2004, **22**(2):214–219.
21. Mann M: **Comparative analysis to guide quality improvements in proteomics.** *Nat Methods* 2009, **6**(10):717–719.
22. Böcker S: **Sequencing from compomers: Using mass spectrometry for DNA de-novo sequencing of 200+ nt.** *J Comput Biol* 2004, **11**(6):1110–1134.
23. Böcker S: **Simulating multiplexed SNP discovery rates using base-specific cleavage and mass spectrometry.** *Bioinformatics* 2007, **23**(2):e5—e12.
24. Böcker S, Kehr B, Rasche F: **Determination of glycan structure from tandem mass spectra.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8**(4):976–986.
25. Goldberg D, Bern MW, Li B, Lebrilla CB: **Automatic determination of O-glycan structure from fragmentation spectra.** *J Proteome Res* 2006, **5**(6):1429–1434.
26. Goldberg D, Bern MW, North SJ, Haslam SM, Dell A: **Glycan family analysis for deducing N-glycan topology from single MS.** *Bioinformatics* 2009, **25**(3):365–371.
27. Baumgaertel A, Scheubert K, Pietsch B, Kempe K, Crecelius AC, Böcker S, Schubert US: **Analysis of different synthetic homopolymers by the use of a new calculation software for tandem mass spectra.** *Rapid Commun Mass Spectrom* 2011, **25**(12):1765–1778.
28. Thalassinos K, Jackson AT, Williams JP, Hilton GR, Slade SE, Scrivens JH: **Novel software for the assignment of peaks from tandem mass spectrometry spectra of synthetic polymers.** *J Am Soc Mass Spectrom* 2007, **18**(7):1324–1331.
29. Katajamaa M, Oresic M: **Data processing for mass spectrometry-based metabolomics.** *J Chromatogr A* 2007, **1158**(1-2):318–328.
30. Wishart DS: **Current progress in computational metabolomics.** *Brief Bioinform* 2007, **8**(5):279–293.
31. Stein SE: **Mass spectral reference libraries: An ever-expanding resource for chemical identification.** *Anal Chem* 2012, **84**(17):7274–7282.
32. Han J, Datla R, Chan S, Borchers CH: **Mass spectrometry-based technologies for high-throughput metabolomics.** *Bioanalysis* 2009, **1**(9):1665–1684.
33. Kind T, Fiehn O: **Advances in structure elucidation of small molecules using mass spectrometry.** *Bioanal Rev* 2010, **2**(1-4):23–60.
34. Xiao JF, Zhou B, Ressom HW: **Metabolite identification and quantitation in LC-MS/MS-based metabolomics.** *Trends Analyt Chem* 2012, **32:**1–14.
35. Fiehn O: **Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry.** *Trends Analyt Chem* 2008, **27**(3):261–269.
36. Valkenborg D, Mertens I, Lemière F, Witters E, Burzykowski T: **The isotopic distribution conundrum.** *Mass Spectrom Rev* 2012, **31**(1):96–109.
37. Neumann S, Böcker S: **Computational mass spectrometry for metabolomics – a review.** *Anal Bioanal Chem* 2010, **398**(7):2779–2788.
38. Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L: **Metabolite profiling: From diagnostics to systems biology.** *Nat Rev Mol Cell Biol* 2004, **5**(9):763–769.
39. Sweedler JV: **Metabolomics in analytical chemistry.** *Anal Chem* 2012, **84**(14):5833.
40. Champarnaud E, Hopley C: **Evaluation of the comparability of spectra generated using a tuning point protocol on twelve electrospray ionisation tandem-in-space mass spectrometers.** *Rapid Commun Mass Spectrom* 2011, **25**(8):1001–1007.
41. Bristow AWT, Webb KS, Lubben AT, Halket J: **Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries.** *Rapid Commun Mass Spectrom* 2004, **18**(13):1447–1454.
42. Halket JM, Waterman D, Przyborowska AM, Patel RKP, Fraser PD, Bramley PM: **Chemical derivatization and mass spectral libraries in metabolic profiling by GC/MS and LC/MS/MS.** *J Exp Bot* 2005, **56**(410):219–243.
43. Goodley P: **Maximizing MS/MS fragmentation in the ion trap using CID voltage ramping.** Technical Report 5988-0704EN, Agilent Technologies, 2007.
44. Hopley C, Bristow T, Lubben A, Simpson A, Bull E, Klagkou K, Herniman J: **Langley J Towards a universal product ion mass spectral library –**

reproducibility of product ion spectra across eleven different mass spectrometers. *Rapid Commun Mass Spectrom* 2008, **22**(12):1779–1786.
45. Palit M, Mallard G: **Fragmentation energy index for universalization of fragmentation energy in ion trap mass spectrometers for the analysis of chemical weapon convention related chemicals by atmospheric pressure ionization-tandem mass spectrometry analysis.** *Anal Chem* 2009, **81**(7):2477–2485.
46. Knochenmuss R, Zenobi R: **MALDI ionization: the role of in-plume processes.** *Chem Rev* 2003, **103**(2):441–452.
47. Kostiainen R, Kotiaho T, Kuuranne T, Auriola S: **Liquid chromatography/atmospheric pressure ionization-mass spectrometry in drug metabolism studies.** *J Mass Spectrom* 2003, **38**(4):357–372.
48. Marchi I, Rudaz S, Veuthey JL: **Atmospheric pressure photoionization for coupling liquid-chromatography to mass spectrometry: a review.** *Talanta* 2009, **78**(1):1–18.
49. Takáts Z, Wiseman JM, Gologan B, Cooks RG: **Mass spectrometry sampling under ambient conditions with desorption electrospray ionization.** *Science* 2004, **306**(5695):471–473.
50. Horvath CG, Lipsky SR: **Use of liquid ion exchange chromatography for the separation of organic compounds.** *Nature* 1966, **211**(5050):748–749.
51. MacNair JE, Lewis KC, Jorgenson JW: **Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns.** *Anal Chem* 1997, **69**(6):983–989.
52. Gao X, Zhang Q, Meng D, Isaac G, Zhao R, Fillmore TL, Chu RK, Zhou J, Tang K, Hu Z, Moore RJ, Smith RD, Katze MG, Metz TO: **A reversed-phase capillary ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) method for comprehensive top-down/bottom-up lipid profiling.** *Anal Bioanal Chem* 2012, **402**(9):2923–2933.
53. Zubarev R, Mann M: **On the proper use of mass accuracy in proteomics.** *Mol Cell Proteomics* 2007, **6**(3):377–381.
54. Cajka T, Hajslova J, Lacina O, Mastovska K, Lehotay SJ: **Rapid analysis of multiple pesticide residues in fruit-based baby food using programmed temperature vaporiser injection-low-pressure gas chromatography-high-resolution time-of-flight mass spectrometry.** *J Chromatogr A* 2008, **1186**(1-2):281–294.
55. Hernández F, Portolés T, Pitarch E, López FJ: **Gas chromatography coupled to high-resolution time-of-flight mass spectrometry to analyze trace-level organic compounds in the environment, food safety and toxicology.** *Trends Anal Chem* 2011, **30**(2):388–400.
56. Hufsky F, Rempt M, Rasche F, Pohnert G, Böcker S: **De novo analysis of electron impact mass spectra using fragmentation trees.** *Anal Chim Acta* 2012, **739:**67–76.
57. Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, Trethewey RN, Lange BM, Wurtele ES, Sumner LW: **Potential of metabolomics as a functional genomics tool.** *Trends Plant Sci* 2004, **9**(9):418–425.
58. Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, Taylor J, Fiehn O, Goodacre R, Bino RJ, Hall R, Kopka J, Lane GA, Lange BM, Liu JR, Mendes P, Nikolau BJ, Oliver SG, Paton NW, Rhee S, Roessner-Tunali U, Saito K, Smedsgaard J, Sumner LW, Wang T, Walsh S, Wurtele ES, Kell DB: **A proposed framework for the description of plant metabolomics experiments and their results.** *Nat Biotechnol* 2004, **22**(12):1601–1606.
59. Board Members MSI, Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O: **The metabolomics standards initiative.** *Nat Biotechnol* 2007, **25**(8):846–848.
60. Goodacre R, Broadhurst D, Smilde AK, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wulfert F: **Proposed minimum reporting standards for data analysis in metabolomics.** *Metabolomics* 2007, **3:**231–241.
61. Sumner LW, Amberg A, Barrett D, Beale M, Beger R, Daykin C, Fan T, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane A, Lindon JC, Marriott P, Nicholls A, Reily M, Thaden J, Viant MR: **Proposed minimum reporting standards for chemical analysis.** *Metabolomics* 2007, **3**(3):211–221.
62. Horai H, Arita M, Nishioka T: **Comparison of ESI-MS spectra in MassBank database.** In *Proc. of Conference on BioMedical Engineering and Informatics (BMEI 2008), volume 2*; 2008:853–857.
63. 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, *et al*: **MassBank: A public repository for sharing mass spectral data for life sciences.** *J Mass Spectrom* 2010, **45**(7):703–714.
64. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G: **METLIN: A metabolite mass spectral database.** *Ther Drug Monit* 2005, **27**(6):747–751.
65. Tautenhahn R, Cho K, Uritboonthai W, Zhu Z, Patti GJ, Siuzdak G: **An accelerated workflow for untargeted metabolomics using the METLIN database.** *Nat Biotechnol* 2012, **30**(9):826–828.
66. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmüller E, Dörmann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D: **GMD@CSB.DB: The Golm Metabolome Database.** *Bioinformatics* 2005, **21**(8):1635–1638.
67. Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, Shinozaki K, Hirai MY, Sakurai T, Kikuchi J, Saito K: **PRIMe: A web site that assembles tools for metabolomics and transcriptomics.** *In Silico Biol* 2008, **8**(3-4):339–345.
68. Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A: **MeltDB: A software platform for the analysis and integration of metabolomics experiment data.** *Bioinformatics* 2008, **24**(23):2726–2732.
69. Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K, Booth T, Bougueleret L, Burns G, Chapman B, Clark T, Coleman LA, Copeland J, Das S, de Daruvar A, de Matos P, Dix I, Edmunds S, Evelo CT, Forster MJ, Gaudet P, Gilbert J, Goble C, Griffin JL, Jacob D, *et al*: **Toward interoperable bioscience data.** *Nat Genet* 2012, **44**(2):121–126.
70. Kind T, Wohlgemuth G, Lee DY, Lu Y, Palazoglu M, Shahbaz S, Fiehn O: **FiehnLib: Mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry.** *Anal Chem* 2009, **81**(24):10038–10048.
71. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC: **On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study.** *J Mass Spectrom* 2009, **44**(4):485–493.
72. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, Csaszar E, Köfeler HC: **On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm.** *J Mass Spectrom* 2009, **44**(4):494–502.
73. Sana TR, Roark JC, Li X, Waddell K, Fischer SM: **Molecular formula and METLIN Personal Metabolite Database matching applied to the identification of compounds generated by LC/TOF-MS.** *J Biomol Tech* 2008, **19**(4):258–266.
74. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, MacInnis GD, *et al*: **HMDB: The human metabolome database.** *Nucleic Acids Res* 2007, **35**(suppl1):D521—526.
75. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, Souza AD, Zuniga A, Dawe M, *et al*: **HMDB: A knowledgebase for the human metabolome.** *Nucleic Acids Res* 2009, **37:**D603—D610.
76. Matsuda F, Yonekura-Sakakibara K, Niida R, Kuromori T, Shinozaki K, Saito K: **MS/MS spectral tag based annotation of non-targeted profile of plant secondary metabolites.** *Plant J* 2008, **57**(3):555–577.
77. Sparkman OD: **Evaluating electron ionization mass spectral library search results.** *J Am Soc Mass Spectrom* 1996, **7**(4):313–318.

78. Hertz HS, Hites RA, Biemann K: **Identification of mass spectra by computer-searching a file of known spectra.** *Anal Chem* 1971, **43**(6):681–691.
79. McLafferty F, Hertel R, Villwock R: **Computer identification of mass spectra: VI. Probability based matching of mass spectra: Rapid identification of specific compounds in mixtures.** *Org Mass Spectrom* 1974, **9**(7):690–702.
80. McLafferty FW, Zhang MY, Stauffer DB, Loh SY: **Comparison of algorithms and databases for matching unknown mass spectra.** *J Am Soc Mass Spectrom* 1998, **9**(1):92–95.
81. Atwater BL, Stauffer DB, McLafferty FW, Peterson DW: **Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra.** *Anal Chem* 1985, **57**(4):899–903.
82. Damen H, Henneberg D, Weimann B: **SISCOM – a new library search system for mass spectra.** *Anal Chim Acta* 1978, **103:**289–302.
83. Sokolow S, Karnofsky J, Gustafson P: **The finnigan library search programs.** Finnigan Application Report 2, Finnigan Corp., 1978.
84. Stein SE, Scott DR: **Optimization and testing of mass spectral library search algorithms for compound identification.** *J Am Soc Mass Spectrom* 1994, **5**(9):859–866.
85. Rasmussen GT, Isenhour TL, Marshall JC: **Mass spectral library searches using ion series data compression.** *J Chem Inf Comput Sci* 1979, **19**(2):98–104.
86. Koo I, Zhang X, Kim S: **Wavelet- and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry.** *Anal Chem* 2011, **83**(14):5631–5638.
87. Kim S, Koo I, Wei X, Zhang X: **A method of finding optimal weight factors for compound identification in gas chromatography-mass spectrometry.** *Bioinformatics* 2012, **28**(8):1158–1163.
88. Stein SE: **Estimating probabilities of correct identification from results of mass spectral library searches.** *J Am Soc Mass Spectrom* 1994, **5**(4):316–323.
89. Jeong J, Shi X, Zhang X, Kim S, Shen C: **An empirical bayes model using a competition score for metabolite identification in gas chromatography mass spectrometry.** *BMC Bioinformatics* 2011, **12:**392.
90. Josephs JL, Sanders M: **Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers.** *Rapid Commun Mass Spectrom* 2004, **18**(7):743–759.
91. Milman BL: **Towards a full reference library of MS^n spectra. Testing of a library containing 3126 MS^2 spectra of 1743 compounds.** *Rapid Commun Mass Spectrom* 2005, **19**(19):2833–2839.
92. Pavlic M, Libiseller K, Oberacher H: **Combined use of ESI-QqTOF-MS and ESI-QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs.** *Anal Bioanal Chem* 2006, **386**(1):69–82.
93. Wan KX, Vidavsky I, Gross ML: **Comparing similar spectra: From similarity index to spectral contrast angle.** *J Am Soc Mass Spectrom* 2002, **13**(13):85–88.
94. Zhou B, Cheema AK, Ressom HW: **SVM-based spectral matching for metabolite identification.** *Conf Proc IEEE Eng Med Biol Soc* 2010, **2010:**756–759.
95. Hansen ME, Smedsgaard J: **A new matching algorithm for high resolution mass spectra.** *J Am Soc Mass Spectrom* 2004, **15:**1173–1180.
96. Matusita K: **Decision rule, based on the distance, for the classification problem.** *Ann Inst Statist Math* 1956, **8**(1):67–77.
97. Mylonas R, Mauron Y, Masselot A, Binz PA, Budin N, Fathi M, Viette V, Hochstrasser DF, Lisacek F: **X-rank: A robust algorithm for small molecule identification using tandem mass spectrometry.** *Anal Chem* 2009, **81**(18):7604–7610.
98. Gergov M, Weinmann W, Meriluoto J, Uusitalo J, Ojanperä I: **Comparison of product ion spectra obtained by liquid chromatography/triple-quadrupole mass spectrometry for library search.** *Rapid Commun Mass Spectrom* 2004, **18**(10):1039–1046.
99. Issaq HJ, Van QN, Waybright TJ, Muschik GM, Veenstra TD: **Analytical and statistical approaches to metabolomics research.** *J Sep Sci* 2009, **32**(13):2183–2199.
100. Böcker S, Lipták Zs: **Efficient mass decomposition.** In *Proc. of ACM Symposium on Applied Computing (ACM SAC 2005)*. New York: ACM press; 2005:151–157.
101. Böcker S, Lipták Zs: **A fast and simple algorithm for the Money Changing Problem.** *Algorithmica* 2007, **48**(4):413–432.
102. Böcker S, Letzel M, Lipták Zs, Pervukhin A: **SIRIUS: Decomposing isotope patterns for metabolite identification.** *Bioinformatics* 2009, **25**(2):218–224.
103. Böcker S, Lipták Zs, Martin M, Pervukhin A, Sudek H: **DECOMP—from interpreting mass spectrometry peaks to solving the money changing problem.** *Bioinformatics* 2008, **24**(4):591–593.
104. Kind T, Fiehn O: **Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry.** *BMC Bioinformatics* 2007, **8:**105.
105. Yergey JA: **A general approach to calculating isotopic distributions for mass spectrometry.** *Int J Mass Spectrom Ion Phys* 1983, **52**(2–3):337–349.
106. Rockwood AL, Van Orden, S L: **Ultrahigh-speed calculation of isotope distributions.** *Anal Chem* 1996, **68:**2027–2030.
107. Rockwood AL, Haimi P: **Efficient calculation of accurate masses of isotopic peaks.** *J Am Soc Mass Spectrom* 2006, **17**(3):415–419.
108. Snider RK: **Efficient calculation of exact mass isotopic distributions.** *J Am Soc Mass Spectrom* 2007, **18**(8):1511–1515.
109. Claesen J, Dittwald P, Burzykowski T, Valkenborg D: **An efficient method to calculate the aggregated isotopic distribution and exact center-masses.** *J Am Soc Mass Spectrom* 2012, **23**(4):753–63.
110. Fernandez-de-Cossio Diaz J, Fernandez-de-Cossio J: **Computation of isotopic peak center-mass distribution by Fourier transform.** *Anal Chem* 2012, **84**(16):7052–7056.
111. Pluskal T, Uehara T, Yanagida M: **Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching.** *Anal Chem* 2012, **84**(10):4396–4403.
112. Matsuda F, Shinbo Y, Oikawa A, Hirai MY, Fiehn O, Kanaya S, Saito K: **Assessment of metabolome annotation quality: A method for evaluating the false discovery rate of elemental composition searches.** *PLoS One* 2009, **4**(10):e7490.
113. Robertson AL, Hamming MC: **MASSFORM: a computer program for the assignment of elemental compositions to high resolution mass spectral data.** *Biomed Mass Spectrom* 1977, **4**(4):203–208.
114. Dromey RG, Foyster GT: **Calculation of elemental compositions from high resolution mass spectral data.** *Anal Chem* 1980, **52**(3):394–398.
115. Fürst A, Clerc JT, Pretsch E: **A computer program for the computation of the molecular formula.** *Chemom Intell Lab Syst* 1989, **5:**329–334.
116. Böcker S, Letzel M, Lipták Zs, Pervukhin A: **Decomposing metabolomic isotope patterns.** In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006), volume 4175 of Lect Notes Comput Sci*. Berlin: Springer; 2006:12–23.
117. Kind T, Fiehn O: **Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm.** *BMC Bioinformatics* 2006, **7**(1):234.
118. Wieser ME: **Atomic weights of the elements 2005 (IUPAC technical report).** *Pure Appl Chem* 2006, **78**(11):2051–2066.
119. Audi G, Wapstra A, Thibault C: **The AME2003 atomic mass evaluation (ii): Tables, graphs, and references.** *Nucl Phys A* 2003, **729:**129–336.
120. de Laeter JR, Böhlke JK, Biévre PD, Hidaka H, Peiser HS, Rosman KJR, Taylor PDP: **Atomic weights of the elements. Review 2000 (IUPAC technical report).** *Pure Appl Chem* 2003, **75**(6):683–800.
121. Biemann K: *Mass Spectrometry: Organic Chemical Applications*. New York: McGraw-Hill; 1962.
122. Kubinyi H: **Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem.** *Anal Chim Acta* 1991, **247:**107–119.
123. Roussis SG, Proulx R: **Reduction of chemical formulas from the isotopic peak distributions of high-resolution mass spectra.** *Anal Chem* 2003, **75**(6):1470–1482.
124. Yamamoto H, McCloskey JA: **Calculations of isotopic distribution in molecules extensively labeled with heavy isotopes.** *Anal Chem* 1977, **49**(2):281–283.

125. Hsu CS: **Diophantine approach to isotopic abundance calculations.** *Anal Chem* 1984, **56**(8):1356–1361.
126. Rockwood AL: **Relationship of fourier transforms to isotope distribution calculations.** *Rapid Commun Mass Spectrom* 1995, **9:**103–105.
127. Rockwood AL, Van Orden, S L, Smith RD: **Rapid calculation of isotope distributions.** *Anal Chem* 1995, **67:**2699–2704.
128. Rockwood AL, Orden SLV, Smith RD: **Ultrahigh resolution isotope distribution calculations.** *Rapid Commun Mass Spectrom* 1996, **10:**54–59.
129. Li L, Kresh JA, Karabacak NM, Cobb JS, Agar JN, Hong P: **A hierarchical algorithm for calculating the isotopic fine structures of molecules.** *J Am Soc Mass Spectrom* 2008, **19**(12):1867–1874.
130. Li L, Karabacak NM, Cobb JS, Wang Q, Hong P, Agar JN: **Memory-efficient calculation of the isotopic mass states of a molecule.** *Rapid Commun Mass Spectrom* 2010, **24**(18):2689–2696.
131. Olson MT, Yergey AL: **Calculation of the isotope cluster for polypeptides by probability grouping.** *J Am Soc Mass Spectrom* 2009, **20**(2):295–302.
132. Böcker S: **Comment on "An efficient method to calculate the aggregated isotopic distribution and exact center-masses" by Claesen et al.** *J Am Soc Mass Spectrom* 2012, **23**(10):1826–1827.
133. Fernandez-de-Cossio J: **Computation of the isotopic distribution in two dimensions.** *Anal Chem* 2010, **82**(15):6726–6729.
134. Claesen J, Dittwald P, Burzykowski T, Valkenborg D: **Reply to the comment on: "An efficient method to calculate the aggregated isotopic distribution and exact center-masses" by Claesen et al.** *J Am Soc Mass Spectrom* 2012, **23**(10):1828–1829.
135. Stoll N, Schmidt E, Thurow K: **Isotope pattern evaluation for the reduction of elemental compositions assigned to high-resolution mass spectral data from electrospray ionization fourier transform ion cyclotron resonance mass spectrometry.** *J Am Soc Mass Spectrom* 2006, **17**(12):1692–1699.
136. Tong H, Bell D, Tabei K, Siegel MM: **Automated data massaging, interpretation, and e-mailing modules for high throughput open access mass spectrometry.** *J Am Soc Mass Spectrom* 1999, **10**(11):1174–1187.
137. Alon T, Amirav A: **Isotope abundance analysis methods and software for improved sample identification with supersonic gas chromatography/mass spectrometry.** *Rapid Commun Mass Spectrom* 2006, **20**(17):2579–2588.
138. Ipsen A, Want EJ, Ebbels TMD: **Construction of confidence regions for isotopic abundance patterns in LC/MS data sets for rigorous determination of molecular formulas.** *Anal Chem* 2010, **82**(17):7319–7328.
139. Rodgers RP, Blumer EN, Hendrickson CL, Marshall AG: **Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement.** *J Am Soc Mass Spectrom* 2000, **11**(10):835–840.
140. Hegeman AD, Schulte CF, Cui Q, Lewis IA, Huttlin EL, Eghbalnia H, Harms AC, Ulrich EL, Markley JL, Sussman MR: **Stable isotope assisted assignment of elemental compositions for metabolomics.** *Anal Chem* 2007, **79**(1):6912–6921.
141. Giavalisco P, Li Y, Matthes A, Eckhardt A, Hubberten HM, Hesse H, Segu S, Hummel J, Köhl K, Willmitzer L: **Elemental formula annotation of polar and lipophilic metabolites using ^{13}C, ^{15}N and ^{34}S isotope labelling, in combination with high-resolution mass spectrometry.** *Plant J* 2011, **68**(2):364–376.
142. Baran R, Bowen BP, Bouskill NJ, Brodie EL, Yannone SM, Northen TR: **Metabolite identification in *Synechococcus sp. PCC 7002* using untargeted stable isotope assisted metabolite profiling.** *Anal Chem* 2010, **82**(21):9034–9042.
143. Jarussophon S, Acoca S, Gao JM, Deprez C, Kiyota T, Draghici C, Purisima E, Konishi Y: **Automated molecular formula determination by tandem mass spectrometry (MS/MS).** *Analyst* 2009, **134**(4):690–700.
144. Konishi Y, Kiyota T, Draghici C, Gao JM, Yeboah F, Acoca S, Jarussophon S, Purisima E: **Molecular formula analysis by an MS/MS/MS technique to expedite dereplication of natural products.** *Anal Chem* 2007, **79**(3):1187–1197.
145. Rojas-Chertó M, Kasper PT, Willighagen EL, Vreeken RJ, Hankemeier T, Reijmers TH: **Elemental composition determination based on MS^n.** *Bioinformatics* 2011, **27:**2376–2383.
146. Zurek G, Krebs I, Götz S, Scheible H, Laufer S, Kammerer B, Albrecht W: **A software solution automatically assigns formulae for construction of fragmentation pathways accelerating drug elucidation with ESI-TOF.** *LCGC Eur Appl Book* 2008, **7:**31–33.
147. Böcker S, Rasche F: **Towards de novo identification of metabolites by analyzing tandem mass spectra.** *Bioinformatics* 2008, **24:**I49—I55.
148. Rasche F, Svatoš A, Maddula RK, Böttcher C, Böcker S: **Computing fragmentation trees from tandem mass spectrometry data.** *Anal Chem* 2011, **83**(4):1243–1251.
149. Singleton KE, Cooks RG, Wood KV: **Utilization of natural isotopic abundance ratios in tandem mass spectrometry.** *Anal Chem* 1983, **55**(4):762–764.
150. Rockwood AL, Kushnir MM, Nelson GJ: **Dissociation of individual isotopic peaks: Predicting isotopic distributions of product ions in MS^n.** *J Am Soc Mass Spectrom* 2003, **14:**311–322.
151. Ramaley L, Herrera LC: **Software for the calculation of isotope patterns in tandem mass spectrometry.** *Rapid Commun Mass Spectrom* 2008, **22**(17):2707–2714.
152. Rogers S, Scheltema RA, Girolami M, Breitling R: **Probabilistic assignment of formulas to mass peaks in metabolomics experiments.** *Bioinformatics* 2009, **25**(4):512–518.
153. Stein SE: **Chemical substructure identification by mass spectral library searching.** *J Am Soc Mass Spectrom* 1995, **6**(8):644–655.
154. Demuth W, Karlovits M, Varmuza K: **Spectral similarity versus structural similarity: Mass spectrometry.** *Anal Chim Acta* 2004, **516**(1-2):75–85.
155. Sheldon MT, Mistrik R, Croley TR: **Determination of ion structures in structurally related compounds using precursor ion fingerprinting.** *J Am Soc Mass Spectrom* 2009, **20**(3):370–376.
156. Werner E, Heilier JF, Ducruix C, Ezan E, Junot C, Tabet JC: **Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends.** *J Chromatogr B* 2008, **871**(2):143–163.
157. Venkataraghavan R, McLafferty FW, van Lear GE: **Computer-aided interpretation of mass spectra.** *Org Mass Spectrom* 1969, **2**(1):1–15.
158. Kwok KS, Venkataraghavan R, McLafferty FW: **Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system.** *J Am Chem Soc* 1973, **95**(13):4185–4194.
159. Scott DR: **Pattern recognition/expert system for mass spectra of volatile toxic and other organic compounds.** *Anal Chim Acta* 1992, **265:**43–54.
160. Scott DR: **Rapid and accurate method for estimating molecular weights of organic compounds from low resolution mass spectra.** *Chemometr Intell Lab* 1992, **16**(3):193–202.
161. Scott DR, Levitsky A, Stein SE: **Large scale evaluation of a pattern recognition/expert system for mass spectral molecular weight estimation.** *Anal Chim Acta* 1993, **278:**137–147.
162. Henneberg D, Weimann B, Zalfen U: **Computer-aided interpretation of mass spectra using databases with spectra and structures. I. Structure searches.** *Org Mass Spectrom* 1993, **28:**198–206.
163. Varmuza K, Werther W: **Mass spectral classifiers for supporting systematic structure elucidation.** *J Chem Inf Comp Sci* 1996, **36**(2):323–333.
164. Schymanski EL, Meinert C, Meringer M, Brack W: **The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis.** *Anal Chim Acta* 2008, **615**(2):136–147.
165. Xiong Q, Zhang Y, Li M: **Computer-assisted prediction of pesticide substructure using mass spectra.** *Anal Chim Acta* 2007, **593**(2):199–206.
166. Zhang L, Liang Y, Chen A: **Selection of neutral losses and characteristic ions for mass spectral classifier.** *Analyst* 2009, **134**(8):1717–1724.
167. Hummel J, Strehmel N, Selbig J, Walther D, Kopka J: **Decision tree supported substructure prediction of metabolites from GC-MS profiles.** *Metabolomics* 2010, **6**(2):322–333.
168. Tsugawa H, Tsujimoto Y, Arita M, Bamba T, Fukusaki E: **GC/MS based metabolomics: Development of a data mining system for**

metabolite identification by using soft independent modeling of class analogy (SIMCA). *BMC Bioinformatics* 2011, **12:**131.
169. Heinonen M, Shen H, Zamboni N, Rousu J: **Metabolite identification and molecular fingerprint prediction via machine learning.** *Bioinformatics* 2012, **28**(18):2333–2341.
170. Kerber A, Laue R, Moser D: **Ein Strukturgenerator für molekulare Graphen.** *Anal Chim Acta* 1990, **235:**221–228.
171. Benecke C, Grund R, Hohberger R, Kerber A, Laue R, Wieland T: **MOLGEN+, a generator of connectivity isomers and stereoisomers for molecular structure elucidation.** *Anal Chim Acta* 1995, **314:**141–147.
172. Kerber A, Laue R, Meringer M, Rücker C: **Molecules in silico: The generation of structural formulae and its applications.** *J Comput Chem Japan* 2004, **3**(3):85–96.
173. Molchanova MS, Shcherbukhin VV, Zefirov NS: **Computer generation of molecular structures by the SMOG program.** *J Chem Inf Comput Sci* 1996, **36**(4):888–899.
174. Fontana P, Pretsch E: **Automatic spectra interpretation, structure generation, and ranking.** *J Chem Inf Comput Sci* 2002, **42**(3):614–619.
175. Gray NAB, Buchs A, Smith DH, Djerassi C: **Computer assisted structural interpretation of mass spectral data.** *Helv Chim Acta* 1981, **64**(2):458–470.
176. Faulon JL: **Stochastic generator of chemical structure: (1) Application to the structure elucidation of large molecules.** *J Chem Inf Comput Sci* 1994, **34:**1204–1218.
177. Peironcely JE, Rojas-Chertó M, Fichera D, Reijmers T, Coulier L, Faulon JL, Hankemeier T: **OMG: open molecule generator.** *J Cheminform* 2012, **4**(1):21.
178. Hill DW, Kertesz TM, Fontaine D, Friedman R, Grant DF: **Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra.** *Anal Chem* 2008, **80**(14):5574–5582.
179. 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.
180. Kangas LJ, Metz TO, Isaac G, Schrom BT, Ginovska-Pangovska B, Wang L, Tan L, Lewis RR, Miller JH: **In silico identification software (ISIS): A machine learning approach to tandem mass spectral identification of lipids.** *Bioinformatics* 2012, **28**(13):1705–1713.
181. Kameyama A, Nakaya S, Ito H, Kikuchi N, Angata T, Nakamura M, Ishida HK, Narimatsu H: **Strategy for simulation of CID spectra of N-linked oligosaccharides toward glycomics.** *J Proteome Res* 2006, **5**(4):808–814.
182. Zhang H, Singh S, Reinhold VN: **Congruent strategies for carbohydrate sequencing. 2. FragLib: An MS^n spectral library.** *Anal Chem* 2005, **77**(19):6263–6270.
183. Chen T, Kao MY, Tepel M, Rush J, Church GM: *A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry*: Society for Industrial and Applied Mathematics; 2000.
184. Dančík V, Addona TA, Clauser KR, Vath JE, Pevzner PA: **De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach.** In *Proc. of Research in Computational Molecular Biology (RECOMB 1999)*:135–144.
185. Taylor JA, Johnson RS: **Sequence database searches via de novo peptide sequencing by tandem mass spectrometry.** *Rapid Commun Mass Spectrom* 1997, **11:**1067–1075.
186. Jalali-Heravi M, Fatemi M: **Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network.** *Anal Chim Acta* 2000, **415**(1-2):95–103.
187. Cooks RG: **Bond formation upon electron-impact.** *Org Mass Spectrom* 1969, **2**(5):481–519.
188. Bandu ML, Watkins KR, Bretthauer ML, Moore CA, Desaire H: **Prediction of MS/MS data. 1. A focus on pharmaceuticals containing carboxylic acids.** *Anal Chem* 2004, **76**(6):1746–1753.
189. Klagkou K, Pullen F, Harrison M, Organ A, Firth A, Langley GJ: **Approaches towards the automated interpretation and prediction of electrospray tandem mass spectra of non-peptidic combinatorial compounds.** *Rapid Commun Mass Spectrom* 2003, **17**(11):1163–1168.
190. Gray NAB, Carhart RE, Lavanchy A, Smith DH, Varkony T, Buchanan BG, White WC, Creary L: **Computerized mass spectrum prediction and ranking.** *Anal Chem* 1980, **52**(7):1095–1102.
191. Clark HA, Jurs PC: **Simulation of mass spectral intensities by regression analysis of calculated structural characteristics.** *Anal Chim Acta* 1981, **132:**75–88.
192. Chen H, Fan B, Xia H, Petitjean M, Yuan S, Panaye A, Doucet JP: **MASSIS: A mass spectrum simulation system 1. Principle and method.** *Eur J Mass Spectrom (Chichester, Eng)* 2003, **9**(3):175–186.
193. Chen H, Fan B, Petitjean M, Panaye A, Doucet JP, Li F, Xia H, Yuan S: **MASSIS: a mass spectrum simulation system. 2: Procedures and performance.** *Eur J Mass Spectrom (Chichester, Eng)* 2003, **9**(5):445–457.
194. Fan B, Chen H, Petitjean M, Panaye A, Doucet JP, Xia H, Yuan S: **New strategy of mass spectrum simulation based on reduced and concentrated knowledge databases.** *Spectrosc Lett* 2005, **38**(2):145–170.
195. Schymanski EL, Meringer M, Brack W: **Matching structures to mass spectra using fragmentation patterns: Are the results as good as they look?** *Anal Chem* 2009, **81**(9):3608–3617.
196. Kerber A, Laue R, Meringer M, Varmuza K: **MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation.** *Adv Mass Spectrom* 2001, **15:**939–940.
197. Kerber A, Meringer M, Rücker C: **CASE via MS: Ranking structure candidates by mass spectra.** *Croat Chem Acta* 2006, **79**(3):449–464.
198. Pelander A, Tyrkkö E, Ojanperä I: **In silico methods for predicting metabolism and mass fragmentation applied to quetiapine in liquid chromatography/time-of-flight mass spectrometry urine drug screening.** *Rapid Commun Mass Spectrom* 2009, **23**(4):506–514.
199. Kumari S, Stevens D, Kind T, Denkert C, Fiehn O: **Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry.** *Anal Chem* 2011, **83**(15):5895–5902.
200. Sweeney DL: **Small molecules as mathematical partitions.** *Anal Chem* 2003, **75**(20):5362–5373.
201. Hill AW, Mortishire-Smith RJ: **Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach.** *Rapid Commun Mass Spectrom* 2005, **19:**3111–3118.
202. Heinonen M, Rantanen A, Mielikäinen T, Pitkänen E, Kokkonen J, Rousu J: **Ab initio prediction of molecular fragments from tandem mass spectrometry data.** In *Proc. of German Conference on Bioinformatics (GCB 2006), volume P-83 of Lecture Notes in Informatics*:40–53.
203. Heinonen M, Rantanen A, Mielikäinen T, Kokkonen J, Kiuru J, Ketola RA, Rousu J: **FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data.** *Rapid Commun Mass Spectrom* 2008, **22**(19):3043–3052.
204. Böcker S, Rasche F, Steijger T: **Annotating fragmentation patterns.** In *Proc. of Workshop on Algorithms in Bioinformatics (WABI 2009), volume 5724 of Lect Notes Comput Sci*. Berlin: Springer; 2009:13–24.
205. Schymanski EL, Gallampois CMJ, 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**(7):3287–3295.
206. Gerlich M, Neumann S: **MetFusion: Integration of compound identification strategies.** *J Mass Spectrom* 2013, **48**(3):291–8.
207. Menikarachchi LC, Cawley S, Hill DW, Hall LM, Hall L, Lai S, Wilder J, Grant DF: **MolFind: A software package enabling HPLC/MS-based identification of unknown chemical structures.** *Anal Chem* 2012, **84**(21):9388–9394.
208. Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, van Schaik R, Vervoort J: **Substructure-based annotation of high-resolution multistage MS^n spectral trees.** *Rapid Commun Mass Spectrom* 2012, **26**(20):2461–2471.
209. Ludwig M, Hufsky F, Elshamy S, Böcker S: **Finding characteristic substructures for metabolite classes.** In *Proc. of German Conference on Bioinformatics (GCB 2012), volume 26 of OpenAccess Series in Informatics (OASIcs)*; 2012:23–38. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.
210. Bode HB, Müller R: **The impact of bacterial genomics on natural product research.** *Angew Chem Int Ed Engl* 2005, **44:**6828–6846.

211. Bandeira N, Ng J, Meluzzi D, Linington RG, Dorrestein P, Pevzner PA: **De novo sequencing of nonribosomal peptides.** In *Proc. of Research in Computational Molecular Biology (RECOMB 2008), volume 4955 of Lect Notes Bioinform*. Berlin: Springer; 2008:181–195.
212. Liu WT, Ng J, Meluzzi D, Bandeira N, Gutierrez M, Simmons TL, Schultz AW, Linington RG, Moore BS, Gerwick WH, Pevzner PA, Dorrestein PC: **Interpretation of tandem mass spectra obtained from cyclic nonribosomal peptides.** *Anal Chem* 2009, **81:**4200–4209.
213. Ng J, Bandeira N, Liu WT, Ghassemian M, Simmons TL, Gerwick WH, Linington R, Dorrestein PC, Pevzner PA: **Dereplication and de novo sequencing of nonribosomal peptides.** *Nat Methods* 2009, **6**(8):596–599.
214. Mohimani H, Yang YL, Liu WT, Hsieh PW, Dorrestein PC, Pevzner PA: **Sequencing cyclic peptides by multistage mass spectrometry.** *Proteomics* 2011, **11**(18):3642–3650.
215. Rojas-Chertó M, Peironcely JE, Kasper PT, van der Hooft JJJ, de Vos RCH, Vreeken RJ, Hankemeier T, Reijmers TH: **Metabolite identification using automated comparison of high-resolution multistage mass spectral trees.** *Anal Chem* 2012, **84**(13):5524–5534.
216. Kasper PT, Rojas-Chertó M, Mistrik R, Reijmers T, Hankemeier T, Vreeken RJ: **Fragmentation trees for the structural characterisation of metabolites.** *Rapid Commun Mass Spectrom* 2012, **26**(19):2275–2286.
217. Rauf I, Rasche F, Nicolas F, Böcker S: **Finding maximum colorful subtrees in practice.** In *Proc. of Research in Computational Molecular Biology (RECOMB 2012), volume 7262 of Lect Notes Comput Sci*. Berlin: Springer; 2012:213–223.
218. Hufsky F, Böcker S: **Comparing fragmentation trees from electron impact mass spectra with annotated fragmentation pathways.** In *Proc. of German Conference on Bioinformatics (GCB 2012), volume 26 of OpenAccess Series in Informatics (OASIcs)*; 2012:12–22. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
219. Scheubert K, Hufsky F, Rasche F, Böcker S: **Computing fragmentation trees from metabolite multiple mass spectrometry data.** In *Proc. of Research in Computational Molecular Biology (RECOMB 2011), volume 6577 of Lect Notes Comput Sci*. Berlin: Springer; 2011:377–391.
220. Scheubert K, Hufsky F, Rasche F, Böcker S: **Computing fragmentation trees from metabolite multiple mass spectrometry data.** *J Comput Biol* 2011, **18**(11):1383–1397.
221. Rasche F, Scheubert K, Hufsky F, Zichner T, Kai M, Svatoš A, Böcker S: **Identifying the unknowns by aligning fragmentation trees.** *Anal Chem* 2012, **84**(7):3417–3426.
222. Hufsky F, Dührkop K, Rasche F, Chimani M, Böcker S: **Fast alignment of fragmentation trees.** *Bioinformatics* 2012, **28:**i265—i273.
223. Rarey M, Dixon JS: **Feature trees: A new molecular similarity measure based on tree matching.** *J Comput Aided Mol Des* 1998, **12**(5):471–490.
224. Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L: **Metabolite profiling for plant functional genomics.** *Nat Biotechnol* 2000, **18**(11):1157–1161.
225. Arkin A, Shen P, Ross J: **A test case of correlation metric construction of a reaction pathway from measurements.** *Science* 1997, **277**(5330):1275–1279.
226. Kose F, Weckwerth W, Linke T, Fiehn O: **Visualizing plant metabolomic correlation networks using clique-metabolite matrices.** *Bioinformatics* 2001, **17**(12):1198–1208.
227. Steuer R, Kurths J, Fiehn O, Weckwerth W: **Observing and interpreting correlations in metabolomic networks.** *Bioinformatics* 2003, **19**(8):1019–1026.
228. Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ: **Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data.** *BMC Syst Biol* 2011, **5:**21.
229. Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP: **Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data.** *Metabolomics* 2006, **2**(3):155–164.
230. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, van der Voort M, Pogliano K, Gross H, Raaijmakers JM, Moore BS, Laskin J, Bandeira N, Dorrestein PC: **Mass spectral molecular networking of living microbial colonies.** *Proc Natl Acad Sci USA* 2012, **109**(26):E1743—E1752.
231. Stein SE: **An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data.** *J Am Soc Mass Spectrom* 1999, **10**(8):770–781.
232. Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, Robert M, Tomita M: **MathDAMP: A package for differential analysis of metabolite profiles.** *BMC Bioinformatics* 2006, **7:**530.
233. Luedemann A, Strassburg K, Erban A, Kopka J: **TagFinder for the quantitative analysis of gas chromatography–mass spectrometry (GC-MS)-based metabolite profiling experiments.** *Bioinformatics* 2008, **24**(5):732–737.
234. Luedemann A, von Malotky L, Erban A, Kopka J: **TagFinder: Preprocessing software for the fingerprinting and the profiling of gas chromatography-mass spectrometry based metabolome analyses.** *Methods Mol Biol* 2012, **860:**255–286.
235. Hiller K, Hangebrauk J, Jäger C, Spura J, Schreiber K, Schomburg D: **MetaboliteDetector: Comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis.** *Anal Chem* 2009, **81**(9):3429–3439.
236. Cuadros-Inostroza A, Caldana C, Redestig H, Kusano M, Lisec J, Peña-Cortés H, Willmitzer L, Hannah MA: **TargetSearch–a Bioconductor package for the efficient preprocessing of GC-MS metabolite profiling data.** *BMC Bioinformatics* 2009, **10:**428.
237. Aggio R, Villas-Bôas SG, Ruggiero K: **Metab: An R package for high-throughput analysis of metabolomics data generated by GC-MS.** *Bioinformatics* 2011, **27**(16):2316–2318.
238. O'Callaghan S, Desouza DP, Isaac A, Wang Q, Hodkinson L, Olshansky M, Erwin T, Appelbe B, Tull DL, Roessner U, Bacic A, McConville MJ, Likic VA: **PyMS: A Python toolkit for processing of gas chromatography–mass spectrometry (GC-MS) data. Application and comparative study of selected tools.** *BMC Bioinformatics* 2012, **13**(1):115.
239. Ni Y, Qiu Y, Jiang W, Suttlemyre K, Su M, Zhang W, Jia W, Du X: **ADAP-GC 2.0: Deconvolution of coeluting metabolites from GC/TOF-MS data for metabolomics studies.** *Anal Chem* 2012, **84**(15):6619–6629.
240. Castillo S, Mattila I, Miettinen J, Orešič M, Hyötyläinen T: **Data analysis tool for comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry.** *Anal Chem* 2011, **83**(8):3058–3067.
241. Benton HP, Wong DM, Trauger SA, Siuzdak G: **XCMS2: Processing tandem mass spectrometry data for metabolite identification and structural characterization.** *Anal Chem* 2008, **80**(16):6382–6389.
242. Benton HP, Want EJ, Ebbels TMD: **Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data.** *Bioinformatics* 2010, **26**(19):2488–2489.
243. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G: **XCMS Online: A web-based platform to process untargeted metabolomic datas.** *Anal Chem* 2012, **84**(11):5035–5039.
244. Alonso A, Julià A, Beltran A, Vinaixa M, Díaz M, Ibañez L, Correig X, Marsal S: **AStream: An R package for annotating LC/MS metabolomic data.** *Bioinformatics* 2011, **27**(9):1339–1340.
245. Wei X, Sun W, Shi X, Koo I, Wang B, Zhang J, Yin X, Tang Y, Bogdanov B, Kim S, Zhou Z, McClain C, Zhang X: **MetSign: A computational platform for high-resolution mass spectrometry-based metabolomics.** *Anal Chem* 2011, **83**(20):7668–7675.
246. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S: **CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets.** *Anal Chem* 2012, **84**(1):283–289.
247. Bueschl C, Kluger B, Berthiller F, Lirk G, Winkler S, Krska R, Schuhmacher R: **MetExtract: A new software tool for the automated comprehensive extraction of metabolite-derived LC/MS signals in metabolomics research.** *Bioinformatics* 2012, **28**(5):736–738.
248. Creek DJ, Jankevics A, Burgess KEV, Breitling R, Barrett MP: **IDEOM: An Excel interface for analysis of LC-MS based metabolomics data.** *Bioinformatics* 2012.
249. Scheltema RA, Jankevics A, Jansen RC, Swertz MA, Breitling R: **PeakML/mzMatch: A file format, Java library, R library, and tool-chain for mass spectrometry data analysis.** *Anal Chem* 2011, **83**(7):2786–2793.
250. Brown M, Wedge DC, Goodacre R, Kell DB, Baker PN, Kenny LC, Mamas MA, Neyses L, Dunn WB: **Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets.** *Bioinformatics* 2011, **27**(8):1108–1112.

251. Brodsky L, Moussaieff A, Shahaf N, Aharoni A, Rogachev I: **Evaluation of peak picking quality in LC-MS metabolomics data.** *Anal Chem* 2010, **82**(22):9177–9187.
252. Katajamaa M, Miettinen J, Oresic M: **MZmine: Toolbox for processing and visualization of mass spectrometry based molecular profile data.** *Bioinformatics* 2006, **22**(5):634–636.
253. Pluskal T, Castillo S, Villar-Briones A, Oresic M: **MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data.** *BMC Bioinformatics* 2010, **11:**395.
254. Broeckling CD, Reddy IR, Duran AL, Zhao X, Sumner LW: **MET-IDEA: Data extraction tool for mass spectrometry-based metabolomics.** *Anal Chem* 2006, **78**(13):4334–4341.
255. Lommen A: **MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing.** *Anal Chem* 2009, **81**(8):3079–3086.
256. Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt WF, Schenck RJ, Trippe AJ: **Structural diversity of organic chemistry: A scaffold analysis of the CAS registry.** *J Org Chem* 2008, **73**(12):4443–4451.

doi:10.1186/1758-2946-5-12
Cite this article as: Scheubert *et al.*: Computational mass spectrometry for small molecules. *Journal of Cheminformatics* 2013 **5**:12.