New kids on the block: Novel informatics methods for natural product discovery

Franziska Hufsky, Kerstin Scheubert, and Sebastian Böcker

Lehrstuhl für Bioinformatik, Friedrich-Schiller-Universität Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany, sebastian.boecker@uni-jena.de

This is a preprint of the article: Franziska Hufsky, Kerstin Scheubert and Sebastian Böcker. New kids on the block: Novel informatics methods for natural product discovery. Nat Prod Rep, 31(6):807-817, 2014.

Abstract. Mass spectrometry is a key technology for the identification and structural elucidation of natural products. Manual interpretation of the resulting data is tedious and time-consuming, so methods for automated analysis are highly sought after. In this review, we focus on four recently developed methods for the detection and investigation of small molecules, namely *MetFrag/MetFusion*, *ISIS*, *FingerID*, and *FT-BLAST*. These methods have the potential to significantly advance the field of computational mass spectrometry for the research of natural products. For example, they may help with the dereplication of compounds at an early stage of the drug discovery process; that is, the detection of molecules that are identical or highly similar to known drugs or drug leads. Furthermore, when a potential drug lead has been determined, these tools may help to identify it and elucidate its structure.

1 Introduction

Metabolomics complements investigation of the genome, transcriptome, and proteome of an organism [32]. Today, the vast majority of metabolites remain unknown; this is particularly the case for non-model organisms and secondary metabolites [2]. The structural diversity of metabolites is extraordinarily large, and much larger than for biopolymers such as proteins. In almost all cases, we cannot deduce the structure of metabolites from genome sequences, as it is done with proteins. Secondary metabolites from plants and microorganisms have a long tradition of being used as therapeutics in medicine. Certain microbial producers, for example the actinomycetes, have been extensively screened over decades and thus chances for the discovery of new drug leads from these organisms are regarded as decreasing. Consequently, alternative sources such as marine bacteria, gliding bacteria and terrestrial microorganisms from unusual ecological niches have come into focus.

Mass spectrometry (MS) is one of the two predominant experimental analysis techniques for detecting and identifying metabolites and other small molecules, the other being nuclear magnetic resonance (NMR). The most important advantage of MS over NMR is that it is orders of magnitude more sensitive, making it the method of choice for first-pass compound detection and identification in medium- to high-throughput screening applications. MS can be coupled with a chromatography separation method to analyze complex mixtures such as cell extracts. Single-stage MS cannot provide information beyond the compound mass; to obtain such information, the analyte is fragmented, and fragment masses are recorded. Typical fragmentation methods include collision-induced dissociation (CID) for tandem MS, and fragmentation during electron ionization (EI).

After some initial progress as part of the DENDRAL project [23], the development of computational methods for analyzing fragmentation patterns of small molecules proceeded only slowly. Today, data analysis is still presumed to be a major bottleneck in metabolomics [28]. Moreover, spectral libraries of reference compounds are vastly incomplete, and this situation is unlikely to change. Three recent approaches (see Table 1) seek to replace searching in spectral libraries by searching in molecular structure databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes) and PubChem, which are and will be more comprehensive than spectral libraries. *MetFusion* [7] combines search results in a molecular structure database and results from searching a spectral library, taking advantage of both resources. *ISIS* [19] generates an *in silico* spectral library from a molecular structure database, using machine learning methods for the prediction of bond cleavage energies. *FingerID* [11] uses machine learning to predict a Tanimoto-style structural fingerprint of the unknown compound, which is then used to search a molecular structure database. A fourth approach, *FT-BLAST* [34], replaces searching in a spectral library by searching in a fragmentation tree library (see Table 1). This allows not only to search for identical compounds but also for chemically similar compounds, and discriminate between true and spurious hits using false discovery rate estimation. Clearly, there exist computational methods beyond those covered here, see the MS/MS network approach [29, 45] for an example; in addition, the presented methods can be seen as a complement to genome-guided metabolite discovery [25, 26]. See Kind and Fiehn [21], Scheubert *et al.* [37] for more comprehensive reviews of the available literature, Stein [41] for searching small molecule MS libraries, and Hufsky *et al.* [17] for the basics of computational approaches dealing with metabolite fragmentation data.

Table 1. Overview on the four approaches described in this review. *FT-BLAST requires a fragmentation tree database which can be created either by hand drawn fragmentation diagrams or by fragmentation tree computation on a spectral library.

	MetFrag [46]/ **MetFusion** [7]	**ISIS** [19]	**FingerID** [11]	**FT-BLAST** [34]
method	combinatorial fragmentation	spectrum prediction	predicting structural properties	fragmentation trees
computational technique	combinatorial optimization	machine learning	machine learning	combinatorial optimization
Molecular structure database required?	yes	yes	yes	no
Spectral database required?	no/yes	no	no	yes*
Discovering unknown unknowns?	no	no	predicts fingerprint of the unknown compound	if similar to known compound

2 Combinatorial Fragmentation: MetFrag and MetFusion

The most common approach for identifying unknown compounds using mass spectrometry is to search for similar fragmentation spectra in a library [31, 42]. Unfortunately, metabolite spectral databases are rather incomplete, so searching in these libraries is often unsuccessful. For CID fragmentation, an additional difficulty is that fragmentation can vary significantly on different instruments [30]. In comparison, molecular structure databases are orders of magnitude larger than spectral libraries. Therefore, it is potentially more powerful to search in molecular structure databases than in spectral libraries. Combinatorial fragmentation attempts to explain the peaks in a query spectrum using sub-molecules of a candidate molecular structure generated by disconnecting bonds [10, 12]. Here, we focus on the method underlying *MetFrag* [46] and *MetFusion* [7]. Whereas *MetFrag* is a “pure” combinatorial fragmenter, *MetFusion* combines *MetFrag* with a “classical” search in spectral libraries.

MetFrag was developed in 2010 by Wolf *et al.* [46]. The input is a measured fragmentation spectrum of an unknown compound, plus a molecular structure database such as KEGG, PubChem, or ChemSpider. We assume that the molecular structure database contains the unknown compound, so only compounds with mass close to the precursor ion mass, or with a particular molecular formula are considered for further analysis. We test each of the remaining entries as a candidate molecular structure of the unknown compound. *MetFrag* simulates the fragmentation by cutting molecular bonds, trying to explain the peaks in the query fragmentation spectrum, see Fig. 1. To find fragments that can explain the peaks, a tree search algorithm is applied: The root of the tree is the intact molecule, edges represent disconnection of one particular bond, and nodes in the tree correspond to the resulting fragments. To process a full molecular structure database, swift running times for a single candidate molecular structure are essential. At the same time, the problem of generating all molecular substructures is computationally challenging. To this end, *MetFrag* uses three heuristics to avoid combinatorial explosion: the number of cleavages allowed is limited by setting a maximum search tree depth; redundant fragments are removed; and finally, only fragments with a mass higher than the lowest mass of any fragment peaks are considered.

To take into account that some bonds break more easily than others, we have to assign costs for cleaving any bond. *MetFrag* uses bond dissociation energies for this purpose. The cost of cutting out a fragment from a molecular structure, is the sum of bond dissociation energies of the disconnected bonds. We assure that each peak in the fragmentation spectrum is explained by the fragment with minimum cost. The score of a candidate molecular structure is a combination of a simple peak counting score and the total bond dissociation energy for generating the corresponding fragments: The higher the bond dissociation energies, the lower the score.

Wolf *et al.* [46] applied *MetFrag* to a dataset of 102 molecules measured on a Micromass Q-TOF II. *MetFrag* retrieved an average of 2508 candidate compounds per molecule from PubChem and returned the correct molecular structure with a median rank of 31. Wolf *et al.* [46] state that *MetFrag* requires a mass accuracy of at least 10 ppm.

There are certain disadvantages for methods based on combinatorial fragmentation: First, fragments resulting from structural rearrangements such as hydrogen relocation can be accounted for only to a low extend, in order to avoid an otherwise inevitable running time explosion. Second, Wolf *et al.* [46] found that the prediction accuracy of *MetFrag* decreases with increasing maximum tree depth. Best results were obtained for maximum tree depth two, meaning that only fragments which are disconnected at one or two bonds, are simulated. This is somewhat unexpected, as one might argue that allowing more bond cleavages should improve results, since more peaks can be explained. Unfortunately, increasing the number of simulated fragmentation steps also results in an increase of “bogus fragments” which can be assigned to some peak in the query spectrum solely by chance. Another point of criticism has been pointed out by Ridder *et al.* [35]. They found that even some simplistic scoring which directly uses bond orders $1, 2, 3$ as the cost of disconnecting some bond, outperforms the more involved cost function of Wolf *et al.* [46]. This indicates that finding a good cost function remains an open problem.

MetFusion [7] is the successor of *MetFrag*, and combines combinatorial fragmentation with searching in spectral libraries. *MetFusion* aims at improving the rank of the correct molecular structure in a list of candidate structures (Figure 1). First, *MetFrag* is used to retrieve and score a list of candidate structures from a molecular structure database. Second, we calculate the spectral similarity between the query spectrum and a set of reference spectra from a spectral library, using cosine distances. These scores are combined with the chemical (i.e., structural) similarity between the candidate structures and the corresponding structures of the reference compounds. The idea is that, for the correct compound, high spectral similarity scores should correlate with high chemical similarity. Accordingly, candidate structures that have high chemical similarity as well as high spectral similarity to compounds in the spectral library are favored. In detail, for each reference

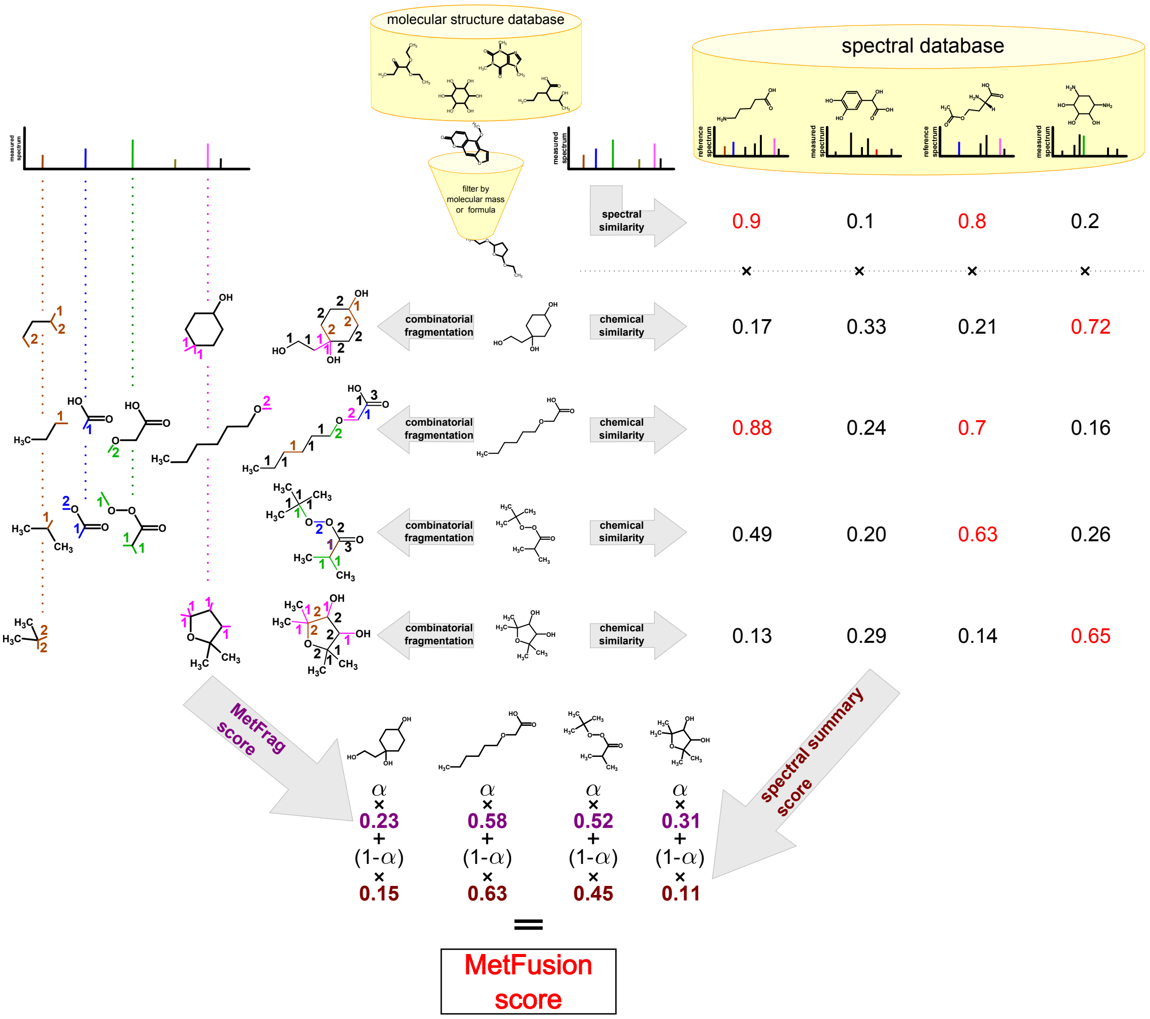


Fig. 1. *MetFusion* is a combination of combinatorial fragmentation by *MetFrag* (left) and searching in a spectral library (right). First, candidate structures from a structural database are filtered by precursor mass or molecular formula. Second, scores are assigned to the remaining candidate structures. These scores take into account how well the peaks in the query spectrum can be explained by combinatorial fragmentation (left), as well as how well the chemical similarity to a reference molecular structure matches the spectral similarity of the corresponding reference spectrum (right). High chemical and spectral similarities are highlighted red. The final score is the sum of these two scores, weighted by a coefficient $\alpha \in [0, 1]$.

compound in the library, we multiply the chemical similarity to the candidate structure and the spectral similarity to the query spectrum, and sum over all reference compounds. The chemical similarity between structures is measured using the Jaccard coefficient. Finally, the *MetFusion* score is the weighted sum of the original *MetFrag* score and the “spectral summary” score.

Gerlich and Neumann [7] applied *MetFusion* to a dataset of 1062 spectra from MassBank, and compared the results of *MetFrag* and *MetFusion.* The spectral library with 5063 compounds was also taken from MassBank; PubChem was used as the molecular structure database. *MetFrag* returns the correct compound with median rank 28, whereas *MetFusion* decreases the median rank to 4. The authors noticed that their evaluation is somewhat biased, as spectral library and test dataset are overlapping and contain many compound pairs with “unusually high” chemical similarity. For a more realistic evaluation, parts of the spectral library with high similarity to the query compound have to be excluded. If compounds with chemical similarity 0.9 or above are excluded from the spectral library, the median rank increases to 7; for chemical similarity above 0.7 this further increases to median rank 10. Results clearly indicate that spectral summary scores improve *MetFrag* results. But results also indicate that performance depends on the comprehensiveness of the spectral library. In cases where the spectral database contains no or only few compounds that are similar to the unknown query compound, the improvement is much less pronounced than for those cases where many similar compounds can be found in the spectral library.

Whereas first ideas related to combinatorial fragmentation can be found in the literature as early as 1980 [8], it took another 30 years until combinatorial fragmentation was made suitable for processing a complete database. In evaluation, *MetFrag* reaches somewhat promising results but is clearly outperformed by its successor *MetFusion.*

3 Predicting fragmentation spectra: ISIS

A rule-based fragmenter simulates a hypothetical fragmentation spectrum for a molecular structure based on manually curated or automatically learned fragmentation rules [5, 6]. The resulting hypothetical fragmentation spectrum can then be compared to the query fragmentation spectrum for searching in a molecular structure database [13]. Unfortunately, the results of *in silico* mass spectra prediction using rule-based fragmentation are in many cases not satisfactory [38]. For many rule-based fragmenters, all predicted peaks have the same intensity as bond cleavage rates are not considered, an example being *Mass Frontier* [19]. Accurate peak intensities (or ion intensities) can, however, greatly improve identification accuracy. The *In silico identification software* (*ISIS*) presented by Kangas *et al.* [19] generates *in silico* spectra by predicting bond cleavage energies through machine learning; doing so, cleavage rates and, hence, peak intensities can be estimated (Figure 2). Note that Kangas *et al.* [19] do not claim these predictions to be “true” fragmentation rules, different from, say, the rules learned during the DENDRAL project, see for example [9].

ISIS predicts fragmentation spectra by simulating the behavior of ions in a linear ion trap using a kinetic Monte Carlo (KMC) simulation. First, the model linear ion trap is filled with replicates of the same ion with internal energies stochastically sampled from an ionization model. Then, an ion is selected for collision based on its cross-section and velocity. The selected ion collides with an ion of the inert gas. Collision energy is thereby transmitted to the model ion. Now, we want to know whether this energy is sufficient to cause a bond to break. To answer this question, an Artificial Neural Network (ANN) predicts bond cleavage energies for all bonds in the molecule. As each bond cleavage is

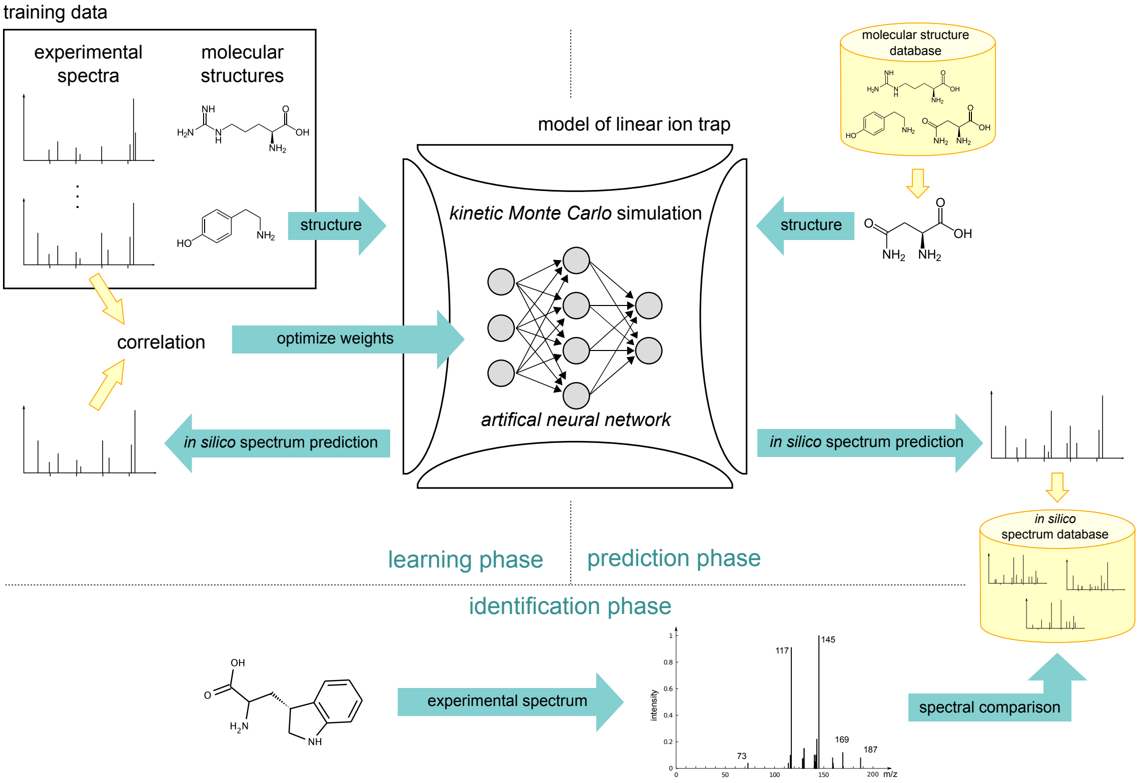


Fig. 2. Predicting fragmentation spectra using the *In silico identification software* (*ISIS*) [19]. A linear ion trap is modeled using kinetic Monte Carlo (KMC) simulation. *Learning phase:* An Artificial Neural Network (ANN) incrementally learns the bond cleavage energies for all bonds in the molecule using a set of training molecules. For each molecule, an *in silico* spectrum is predicted and compared to the corresponding experimental spectrum. The resulting correlation is used to optimize the weights in the ANN to improve prediction. *Prediction phase:* For each molecule in a molecular database an *in silico* spectrum is predicted. *Identification phase:* The experimental spectrum of an unknown compound is matched against the *in silico* database.

assumed to be an independent event, either no bond or exactly one bond is cleaved. Note that rearrangements of atoms or bonds are not covered by this model. If a bond cleavage leads to fragmentation, a charge prediction model labels one fragment as the ion and the other one as neutral. In case the resulting fragment has mass below some threshold, it looses its velocity and is not considered for further fragmentation. After each fragmentation event, properties of all fragments have to be recomputed to predict new bond cleavage rates. Finally, a cooling schedule decreases the internal energies of ions, leading the fragmentation process to an end.

For the prediction of bond cleavage energies, an ANN needs to undergo a learning phase using molecules and their experimental spectra. For each molecule, an *in silico* spectrum is predicted using the KMC simulation described above. This *in silico* spectrum is compared to the experimental spectrum. The resulting correlation is used to optimize the weights of the ANN to predict refined bond cleavage energies, which in turn improve the prediction of the *in silico* spectra. This iterative process is repeated: The *ISIS* learning phase required four months of computation [19].

Kangas *et al.* [19] trained and evaluated *ISIS* exclusively for lipids. The authors optimized ANN weights using mass spectral data of 22 lipids. After four months of training, the Pearson correlation coefficient between the predicted spectra and experimental spectra reached $r^2 = 0.97$ and the training algorithm was stopped. After the learning phase, the software was used to build an *in silico* spectral test database of about 18 000 lipids. KMC simulation of fragmentation spectra required about one minute per spectrum. For 45 lipids from the test database, an experimental spectrum was measured to search against the *in silico* spectra. This screening resulted in the correct identification of 40 out of 45 lipids, while for the remaining five lipids, the correct answer was ranked second. In comparison, *MetFrag* (see Section 2) ranked only 21 lipids in the top position, 8 in the second position, and the remaining ones in the third and fourth positions [19].

Kangas *et al.* [19] have, for the first time, simulated fragmentation spectra using machine learning methods. As this approach is still relatively young, it is difficult to predict its progress and success in the years to come. Kangas *et al.* mention two important additions they are planning to integrate into simulations: modeling rearrangements, which are especially important for certain fragmentation techniques such as electron ionization; and, processing of different (not only hydrogen) adducts. It is also very likely that the *ISIS* method will require profound changes and extensive testing to be applicable for other molecule classes than lipids. Nevertheless, this approach appears to be promising for small molecule identification.

A more recent approach by Allen *et al.* [1], called *Competitive Fragmentation Modeling*, predicts spectra using a probabilistic generative model for the MS/MS fragmentation process and is not limited to lipids.

4 Fingerprint Prediction: FingerID

There are different ways to apply machine learning techniques to mass spectral data. Rather than predicting fragmentation mass spectra from molecular structures, Heinonen *et al.* [11] use machine learning “the other way around”: Namely, they predict structural properties (*fingerprints*) of the unknown compound from the mass spectrum. In a second step, they use the predicted fingerprints to search the unknown compound in a molecular structure database.

To learn and predict structural properties from mass spectral data, spectra need to be transformed to a set of numerical *features* characterizing the data. This idea was pioneered by Venkataraghavan *et al.* [44] in 1969. It is understood that an appropriate transformation of the mass spectral data into informative features, is essential for a good prediction of structural properties [43]. Heinonen *et al.* [11] use three classes of features extracted from the mass spectra: peaks, primary fragmentation reactions, and secondary fragmentation reactions, see Figure 3. All spectra are transformed to feature vectors in the same manner.

The prediction of structural fingerprints is learned using Support Vector Machines (SVM). For a training set of known molecular structures, the respective mass spectra are transformed to feature vectors as described above. Each feature vector is an object (data point) in a high-dimensional vector space, see Figure 3. For each structural fingerprint, feature vectors are marked as belonging to one of two categories: those possessing this property and those not possessing it. By optimization, an SVM finds a hyperplane in the vector space that separates these two classes. New feature vectors of unknown compounds are classified according to which side of the hyperplane they fall.

Structural fingerprints could be directly used to characterize the class and properties of unknown metabolites; note, though, that the accuracy of the predictors presented in [11] is most likely insufficient for such a manual analysis. To this end, Heinonen *et al.* [11] use the predicted fingerprint vectors to retrieve and score candidate molecules from a

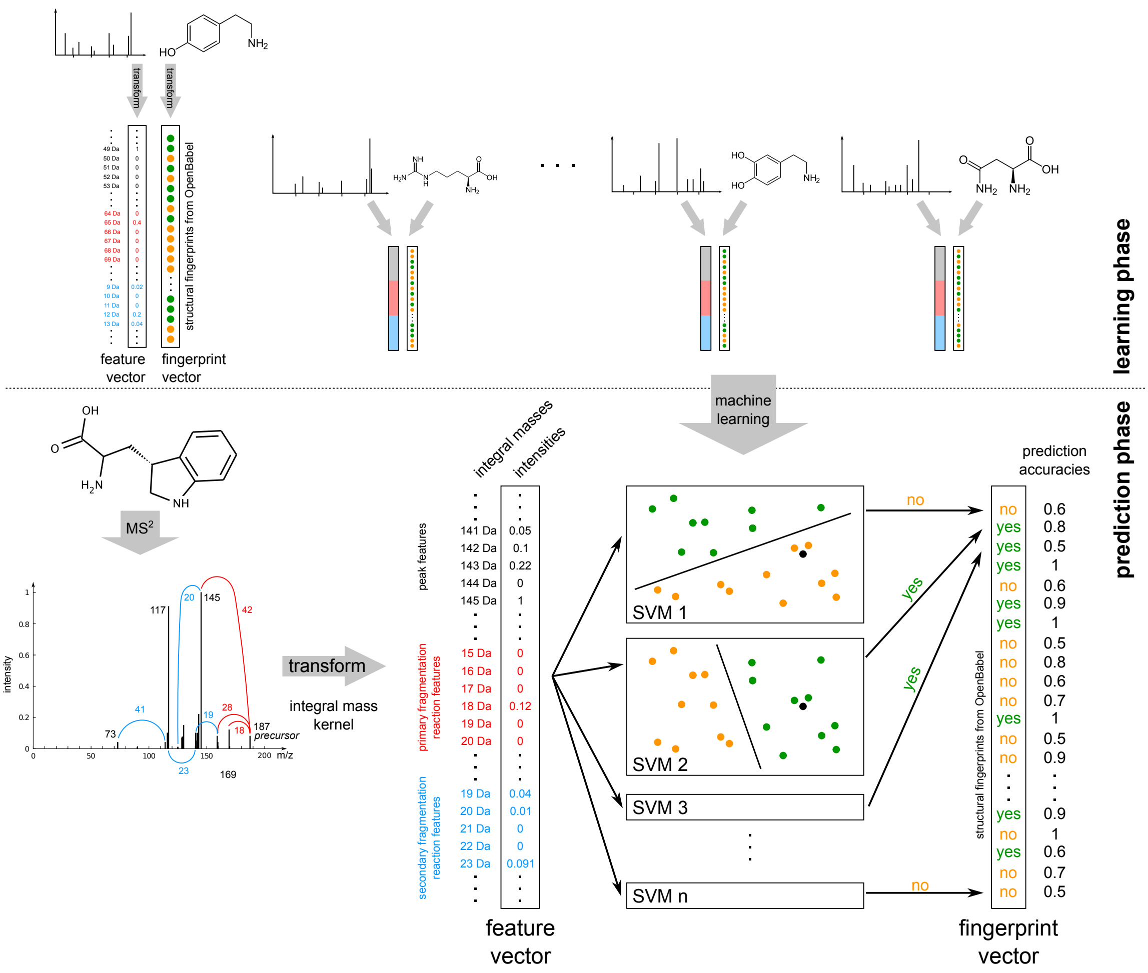


Fig. 3. Predicting fingerprints for the identification of unknown compounds. A spectrum is transformed to a feature vector using three classes of features: peaks (black), primary fragmentation reactions (red) and secondary fragmentation reactions (blue). In the training phase, each spectrum in the training dataset is transformed to a feature vector, which reside in the high-dimensional vector space of all possible features. For each structural fingerprint, feature vectors are marked as possessing this property (green) or not possessing it (orange). A Support Vector Machine is optimized on this training data to separate the vector space into these two classes. The unknown object (black) is classified based on which side of the hyperplane it falls on.

large molecular structure database such as KEGG and PubChem [11]. The structural fingerprint vectors of the database entries can be directly assessed from their molecular structure, without the need of prediction. The query spectrum of an unknown compound is transformed to a feature vector in the same way as for the training data. The structural fingerprints of the unknown compound are then predicted using the SVMs. Reliability of predictions varies for different fingerprints. For ranking candidate molecules, Heinonen *et al.* [11] use a vector of prediction accuracies (obtained, e.g., from cross-validation experiments) to give more weight to more reliably-predicted fingerprints.

Heinonen *et al.* [11] used two different datasets for training and evaluating their method: nominal mass accuracy *QqQ* data and ultra-high mass accuracy *LTQ* data. In fact, processing this data is achieved using two different kernels, we leave out the technical details. They used the predicted fingerprints to search in the KEGG database. For the *QqQ* data, *FingerID* ranked the correct molecular structure within the top 10 for 85 % of the 514 metabolites; on average, there were 25 candidate molecular structures to rank. For the more accurate *LTQ* data, the method ranked the correct molecular structure within the top 10 for 92 % of the 293 metabolites, from an average of 27 candidates. Heinonen *et al.* also compared *FingerID* to *MetFrag* (see Section 2) on a randomly selected subset of the data, and found *FingerID* to obtain better results than *MetFrag* in most cases [11].

Automated prediction of structural properties from mass spectral data using machine learning methods has been studied for several years, but solely for EI fragmentation data [22, 39, 43, 44]. The approach of Varmuza and Werther [43] has become widely used in the community, as it has been included in the NIST software. For tandem MS data, machine learning approaches have not been as widely studied. Heinonen *et al.* [11] showed that it is possible to learn structural properties from tandem mass spectral data, and that these properties can be used to score candidate molecules for compound identification. The performance of these identifications depends on both an adequate transformation of the spectral data to feature vectors, and the selection of a "good" set of fingerprints. Heinonen *et al.* [11] report that, for example, including secondary fragmentation reactions in the feature vector seem not to affect identification accuracy. For the selection of "good" fingerprints, a balance between the uniqueness of fingerprints for particular sets of metabolites, and ability to predict these fingerprints needs to be found.

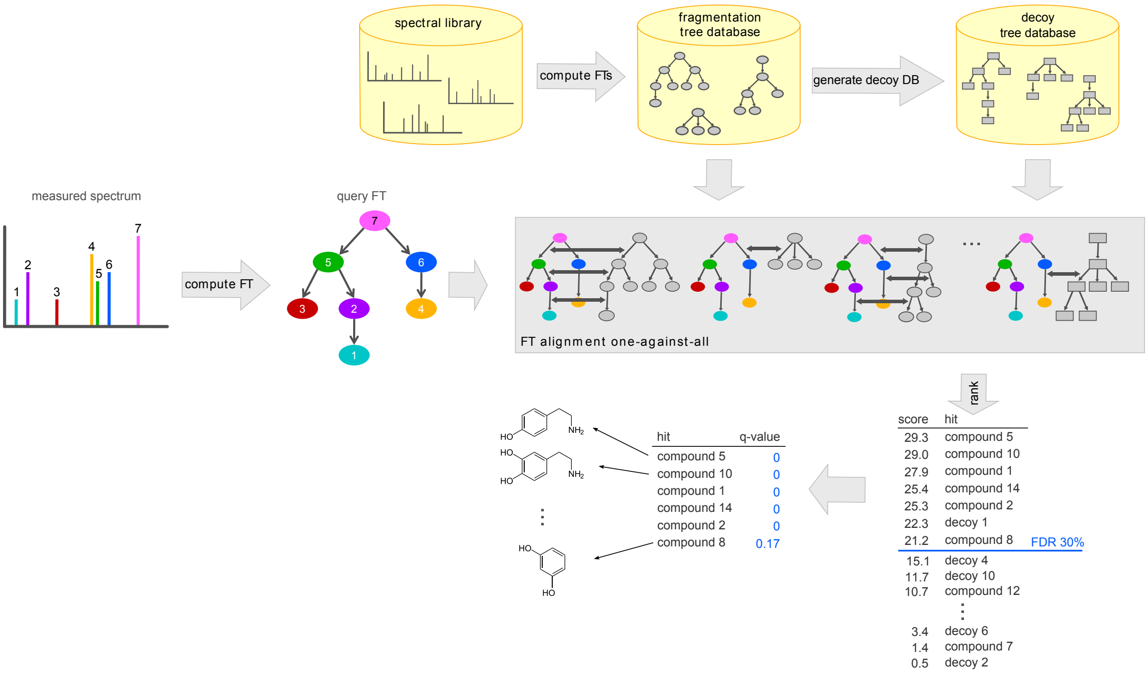
More recently, Shen *et al.* [40] integrated fragmentation tree computations into kernel-based machine learning to predict molecular fingerprints and identify molecular structures and showed that the new method significantly improves molecular fingerprint prediction accuracy.

5 Fragmentation Trees: FT-BLAST

The classical approach for identifying compounds is to look up spectra in a spectral library [42]. We mentioned above that the coverage of metabolites in spectral libraries is limited and, for CID fragmentation, spectra are often not reproducible on different instruments [30]. Certain approaches attempt not to find only the true compound, which is assumed to be absent from the database, but also compounds that have high chemical similarity to the unknown query [4, 31]. In this case, the result of querying the database with a single spectrum is a sorted list of compounds and corresponding scores; the goal is that compounds chemically similar to the query are listed first.

One open problem of this approach is how to estimate statistical significance: Using the above scores, it is highly non-trivial to decide whether hits are true or spurious. For peptide tandem mass spectra, there exist several methods for the statistical evaluation of hits, such as False Discovery Rates (FDR) based on a target-decoy approach [18], or Markov chains to directly compute the probability of a single peptide spectrum match [27]. A major challenge of target-decoy approaches is the generation of a "good" decoy database: The decoy spectra have to be similar, but not too similar to real spectra. Peptides are linear chains of amino acids, and it turns out that a good decoy database can be generated by simply reversing peptides (except for the last amino acid). In contrast to peptides, small molecules show more diverse structures and it remains an open issue how to produce good decoy databases [41]. It should be, however, noted that target-decoy approaches are best suited for statistical analysis of large spectral data sets, not for processing a single compound at a time.

The *Fragmentation Tree Basic Local Alignment Search Tool (FT-BLAST)* [34] (Figure 4) is based on the calculation of fragmentation trees (FTs) and FT alignments. Instead of



score	hit
29.3	compound 5
29.0	compound 10
27.9	compound 1
25.4	compound 14
25.3	compound 2
22.3	decoy 1
21.2	compound 8
15.1	decoy 4
11.7	decoy 10
10.7	compound 12
	⋮
3.4	decoy 6
1.4	compound 7
0.5	decoy 2

hit	q-value
compound 5	0
compound 10	0
compound 1	0
compound 14	0
compound 2	0
compound 8	0.17

Fig. 4. FT library search with *FT-BLAST*. An FT is constructed for a query spectrum, then aligned to all FTs of the target and decoy databases. The combined result list is sorted by alignment score. FDRs are calculated and the lowest score with an FDR below 30% is used as a threshold. All hits with a higher score than this threshold are reported in the result list. Corresponding q-values are calculated as the smallest FDR value for which some hit would be reported. The reported compounds show high chemical similarity to the query compounds.

searching in a spectral library, *FT-BLAST* searches in a fragmentation tree library. One direct advantage of this approach is that it is easy to generate a decoy FT database and, hence, to determine FDRs for hits in the target database. FTs were introduced in 2008 by Böcker and Rasche, and model the fragmentation cascades during tandem MS. The nodes of an FT are labeled with the molecular formulas of the compound and its fragments, and edges correspond to losses. Given the fragmentation spectrum of an unknown compound, an FT is determined by an optimization algorithm that searches for the FT that best explains the data [3]. Related algorithms were developed for multiple MS [36] and EI fragmentation [16]. Although FTs do not necessarily reflect the true fragmentation processes, but rather provide a fragmentation pattern for further analysis, they are in close agreement with MS expert interpretation of the fragmentation spectra [14, 33].

The basic idea behind *FT-BLAST* is that fragmentation pattern similarities are correlated with the chemical similarity of the corresponding compounds. Thus, searching in a spectral library for a query spectrum can be replaced by searching in an FT library for the corresponding query FT calculated from the query spectrum. Similarity between FTs is determined by local tree alignments [34]: The local tree alignment of two trees contains those parts where similar fragmentation cascades occur. The problem of finding the best local tree alignment is computationally hard, but using efficient algorithms, it can be solved for most instances in a matter of milliseconds [15, 34].

One innovation of *FT-BLAST* is the possibility to calculate FDRs based on the target-decoy approach. FDRs are designed to control the expected proportion of false positive hits. The decoy database of *FT-BLAST* is generated as follows: for each FT in the target database, an FT with the same number of edges and the same compound molecular formula is constructed for the decoy database. FTs of an independent dataset serve as scaffolds for these decoy trees. Losses in the decoy FT are randomly drawn from losses in the target database, respecting the multiplicity of losses. The query FT of the unknown compound is aligned to each FT in the combined database that contains target as well as decoy FTs. Hits (pairs of query and reference FT) are sorted by score. For any score threshold, we can estimate the number of false hits in the target database; for a given FDR, we report the largest list of hits such that the estimated ratio of false hits is below the given FDR threshold. Rasche *et al.* [34] specified an FDR of 30 %, thus obtaining a list of reliable hits where only 30 % of the hits are "probably wrong". Note that there is no direct definition of a "true hit" or "false hit" when searching for structurally *similar* compounds, see below. Different from the field of proteomics, where an FDR of 1 %-5 % is commonly used to identify a compound, searching for similar rather than identical compounds means that we have to use a higher FDR. In practice, the user has to specify the FDR depending on his needs. It should, however, be noted that with decreasing FDR, the number of true positives also decreases.

FT-BLAST was tested on a dataset of 97 reference compounds and 32 compounds from *P. nudicaule* measured on an Orbitrap XL instrument. Using the reference compounds as query compounds in a leave-one-out evaluation, many hit lists contained compounds mostly from the same class or with high chemical similarity to the query compound. The mean Tanimoto similarity of a query compound to all compounds returned by *FT-BLAST* was 0.76. Furthermore, most compounds in the hit lists belonged to the same or a similar compound class as the query compound. Out of the 652 total hits, only 31 are *clear* false hits; this is less than one could expect for the given FDR of 30 %. For the *P. nudicaule* dataset of 32 unknown compounds, eight compounds were manually identified. For four of these unknowns, reference measurements were available in the dataset, and the top hit of FT-BLAST was the correct compound. The other four compounds were not in the reference dataset; here, search results of *FT-BLAST* included compounds from the biosynthetic pathways of the query compounds, or compounds that shared large substructures with these. First steps have been made towards an automated analysis of the *FT-BLAST* hit lists, searching for characteristic substructures in these lists [24], but the results are currently not chemically sound.

In contrast to classical spectral search methods, *FT-BLAST* searches a library not only for identical, but also for similar compounds. Using a decoy database, the method can estimate FDRs and q-values for each hit. This may help to distinguish true hits from spurious hits. Result lists of *FT-BLAST* are of variable lengths, depending on both the query and the FT library. If the FT library does not contain any similar compounds, the result list might be empty for certain queries. Compounds in the *FT-BLAST* hit lists show high chemical similarity to the query compounds and, for a biological sample, the result lists even contained precursor compounds of the biosynthesis. This shows the potential of *FT-BLAST* for the identification of unknown compounds in practice.

6 Conclusion

It has been recognized that searching in spectral libraries may often be insufficient for the identification of small molecules, such as novel metabolites which serve as drug leads. Truly unknown metabolites are obviously not contained in such libraries, and even searching for similar compounds carries some problems, such as the differentiation between true and spurious hits.

For several decades, not much progress has been made in the automated analysis of fragmentation mass spectra of small molecules, despite searching spectral libraries. But over the last five years, several new ideas and approaches have surfaced: Three approaches share the idea to replace searching in spectral libraries by searching in the more comprehensive molecular structure databases; another approach uses fragmentation trees to "annotate" fragmentation spectra, which facilitates the search for similar compounds. The presented approaches rely on established computational techniques, most notably machine learning and combinatorial optimization.

One may be tempted to replace molecular structure databases by the fully comprehensive molecular structures generated by a molecular isomer generators [20]. It must be noted, though, that the approaches presented here heavily rely on the sparsity of molecular structure databases: Kerber *et al.* [20] report that there are more than 100 million molecular structures for the molecular formula $C_8H_6N_2O$ with mass 146 Da; in contrast, PubChem contains only 323 hits.

As all approaches presented here are relatively young, it is difficult to predict their future development in the years to come. But as these approaches tackle the problem from new and different directions, they appear to be promising tools for, say, the dereplication of compounds and the identification of novel drug leads.

Funding Information: F.H. funded by DFG project BO 1910/16.

Bibliography

[1] F. Allen, R. Greiner and D. Wishart. Competitive Fragmentation Modeling of ESI-MS/MS spectra for metabolite identification. Preprint, Cornell University Library, 2013. arXiv:1312.0264.

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

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

[4] W. Demuth, M. Karlovits and K. Varmuza. Spectral similarity versus structural similarity: Mass spectrometry. *Anal Chim Acta*, 516(1-2):75–85, 2004.

[5] B. Fan, H. Chen, M. Petitjean, A. Panaye, J.-P. Doucet, H. Xia and S. Yuan. New strategy of mass spectrum simulation based on reduced and concentrated knowledge databases. *Spectrosc Lett*, 38(2):145–170, 2005.

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

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

[8] N. A. B. Gray, R. E. Carhart, A. Lavanchy, D. H. Smith, T. Varkony, B. G. Buchanan, W. C. White, and L. Creary. Computerized mass spectrum prediction and ranking. *Anal Chem*, 52(7):1095–1102, 1980.

[9] N. A. B. Gray, A. Buchs, D. H. Smith and C. Djerassi. Computer assisted structural interpretation of mass spectral data. *Helv Chim Acta*, 64(2):458–470, 1981.

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

[11] M. Heinonen, H. Shen, N. Zamboni and J. Rousu. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics*, 28(18):2333–2341, 2012. Proc. of *European Conference on Computational Biology* (ECCB 2012).

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

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

[14] F. Hufsky and S. Böcker. 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)*, pages 12–22. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012.

[15] F. Hufsky, K. Dührkop, F. Rasche, M. Chimani and S. Böcker. Fast alignment of fragmentation trees. *Bioinformatics*, 28(12):i265–i273, 2012. Proc. of *Intelligent Systems for Molecular Biology* (ISMB 2012).

[16] F. Hufsky, M. Rempt, F. Rasche, G. Pohnert and S. Böcker. De novo analysis of electron impact mass spectra using fragmentation trees. *Anal Chim Acta*, 739:67–76, 2012.

[17] F. Hufsky, K. Scheubert and S. Böcker. Computational mass spectrometry for small molecule fragmentation. *Trends Anal Chem*, 53:41–48, 2014.

[18] K. Jeong, S. Kim and N. Bandeira. False discovery rates in spectral identification. *BMC Bioinformatics*, 13 Suppl 16:S2, 2012.

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

[20] A. Kerber, R. Laue, M. Meringer and C. Rücker. Molecules in silico: The generation of structural formulae and its applications. *J Comput Chem Japan*, 3(3):85–96, 2004.

[21] T. Kind and O. Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal Rev*, 2(1-4):23–60, 2010.

[22] K.-S. Kwok, R. Venkataraghavan and F. W. McLafferty. Computer-aided interpretation of mass spectra. III. Self-training interpretive and retrieval system. *J Am Chem Soc*, 95(13):4185–4194, 1973.

[23] J. Lederberg. Topological mapping of organic molecules. *Proc Natl Acad Sci U S A*, 53(1):134–139, 1965.

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

[25] M. H. Medema, K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano, and R. Breitling. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res*, 39(Web Server issue):W339–W346, 2011.

[26] H. Mohimani, W.-T. Liu, J. S. Mylne, A. G. Poth, M. L. Colgrave, D. Tran, M. E. Selsted, P. C. Dorrestein, and P. A. Pevzner. Cycloquest: identification of cyclopeptides via database search of their mass spectra against genome databases. *J Proteome Res*, 10(10):4505–4512, 2011.

[27] H. Mohimani, S. Kim and P. A. Pevzner. A new approach to evaluating statistical significance of spectral identifications. *J Proteome Res*, 12(4):1560–1568, 2013.

[28] S. Neumann and S. Böcker. Computational mass spectrometry for metabolomics – a review. *Anal Bioanal Chem*, 398(7):2779–2788, 2010.

[29] D. D. Nguyen, C.-H. Wu, W. J. Moree, A. Lamsa, M. H. Medema, X. Zhao, R. G. Gavilan, M. Aparicio, L. Atencio, C. Jackson, J. Ballesteros, J. Sanchez, J. D. Watrous, V. V. Phelan, C. van de Wiel, R. D. Kersten, S. Mehnaz, R. De Mot, E. A. Shank, P. Charusanti, H. Nagarajan, B. M. Duggan, B. S. Moore, N. Bandeira, B. Ø. Palsson, K. Pogliano, M. Gutiérrez, and P. C. Dorrestein. MS/MS networking guided analysis of molecule and gene cluster families. *Proc Natl Acad Sci U S A*, 110(28):E2611–E2620, 2013.

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

[31] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, E. Csaszar, and H. C. Köfeler. 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*, 44(4):494–502, 2009.

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

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

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

[35] L. Ridder, J. J. J. van der Hooft, S. Verhoeven, R. C. H. de Vos, R. van Schaik and J. Vervoort. Substructure-based annotation of high-resolution multistage MS^n spectral trees. *Rapid Commun Mass Spectrom*, 26(20):2461–2471, 2012.

[36] K. Scheubert, F. Hufsky, F. Rasche and S. Böcker. Computing fragmentation trees from metabolite multiple mass spectrometry data. *J Comput Biol*, 18(11):1383–1397, 2011.

[37] K. Scheubert, F. Hufsky and S. Böcker. Computational mass spectrometry for small molecules. *J Cheminform*, 5:12, 2013.

[38] E. L. Schymanski, M. Meringer and W. Brack. Matching structures to mass spectra using fragmentation patterns: Are the results as good as they look? *Anal Chem*, 81 (9):3608–3617, 2009.

[39] D. R. Scott. Rapid and accurate method for estimating molecular weights of organic compounds from low resolution mass spectra. *Chemometr Intell Lab*, 16(3):193–202, 1992.

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

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

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

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

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

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

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