Identifying the unknowns by aligning fragmentation trees

Florian Rasche,[†] Kerstin Scheubert,[†] Franziska Hufsky,[†,§] Thomas Zichner,[‡] Marco Kai,[¶] Aleš Svatoš,[¶] and Sebastian Böcker[*,†]

Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany, Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, and Research Group Mass Spectrometry and Proteomics, Max Planck Institute for Chemical Ecology, Jena, Germany

E-mail: sebastian.boecker@uni-jena.de

Preprint of: Florian Rasche, Kerstin Scheubert, Franziska Hufsky, Thomas Zichner, Marco Kai, Aleš Svatoš, and Sebastian Böcker. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.*, 84(7):3417-3426, 2012.

Abstract

Mass spectrometry allows sensitive, automated and high-throughput analysis of small molecules. In principle, tandem mass spectrometry allows us to identify "unknown" small molecules not in any database, but the automated interpretation of such data is in its infancy. Fragmentation trees have recently been introduced for the automated analysis of the fragmentation patterns of small molecules. We present a method for the automated comparison of such fragmentation patterns, based on aligning the compounds' fragmentation trees. We cluster compounds based solely on their fragmentation patterns, and show a good agreement with known compound classes. Fragmentation pattern similarities are strongly correlated with the chemical similarity of molecules. We present a tool for searching a database for compounds with fragmentation pattern similar to an unknown sample compound. We apply this tool to metabolites from Icelandic poppy. Our method allows fully automated computational identification of small molecules that cannot be found in any database.

Mass spectrometry (MS) is a key analytical technology for detecting and identifying small molecules such as metabolites.[1–3] It is orders of magnitude more sensitive than nuclear magnetic

[*]To whom correspondence should be addressed
[†]Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany
[‡]Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[¶]Research Group Mass Spectrometry and Proteomics, Max Planck Institute for Chemical Ecology, Jena, Germany
[§]Max Planck Institute for Chemical Ecology, Jena, Germany

resonance (NMR). Several analytical techniques have been developed, most notably gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). We can analyze thermally unstable metabolites using LC-coupled tandem MS. This technique is usually combined with a gentle ionization, that results in minimal fragmentation of the adduct ions formed. In addition, LC-MS requires less sample preparation as no derivatization step is needed, and is more sensitive and quantitative more accurate.[4] Molecules are mass-selected, fragmented, and the mass-to-charge ratios (m/z) of the resulting fragments recorded. This analytical technique has been applied for many years in proteomics.[5,6]

Computational methods for analyzing fragmentation spectra of small molecules were developed as part of the DENDRAL project.[7] Unfortunately, the project failed to achieve its major objective of automated structure elucidation using MS data. The computational analysis of GC-MS electron impact (EI) fragmentation spectra of small molecules is presumably simpler, as fragmentation is largely reproducible between instruments, and mostly independent of MS model or manufacturer. Computational methods have been developed for searching for similar compounds in a spectral library: In particular, Demuth *et al.*[8] propose a method that aims at finding similar molecules in case a database does not contain the sample molecule; and Stein[9] and Varmuza and Werther[10] present methods to identify chemical substructures of the unknown sample molecule, see ref.[11,12] for similar studies. All of these methods are based on the direct comparison of fragmentation spectra. Even for GC-EI-MS, the resulting computational problems are still far from being "solved". See Kind and Fiehn[13] for a recent review.

Fragmentation in LC-MS experiments (usually collision-induced dissociation (CID)) is less reproducible than fragmentation by electron ionization for GC-MS. Even the time-consuming manual analysis of such data,[14] as well as searching in spectral libraries, are major problems.[15] Apart from a few pioneering studies (e.g. ref.[16–18]), there are few computational methods for the automated analysis of tandem MS data from small molecules. For multiple MS, Sheldon *et al.*[19] proposed a method that takes into account tandem MS spectra of fragments in a "spectral tree". Also, methods exist for *de novo* sequencing of linear or cyclic non-ribosomal peptides,[20–22] but these polymers are structurally strongly restricted.

For decades, MS experts have manually determined fragmentation pathways to explain tandem MS data and determine the molecular structure. In 2008, Böcker and Rasche[23] presented an automated and swift method for annotating tandem MS data using a hypothetical *fragmentation tree* (FT). Tree nodes are annotated with the molecular formulas of the fragments and the edges represent (neutral or radical) *losses*. Computing FTs does not require databases of compound structures or of mass spectra. Neither does it require, apart from lists of common and implausible losses, expert knowledge of fragmentation. Expert evaluation suggests that the FTs are of very good quality.[24] FTs can also be computed from multiple MS data.[25] Rojas-Chertó *et al.*[26] use multiple MS data to derive molecular formulas; note that their "fragmentation trees" are not related to the FTs used herein, but rather to spectral trees.[19] Similar FTs can be identified using visual comparison, which indicates some similarity in the structure of the underlying compounds. Unfortunately, "manual comparison of FTs is also laborious and time-consuming".[24]

Here, we present an automated method for comparing the FTs of two compounds. This allows us to use FTs in applications such as database searching, where we replace the direct comparison of mass spectra by the comparison of the (annotated and more informative) FTs. Our method is based on local tree alignments, generalizing local sequence alignments. We assume that structural similarity is inherently coded in the CID spectra fragments. FT similarity is defined by its edges,

which represent losses and nodes, representing fragments. The local tree alignment contains those parts of the two trees where similar fragmentation cascades occurred.

Aligning FTs when the molecular structure of one compound is known can help elucidate the structure of the unknown compound. We concentrate on the pairwise similarity scores between FTs because these are simple numerical values easily susceptible to *automated* downstream analysis. We present three workflows based on similarity scores. First, we compute pairwise tree alignments for all compounds and so generate a pairwise similarity matrix. We then cluster the compounds based solely on this similarity measure. We find that the clusters that result agree well with the structural properties of the compounds. Second, we calculate pairwise FT alignment similarities and pairwise Tanimoto structural similarities of a dataset of knowns. These similarities are strongly correlated, reaching Pearson correlation coefficients up to $r = +0.68$ ($r^2 = 0.46$) and Spearman correlation coefficients up to $\rho = +0.71$ ($\rho^2 = 0.50$) for certain compound subsets. Third, we determine the similarities of a fragmentation tree from an unknown compound with all trees in a database, to search for related compounds. To filter out spurious hits, we present a statistical evaluation based on decoy database searching. These database hitlists can reveal structural features of the unknown.[27] We name this approach *fragmentation tree basic local alignment search tool* or FT-BLAST for short. Finally, as a proof of principle we show how biological samples from Icelandic poppy (*P. nudicaule*) are analyzed in this framework.

We have elaborated suitable workflows for the process of clustering, database searching, and correlation with chemical similarity (Figure 1). Apart from the need to choose easily accessible parameters for the analysis no user interaction is required, as all workflows are fully automated. Fragmentation tree alignment provides solutions to a major problem in identifying small molecules and it makes possible high throughput computational identification of small molecules even when they have not been databased.

Methods

We analyzed spectra from three reference datasets (Table 1). The *Orbitrap* dataset contains 97 compounds, measured on a Thermo Scientific Orbitrap XL instrument. The *MassBank* dataset[28] consists of 370 compounds measured on a Waters Q-Tof Premier spectrometer. The *QSTAR* dataset contains 44 compounds measured on an API QSTAR QTOF spectrometer by Applied Biosystems.[24] The masses of all compounds ranged from 75 Da to 1258 Da. The supplementary material contains a detailed description of the computational methods.

Acquisition of mass spectra.

For the Orbitrap dataset, 37 compounds were previously measured and used for fragmentation tree evaluation.[24] The remaining compounds originated from our stock, were purchased or donated by M. Strnad (Palacký University, Olomouc, Czech Republic). Some compounds were isotopically labeled with deuterium. The samples were dissolved in methanol (ca. 1 mg/1 mL). They were either introduced into electrospray sources using a built-in infusion pump or mixed and separated by liquid chromatography, then measured on an Orbitrap XL instrument (Thermo Fisher Scientific, Bremen, Germany). Full-scan and CID mass spectra were generated using 30 000 and 7500 full width at half maximum (FWHM) resolution, respectively. The activation time was set at 30 ms with

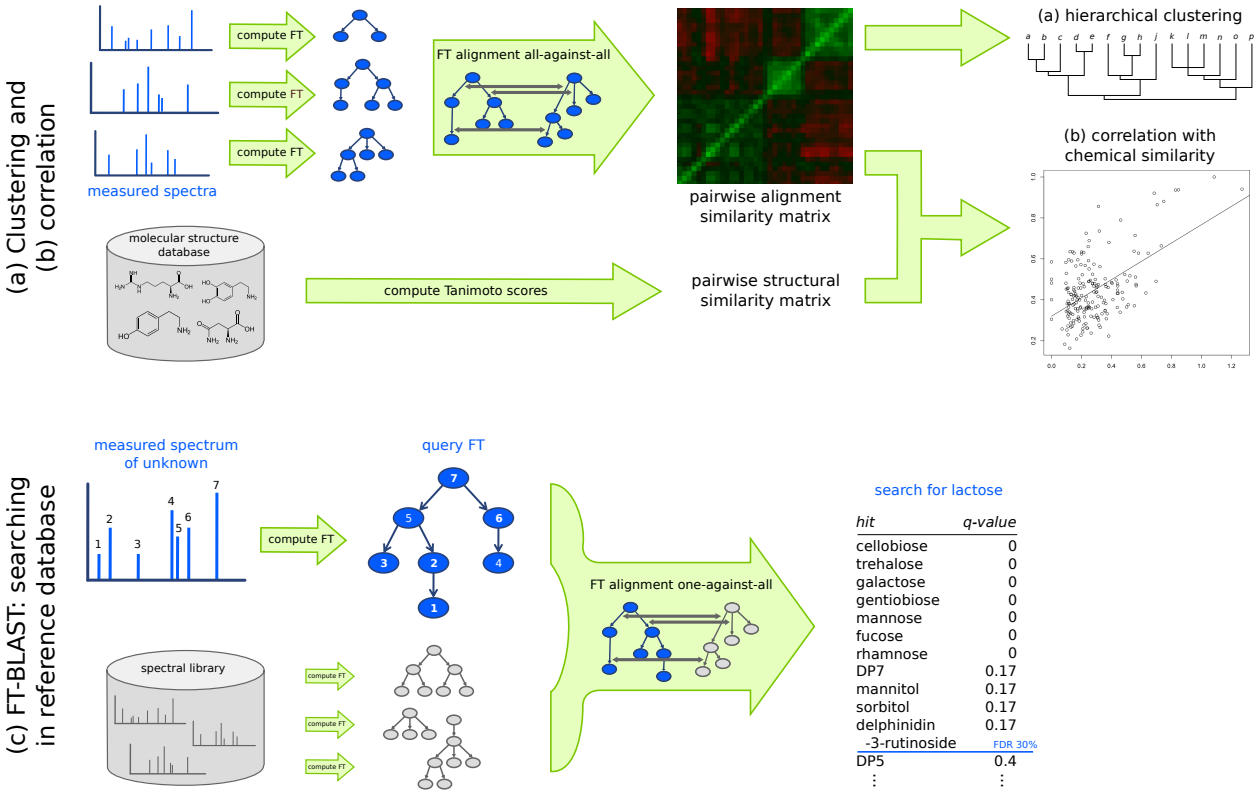


Figure 1: Workflows elaborated for the analysis of tandem MS data. Apart from choosing analysis parameters such as mass accuracy, no user interaction is required. Workflows (a) and (c) are targeted at compounds that are *not* in any database. (a) Clustering of known and unknown compounds using an all-against-all pairwise FT alignment, followed by hierarchical clustering. (b) Correlating FT alignment similarities and chemical similarities for a set of reference compounds. (c) Searching for an unknown compound in databases of reference compounds (either tandem mass spectra or fragmentation trees) using FT-BLAST. This method will return hits (similar compounds) even if the true compound is not in the database. Molecular structures are required only to compute chemical similarities (correlation analysis) or to annotate FT-BLAST hits.

the activation parameter $q = 0.25$. An isolation window of 1.5 mass units was used. Fragmentation was performed using Collision Induced Dissociation (CID) or High-energy Collision Dissociation (HCD). Peak picking was done by the vendor software.

The MassBank dataset was downloaded from the MassBank database[28] with accession numbers PR100001 to PR101056. We discarded compounds with precursor mass deviations above 10 ppm. Mass accuracy 50 ppm for the analysis was chosen by manual inspection of the data. The QSTAR dataset was also from a published source.[24] Peak lists at different collision energies were merged using a window of 50 mDa. This window was determined through visual inpection of a few compounds from the QSTAR dataset. A too small window might cause a few wrong additional fragments to appear in the trees from less accurate datasets, whereas a too wide window results in fragments not appearing in the trees from more accurate datasets. Our alignment approach can compensate for such errors, however.

Table 1: Datasets used in this study. The QSTAR dataset and 38 compounds from the Orbitrap dataset were used for evaluating FTs in ref.[24] The MassBank dataset was downloaded from the MassBank database[28] (`http://www.massbank.jp/`), accession numbers PR100001 to PR101056. We discarded compounds where the measurement of the unfragmented molecule mass deviated more than 10 ppm from the theoretical mass. The MassBank dataset consists of ramp spectra; the other datasets were measured at discrete collision energies. 26 compounds of the Orbitrap dataset were fragmented using higher-energy collisional dissociation (HCD). For these compounds we used fragmentation energies between 5 and 95 arbitrary units. [a]Expert estimate of measurement accuracy. [b]Between 1 and 20 different collision energies. 41 compounds (zeatins, sugars, lipids, bicuculline) were measured at a single collision energy. [c]Some compounds were also measured at 30 eV discrete collision energy. [d]Three to five distinct collision energies for each compound; four compounds measured at a single collision energy.

Name	Orbitrap	MassBank	QSTAR
Mass accuracy (ppm)[a]	< 5	≈ 50	20
collision energy (eV)	between 5 and 150[b]	ramp 5–60, 30[c]	15,25,45,55,90[d]
Number of compounds	97	370	44
Mass range (dalton)	75.0 – 1257.4	90.0 – 822.4	89.0 – 450.2
Median / average mass	342.1 / 346.2	230.0 / 298.0	174.6 / 212.1
FTs with 1+, 3+, 5+, 7+ losses	93, 77, 65, 51	343, 242, 157, 103	44, 43, 32, 28
Major compound classes	zeatins (24), amino acids (19), glucosinolates (14), sugars (12), benzopyrans (11)	flavonoids (85), carboxylic acids (76), amino acids (73), nucleotides (65), sugars (22)	amino acids (21), cholines (18), amines (4)
Compound details	Table 2 and Suppl. Table 7	Suppl. Table 8	Suppl. Table 9

Fragmentation trees and molecular formulas.

For Orbitrap and QSTAR data, we identified molecular formulas following a published method.[24] For each compound, we computed a hypothetical FT, annotating fragment peaks with molecular formulas and modeling fragmentation reactions through dependencies between fragment ions (Figure 2). We performed calculations as described in ref.[24] using a revised and somewhat simplified scoring. The automated computation proceeded in three steps. First, we created a graph containing all molecular formulas that might explain each fragment peak and all potential fragmentation reactions between these formulas. Next, fragmentation reactions were scored, so that the more likely it was that a hypothetical fragmentation reaction was "real", the higher its score. Common losses such as H_2O were given a bonus (Supplementary Table 1). In contrast to the published method[24] we penalized implausible losses (Supplementary Table 2) and we allowed radicals as fragments (Supplementary Table 3). From this graph, we computed the FT with maximum score, annotating every peak once at most. We used an exact method to compute optimal FTs (Supplementary Fig. 10,11,12).

Aligning fragmentation trees.

For the automated comparison of FTs we followed the paradigm of pairwise *local alignments*. We defined a simple similarity measure on the edges (losses) and nodes (fragments) of the two FTs

(Supplementary Table 4). We generalized this similarity measure to trees of identical topology and summed the similarity of tree edges. We also allowed for the insertion and deletion of edges. We searched for *subtrees* in the two FTs that maximized our similarity measure. The rationale for doing so was the same as in the case of local sequence alignments. It is because the molecular structures are not *identical* but subtree similarity indicates structural resemblance.

Tree alignments have been proposed in the context of RNA structure comparison and efficient algorithms have been developed to compute them.[29] In contrast to RNA trees, FTs are unordered, as there is no meaningful ordering of the losses of some fragments. Aligning unordered trees is computationally hard.[29] To compute alignments of unordered trees, we used an exact algorithm based on dynamic programming that guarantees the optimal solution is found. Computational complexity is not usually an issue as the algorithm is efficient if the trees do not contain nodes with many outgoing edges.

Similarity of subtrees was defined as the sum of similarities of edges which, in turn, was chosen to reward identical losses and penalize distinct losses and insertions or deletions. Edge similarities were modified based on the number of non-hydrogen atoms contained. Similarity between fragments (nodes) was also rewarded or penalized (Supplementary Table 4). We modified the published recurrence[29] for solving the problem in three ways. First, we also considered edge similarities. Second, we computed local alignments for maximum subtree similarity by adding a "zero-case" to the recurrence, corresponding to the leaves of the subtree. Third, we scored *join nodes* where two losses were combined into one, corresponding to the non-appearance of intermediate fragmentation steps. Alignment scores will clearly be large for large trees and small for small trees, so we normalized similarities by perfect match scores. To do this we computed for each FT the alignment score against itself, then used the minimum of the two scores, taken to the power of 0.5. We refrained from using the similarity matrix directly. Instead, for each compound we viewed its similarity matrix column as a fingerprint (or feature vector), as is done with gene expression data. We computed the Pearson correlation for any two fingerprints, and processed the resulting *fingerprint similarities*. We implemented all algorithms in Java 1.6.

Clustering.

For each dataset, we performed all-against-all pairwise alignments. We limited calculations to FTs with three and more losses (3+ losses), as smaller trees do not contain sufficient information for clustering. We applied hierarchical clustering[30] (Unweighted Pair Group Method with Arithmetic Mean, UPGMA) to the FT fingerprint similarities using EPoS.[31] Mostly homogeneous clusters were collapsed based on visual inspection.

Correlation with chemical similarity.

Since the chemical structures are known for all reference compounds in our spectral datasets, we can correlate FT fingerprint similarity and chemical similarity. We chose the PubChem/Tanimoto[32,33] measure of chemical similarity because it is the most widely used. We used the Chemistry Development Toolkit[34] (version 1.3.37) to calculate the scores. We did not include any FTs with fewer than one loss.

It is important to note that we did not compare any compound against itself, which trivially results in identical fragmentation patterns, FTs, and molecular structures (including self-comparisons

would result in stronger correlations). We estimated Pearson and Spearman correlation coefficients for all datasets and restrictions using the programming language R. We also performed a between-datasets analysis, where we only considered compound pairs from different datasets.

To evaluate our results, we also tested the correlation of chemical similarity and the classic peak counting score, as well as many of its variants.

FT-BLAST.

The classic way of analyzing tandem MS data is database searching and FT alignments can be used for this task. Given the tandem MS spectra of an unknown compound, we computed its FT, then aligned it to all FTs in our target database, and ranked hits according to fingerprint similarity. Target FTs are constructed from tandem MS data, possibly on the fly. Searching for a "known" compound in a target database is a task that has already been thoroughly studied. We concentrated on the much more intriguing case of where we could not find the query compound in the target database.

An important point is to differentiate between true and spurious hits. We employ a *decoy database strategy* where for each FT in the target database, a similar FT in the decoy database was generated.[35] We created decoy fragmentation trees by using the backbones of real fragmentation trees from another dataset.[16] We searched in the combined target and decoy database, and sorted results with respect to score. We reported hits from the true database only and displayed all hits up to a False Discovery Rate (FDR) of 30%. For each compound hit we can also compute an individual q-value, that is, the smallest FDR for which the hit is included in the output list.

We evaluated FT-BLAST by a *leave-one-out strategy* on the Orbitrap dataset. For each compound we removed the correct answer from the database and searched for the compound in the remainder.

Poppy data.

Surface extracts of *P. nudicaule* were made using methanol: 1% acetic acid 2:1 mixture. The extracts were directly infused using a Nanomate Triversa system (Advion, Ithaca, NY) on a Nanomate nanoelectrospray chip and analyzed on an Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany). Measurements were conducted in both positive and negative mode using several collision energies. Precursor ions were manually selected based on ion intensities and expected masses obtained from literature, and HCD-fragmented. We used a published method[24] to determine molecular formulas. We separately considered the results of the isotope analysis and the compound was kept in the fragmentation analysis only if the sum formula identified was among the top five hits in both cases. FTs, FT alignments, and FT fingerprint similarities were calculated as previously described. We included FTs from unknowns in the fingerprints. We ran FT-BLAST and hierarchical clustering as described above.

Results

In this analysis, we assume that we know the correct molecular formula of each compound. Computing molecular formulas is possible through combining isotope and fragmentation pattern data.[24]

For the QSTAR dataset, identifying the correct molecular formula is possible in all cases.[24] Isotope patterns were available for 51 compounds from the Orbitrap dataset and for 47 of them we identified the correct molecular formula. There are no isotope pattern data available for the MassBank dataset so no molecular formulas were determined.

FTs with fewer losses contain less information and were therefore excluded from our analysis in some cases (Table 1). Computational complexity was not an issue, as running times increase primarily with the out-degree of nodes and, in our experience, FTs rarely contain nodes of out-degree six or higher (example in Figure 2). The average running time for each alignment was below 4 milliseconds on a laptop computer.

Clustering.

Figure 3 and Supplementary Figures 1–4 show the results of the clustering analysis. For Orbitrap data, sugars, zeatins, glucosinolates, amino acids and benzopyrans formed almost perfect clusters. For the MassBank dataset, flavonoids formed one large (64 flavonoids, two other) and three small clusters (12 flavonoids total, one other). Groups of nucleotides, carboxylic acids, sugars, and amino acids formed well-partitioned clusters. For QSTAR data, we observed good partitioning into amino acids, amines, and cholines.

To show how our method applies with measurements from different instruments, we performed combined dataset clustering, in which we clustered all FTs with five and more losses (5+ losses) from the three datasets (Figure 3). We observed many perfect or almost perfect clusters. In addition, compounds of the same class but from different datasets clustered together.

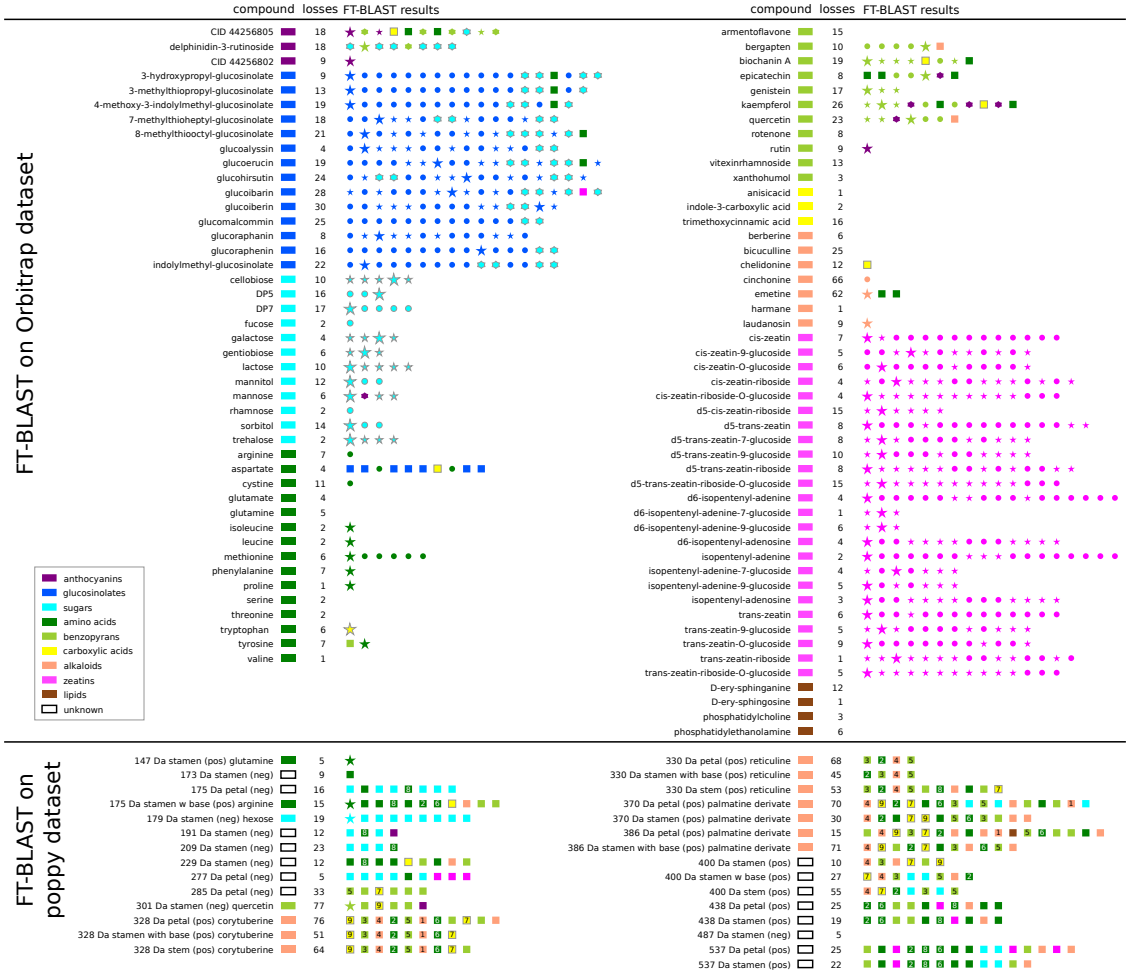
Correlation with chemical similarity.

For the Orbitrap dataset, the Pearson correlation between the FT fingerprint score and the PubChem/Tanimoto score was $r = +0.65$ ($r^2 = 0.42$); this correlation increased slightly for FTs with 3+ losses (Figure 4). For the MassBank dataset, Pearson correlation was $r = +0.50$ ($r^2 = 0.25$). The correlation increased if we restricted ourselves to compounds with more losses. For FTs with seven and more losses (7+ losses) the Pearson correlation was $r = +0.68$ ($r^2 = 0.46$) and Spearman correlation $\rho = +0.71$ ($\rho^2 = 0.50$) (Supplementary Fig. 5). For the QSTAR dataset, the Pearson correlation was $r = +0.63$ ($r^2 = 0.40$) (Supplementary Fig. 6). All correlation results can be found in Supplementary Table 5.

We also performed a between-datasets analysis in which each compound from each dataset (Orbitrap, MassBank, QSTAR) was compared to every compound from the other two datasets. We explicitly excluded comparisons between two compounds from the same dataset. The Pearson correlation was $r = +0.49$ ($r^2 = 0.24$) for the complete datasets and $r = +0.58$ ($r^2 = 0.34$) for FTs with 7+ losses (Figure 4).

We found that correlation between the classical peak counting score and chemical similarity is much weaker than for the FT fingerprint similarity (Supplementary Fig. 9 and Supplementary Table 6).

Table 2: Top: Results of the FT-BLAST analysis for the Orbitrap dataset, compounds with at least one loss ($N = 93$). For each compound, we report results of the leave-one-out search in the database *not* containing the compound we search for. The FDR threshold is set to 30 %. Results are ordered according to fingerprint similarity score. Circles correspond to hits in the same compound class as the query compound, hexagons to hits from a "similar" compound class. Since anthocyanins are made up of sugars and benzopyrans, they are regarded as being similar to both classes; as glucosinolates contain a sugar moiety, these classes are also regarded as being similar. Boxes correspond to hits from all other classes. A large asterisk indicates the compound with the highest chemical similarity (PubChem/Tanimoto), and small asterisks indicate other hits with chemical similarity above 0.85. Symbols are colored by the class of the compound. Overall, we return 557 compounds from the same group, 63 compounds from a similar group, 270 compounds with best or high PubChem/Tanimoto score, and only 31 compounds which do not fall into any of the above categories. In 33 cases (35%) we return the compound with highest chemical similarity at the top position; in 56 cases (60%) this compound is in the TOP 3. Bottom: Searching poppy data in the Orbitrap dataset. A large asterisk indicates the correct identification. Search results mentioned in text and frequent search results are indicated by a boxed number, namely chelidonine (1), phenylalanine (2), laudanosine (3), rotenone (4), bergapten (5), tyrosine (6), trimethoxycinnamic acid (7), glutamate (8), and anisic acid (9).



FT-BLAST.

Table 2 shows the results of the *leave-one-out* FT-BLAST search on the Orbitrap dataset. For each compound we removed the correct answer from the database and searched for the compound in the remainder. For each hit we verified whether it belonged to the same or a chemically "similar" compound class as the query. We also verified whether it had high (PubChem/Tanimoto at least 0.85) or the highest chemical similarity to the query. Many hit lists contained compounds mostly from the same class or with high chemical similarity; other hit lists were short or empty. Only a

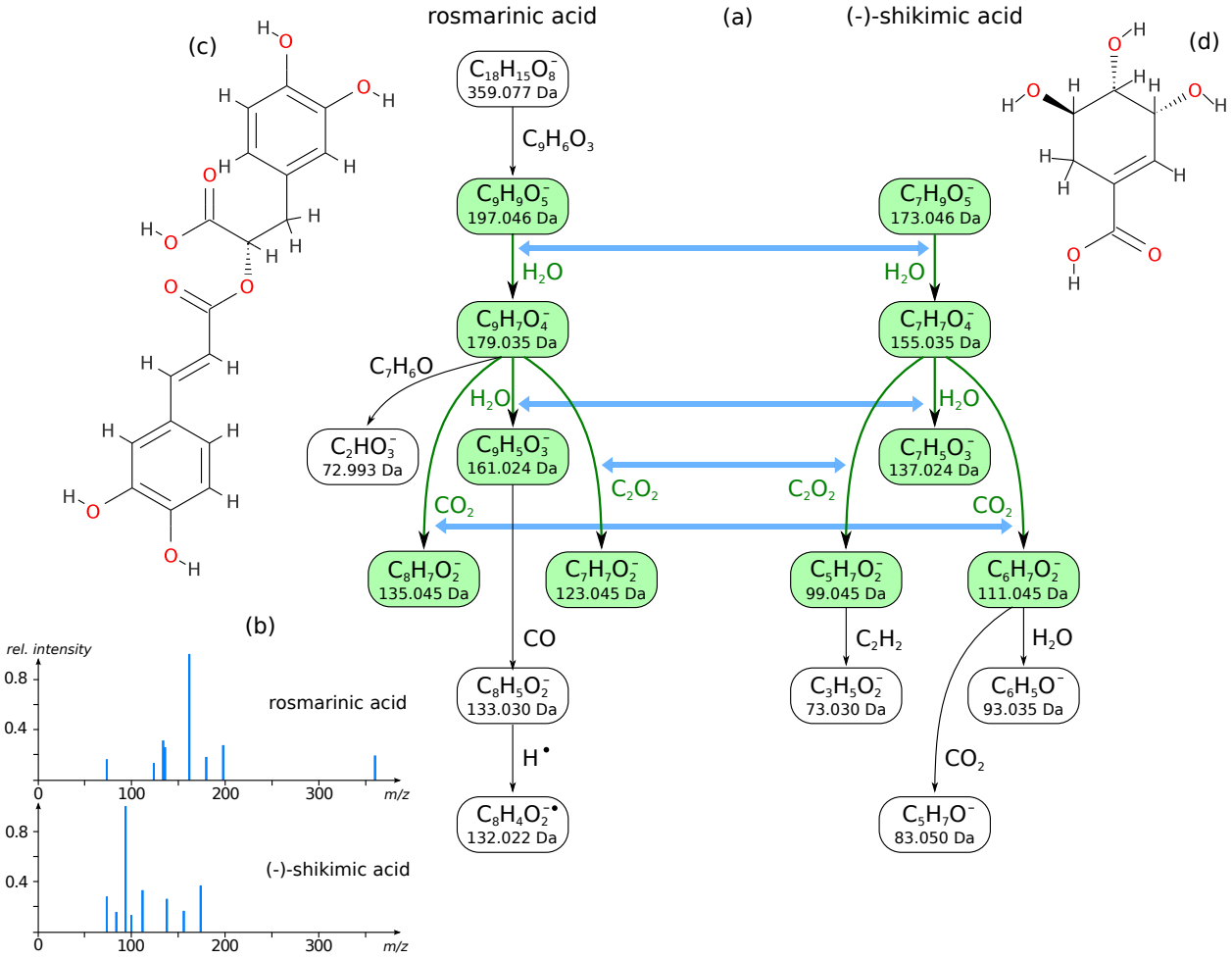


Figure 2: Optimal FT alignment for rosmarinic acid (8 losses) and (-)-shikimic acid (7 losses) from the MassBank dataset (a). The FT fingerprint similarity (from -1 to $+1$) of the mass spectra is $+0.24$. (b) Fragmentation mass spectra of rosmarinic acid and (-)-shikimic acid used for computing FTs. The mass spectra do not share common peaks. Molecular structures of rosmarinic acid (c) and (-)-shikimic acid (d). PubChem Tanimoto score of the compounds is 0.50. The molecular structures are not known to the alignment method. We find that the FT alignment reproduces the key structural similarity of the two compounds: rosmarinic acid loses dehydrocaffeic acid and the anion formed loses two water molecules and carbon dioxide. The (-)-shikimic acid behaves similarly. The key C_2O_2 loss originates from $n, n+1$ dihydroxylation of the aromatic rings. The compounds share a common biosynthetic polyketide origin.

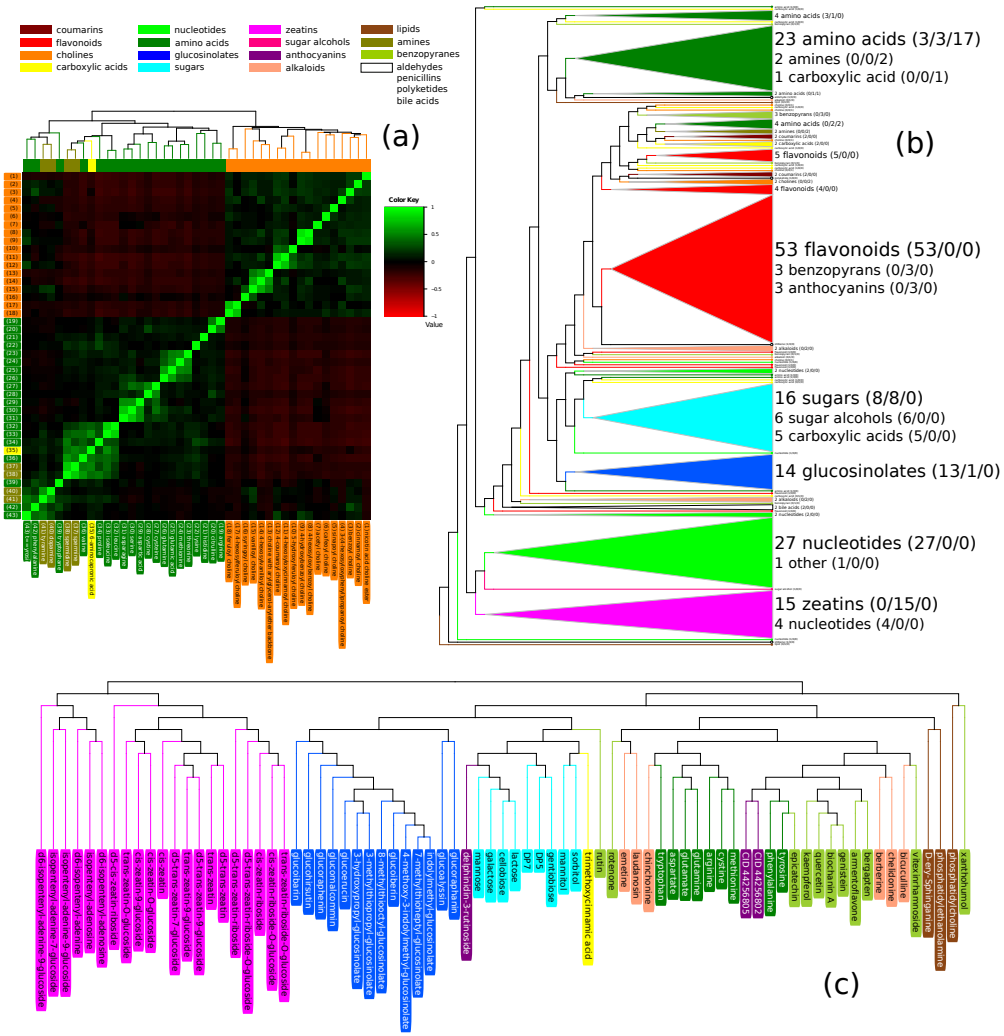


Figure 3: Clustering results based on FT fingerprint similarities. (a) Heat map and hierarchical clustering of the QSTAR dataset, FTs with 3+ losses, $N = 43$. We observe good partitioning of the compounds into amino acids, amines, and cholines. (b) Combined dataset clustering, FTs with 5+ losses, $N = 254$. For better visualization, we have collapsed mostly homogeneous clusters; compounds from different classes are reported as "others". Number of compounds from different datasets are given as "(MassBank/Orbitrap/QSTAR)". Compounds of the same or similar classes but from different datasets, such as amino acids or sugars, cluster together. A nucleotide cluster (from MassBank) forms a subcluster of the zeatin cluster (from Orbitrap). (c) Hierarchical clustering of the Orbitrap dataset, FTs with 3+ losses, $N = 77$. Glucosinolates and zeatins form perfect clusters, all sugars form a cluster together with two other compounds, and large groups of amino acids and benzopyrans form almost perfect clusters.

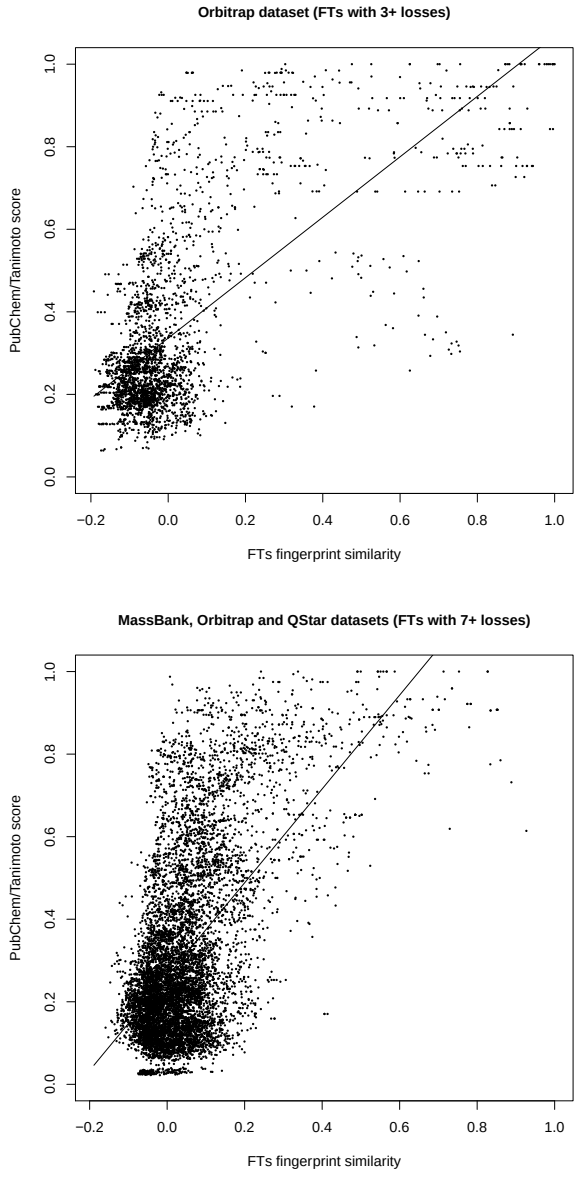


Figure 4: Correlation and regression line: FT fingerprint similarity (x-axis) plotted against chemical similarity measured by PubChem/Tanimoto score (y-axis). Left: Orbitrap dataset, FTs with 3+ losses, $N = 2926$. Pearson correlation is $r = +0.67$ ($r^2 = 0.45$) and Spearman correlation is $\rho = +0.47$ ($\rho^2 = 0.22$). Right: between-datasets analysis, each compound from one dataset is compared to all compounds from the other two datasets. Only FTs with 7+ losses are considered, $N = 9565$. Pearson correlation is $r = +0.58$ ($r^2 = 0.34$) and Spearman correlation is $\rho = +0.43$ ($\rho^2 = 0.18$).

few queries resulted in hit lists with several hits from incorrect compound classes. In fact, only 5% of the hits must be regarded as "wrong". We can use q-values to discriminate further between true and spurious hits. They are omitted from Table 2 solely for the sake of readability.

Similar to Demuth *et al.*[8] we also estimated the average chemical similarity of the query compounds to all compounds returned by FT-BLAST: The mean PubChem/Tanimoto similarity for the complete dataset, using the leave-one-out strategy described above, is 0.76. If we ignore the decoy analysis of FT-BLAST and average over the TOP 5 hits of our search, the mean similarity drops to 0.67; if we combine both approaches and take at most the TOP 5 hits of FT-BLAST, the similarity increases to 0.78.

Identifying unknowns from a biological sample.

As a real-world example of using our method we analyzed several extracts from Icelandic poppy (*P. nudicaule*) in an Orbitrap mass spectrometer. We found 89 features and identified their molecular formulas following a published method.[24] After manual inspection, we selected 32 features with reliably identified molecular formulas. In other cases the isotope patterns of the features were often only faint. FTs were calculated and compared with the Orbitrap dataset using FT-BLAST (Table 2). Eight compounds from the dataset were manually identified. For arginine, glutamine, quercetin and a hexose the top hit was the correct compound from the Orbitrap dataset. FT-BLAST results for reticuline (330.17 Da) and corytuberine (328.15 Da) included laudanosine, several benzopyrans, and phenylalanine, from which these alkaloids are synthesized. Search results for corytuberine also included chelidonine. These two alkaloids share a large substructure. Two other unknowns (370 and 386 Da) were manually classified as palmatine-derivatives. The structurally very similar alkaloid laudanosine was the first or second search result and the other hits were similar to those above. We are currently analyzing the extract using NMR spectrometry to obtain further data for the identification of the novel compounds.

We clustered poppy unknowns together with the Orbitrap dataset (Supplementary Fig. 8). Reticuline, corytuberine, the two palmatine derivatives, and one unknown clustered together with many alkaloids. Other unknowns fell into the amino acid or sugar cluster. A contaminant at m/z 338 (erucamide) was classified as lipid. No unknowns clustered with glucosinolates or zeatins.

Discussion

To achieve the full potential of small molecule MS analysis and to overcome limitations of spectral libraries, we need methods for the computational analysis of fragmentation spectra from unknown compounds. Rule-based approaches for analyzing compound fragmentation spectra may suffer from the tremendous number of rules, both known and unknown. In addition, completely unknown compounds may not necessarily follow the known rules of fragmentation. MS experts have come up with rules for classifying compounds, such as a water and ammonia loss for amino acids. However, real fragmentation patterns are far more complicated, and new "rules" are constantly being introduced. This makes manual compound classification and structure elucidation cumbersome as they require a thorough understanding of fragmentation patterns and profound knowledge of gas-phase ion chemistry and energetics. In contrast, the approach presented here is fully automated and

"rule-free", both when computing and aligning FTs. It only requires sufficiently information-rich fragmentation spectra.

Clustering results show the potential of the method to differentiate compound classes. For the QSTAR dataset, we found good separation into the three compound classes. For Orbitrap data, large compound classes formed almost perfectly separated clusters. Smaller compound classes were distributed among several clusters, but clusters contained few outliers. For the MassBank dataset, flavonoids were perfectly clustered, whereas other compound classes were distributed among several well-separated and homogeneous clusters. Importantly, in the combined dataset clustering, compounds of the same class but from different datasets clustered together. Hierarchical clustering was applied as a proof-of-concept and to demonstrate clustering results. Better results can possibly be achieved by other clustering methods and supervised Machine Learning. Nevertheless, our results indicate how to deduce the compound class of an unknown when a reasonable number of knowns are clustered simultaneously. Even amino acids that did not show the characteristic losses were recognized by the method, such as *N*-formyl-L-methionine and *N*-tigloylglycine (MassBank, 3+ losses, Supplementary Fig. 3).

We found strong correlation between FT similarity and chemical similarity. This is true even for the QSTAR dataset that contained only two major compound classes, and for the MassBank dataset with mass accuracy much lower than 10 ppm. We observed a slight drop in correlation for Orbitrap and QSTAR data for FTs with more losses but assume that this is an artifact (see the Supplementary Material). The correlation between two different measures of chemical similarity (PubChem/MACCS Tanimoto scores) was at most $r = +0.82$ for our datasets, emphasizing the quality of the above results. FT similarity must not be understood as a *prediction* of chemical similarity in the sense of Machine Learning methods. However, FT similarity, expert knoledge, and other sources of information can be combined to permit the accurate prediction of chemical similarity for many compounds.

FT-BLAST achieves a "larger profit" than classical spectral comparison methods, as it searches for similar, not identical, compounds. For the Orbitrap dataset, we achieved excellent search results for most compounds. Even when FT-BLAST returned only a single hit it was often meaningful. Cases where no hits or spurious hits were returned could often be attributed to small FTs, low quality measurements, or the absence of similar compounds from the database. Carboxylic acids and aromatic amino acids were harder to identify as their fragmentation patterns appeared to be more diverse. Results for the smaller QSTAR dataset were of comparable quality. We also found chemically similar hits in the MassBank dataset but the relationships were more complicated than membership in a compound class or Tanimoto similarity.

FT-BLAST individually selects the size of the output for each query compound. For this purpose, we have proposed a method for generating a decoy database of FTs that can be searched simultaneously. Database searching by spectral comparison has been in use for decades; but even today, no sensible methods for generating decoy databases for spectral comparisons have been developed. Although FT-BLAST returned an average of 8.4 compounds per query on the Orbitrap dataset, the average similarity of these results is much higher than when choosing the TOP 5 in all cases (see also Supplementary Fig. 7). The chemical similarities reported above (0.67 to 0.78) compare well to the numbers from ref.[8]: the highest TOP 5 chemical similarity reported there is 0.605, obtained after extensive parameter optimization. But clearly, we cannot rule out that the improved performance is due to different database sizes, content, or spectral qualities.

By applying FT-BLAST and clustering to an unknown sample from poppy, we confirmed

eight manual identifications and suggested compound classes for some other unknowns, as they were unquestionably members of a well-defined cluster. Particularly remarkable was that we also identified the biosynthetic precursor of several alkaloids, which come from mixed biosynthetic pathways. The analysis of unknowns will become more powerful as more reference compounds become available. Our results may also simplify downstream NMR analysis.

The results presented are of good quality, but further improvements are possible with better scoring and when more data becomes available. We found that optimizing sample preparation and instrument settings to obtain fragment rich CID spectra could be advantageous. With compounds for which tandem MS does not produce a sufficient number of fragments, computing FTs from multiple MS spectra may be beneficial.[25] Other fragmentation techniques, such as Electron Transfer Dissociation (ETD), can be analyzed by FT alignments, as our method is not limited specifically to CID fragmentation. In the future, we want to include more expert knowledge on characteristic losses and ions.

FT alignments open a way to a fast classification/identification of metabolites, limiting work spent on ubiquitously occuring "uninteresting" molecules. Areas of application include natural product discovery, identifying catabolic processing of drugs, dereplication and searching for biomarkers.[36] In future, the systems biology approach of inferring biosynthetic pathways and metabolic networks from tandem MS data might be improved by using FT similarities instead of spectral comparisons.[37,38]

Acknowledgement

KS funded by Deutsche Forschungsgemeinschaft (BO 1910/10-1). FH supported by the International Max Planck Research School, Jena. We thank Masanori Arita (University of Tokyo, Japan) for providing the MassBank data, Ravi Kumar Maddula for measuring some of the compounds in the Orbitrap dataset, Miroslav Strnad (Palacký University, Olomouc, Czech Republic) for supplying the zeatins, and Evangelos Tatsis (MPI-CE Jena) for poppy samples. Financial support from the Max Planck Society is acknowledged.

Supporting Information Available

This material is available free of charge via the Internet at `http://pubs.acs.org`.

References

(1) Fernie, A. R.; Trethewey, R. N.; Krotzky, A. J.; Willmitzer, L. *Nat. Rev. Mol. Cell Biol.* **2004**, *5*, 763–769.

(2) Last, R. L.; Jones, A. D.; Shachar-Hill, Y. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 167–174.

(3) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26*, 162–164.

(4) Josephs, J. L.; Sanders, M. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 743–759.

(5) Cravatt, B. F.; Simon, G. M.; Yates, J. R. *Nature* **2007**, *450*, 991–1000.

(6) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.

(7) Lederberg, J. *Proc. Natl. Acad. Sci. U. S. A.* **1965**, *53*, 134–139.

(8) Demuth, W.; Karlovits, M.; Varmuza, K. *Anal. Chim. Acta.* **2004**, *516*, 75 – 85.

(9) Stein, S. *JASMS* **1995**, *6*, 644–655.

(10) Varmuza, K.; Werther, W. *J. Chem. Inf. Comp. Sci.* **1996**, *36*, 323–333.

(11) Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. *Metabolomics* **2010**, *6*, 322–333.

(12) Tsugawa, H.; Tsujimoto, Y.; Arita, M.; Bamba, T.; Fukusaki, E. *BMC Bioinformatics* **2011**, *12*, 131.

(13) Kind, T.; Fiehn, O. *Bioanal. Rev.* **2010**, *2*, 23–60.

(14) Werner, E.; Heilier, J.-F.; Ducruix, C.; Ezan, E.; Junot, C.; Tabet, J.-C. *J. Chromatogr. B* **2008**, *871*, 143–163.

(15) Oberacher, H.; Pavlic, M.; Libiseller, K.; Schubert, B.; Sulyok, M.; Schuhmacher, R.; Csaszar, E.; Köfeler, H. C. *J. Mass Spectrom.* **2009**, *44*, 485–493.

(16) Hill, D. W.; Kertesz, T. M.; Fontaine, D.; Friedman, R.; Grant, D. F. *Anal. Chem.* **2008**, *80*, 5574–5582.

(17) Pelander, A.; Tyrkkö, E.; Ojanperä, I. *Rapid Commun. Mass Spectrom.* **2009**, *23*, 506–514.

(18) Heinonen, M.; Rantanen, A.; Mielikäinen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043–3052.

(19) Sheldon, M. T.; Mistrik, R.; Croley, T. R. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 370–376.

(20) Ng, J.; Bandeira, N.; Liu, W.-T.; Ghassemian, M.; Simmons, T. L.; Gerwick, W. H.; Linington, R.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Methods* **2009**, *6*, 596–599.

(21) Bandeira, N.; Pham, V.; Pevzner, P.; Arnott, D.; Lill, J. R. *Nat. Biotechnol.* **2008**, *26*, 1336–1338.

(22) Mohimani, H.; Yang, Y.-L.; Liu, W.-T.; Hsieh, P.-W.; Dorrestein, P. C.; Pevzner, P. A. *Proteomics* **2011**, *11*, 3642–3650.

(23) Böcker, S.; Rasche, F. *Bioinformatics* **2008**, *24*, I49–I55, Proc. of *European Conference on Computational Biology* (ECCB 2008).

(24) Rasche, F.; Svatoš, A.; Maddula, R. K.; Böttcher, C.; Böcker, S. *Anal. Chem.* **2011**, *83*, 1243–1251.

(25) Scheubert, K.; Hufsky, F.; Rasche, F.; Böcker, S. *J. Comput. Biol.* **2011**, *18*, 1383–1397.

(26) Rojas-Chertó, M.; Kasper, P. T.; Willighagen, E. L.; Vreeken, R. J.; Hankemeier, T.; Reijmers, T. H. *Bioinformatics* **2011**, *27*, 2376–2383.

(27) Henneberg, D.; Weimann, B.; Zalfen, U. *Org. Mass Spectrom.* **1993**, *28*, 198–206.

(28) 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. et al. *J. Mass Spectrom.* **2010**, *45*, 703–714.

(29) Jiang, T.; Wang, L.; Zhang, K. *Theor. Comput. Sci.* **1995**, *143*, 137–148.

(30) D'haeseleer, P. *Nat. Biotechnol.* **2005**, *23*, 1499–1501.

(31) Griebel, T.; Brinkmeyer, M.; Böcker, S. *Bioinformatics* **2008**, *24*, 2399–2400.

(32) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

(33) Rogers, D. J.; Tanimoto, T. T. *Science* **1960**, *132*, 1115–1118.

(34) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.

(35) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.

(36) Kulasingam, V.; Diamandis, E. P. *Nat. Clin. Pract. Oncol.* **2008**, *5*, 588–599.

(37) Berthoumieux, S.; Brilli, M.; de Jong, H.; Kahn, D.; Cinquemani, E. *Bioinformatics* **2011**, *27*, i186–i195.

(38) Krumsiek, J.; Suhre, K.; Illig, T.; Adamski, J.; Theis, F. J. *BMC Systems Biology* **2011**, *5*, 21.

Identifying the unknowns by aligning fragmentation trees Supplementary methods and material

Florian Rasche[1], Kerstin Scheubert[1], Franziska Hufsky[1,4], Thomas Zichner[2], Marco Kai[3], Aleš Svatoš[3], and Sebastian Böcker[1]

[1] Chair for Bioinformatics, Friedrich Schiller University, Jena, Germany
[2] Genome Biology Research Unit, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany
[3] Research Group Mass Spectrometry and Proteomics, Max Planck Institute for Chemical Ecology, Jena, Germany
[4] Max Planck Institute for Chemical Ecology, Jena, Germany

Table of Contents

1 Experimental section

To show the applicability of our method to diverse types of MS data, we use three datasets in this study (Table 1). The first dataset consists of 97 compounds measured on an Orbitrap mass spectrometer. The second dataset was downloaded from the MassBank database. The third dataset contains 44 compounds measured on an API QSTAR. Supplementary Tables 7, 8, 9 list the compounds in the datasets.

For the Orbitrap dataset, 37 compounds were previously measured and used for fragmentation tree evaluation [17]. The remaining 60 substances were from our laboratory stocks and were either previously purchased or isolated from natural sources. The Orbitrap dataset mainly contains zeatins, amino acids, glucosinolates, sugars and benzopyrans. For 41 compounds (zeatins, sugars, lipids, bicuculline) only a single fragmentation energy was used.

The MassBank dataset was downloaded from the MassBank database [9] at `http://www.massbank.jp/`, accession numbers PR100001 to PR101056. These spectra were measured on a Waters Q-Tof Premier instrument at the RIKEN Plant Science Center (Yokohama, Japan) by F. Matsuda, M. Suzuki, and Y. Sawada. We discarded 47 compounds where the measurement of the *unfragmented* molecule mass deviated more than 10 ppm from the theoretical mass, leaving us with 370 compounds. We stress that mass accuracy of fragment ions is worse than 10 ppm: The instrument configures itself to be most accurate in the mass range of the precursor mass. Thus, even if the mass accuracy of the *unfragmented* molecule mass is below 10 ppm, the fragment ions distributed over the full mass range may be much more inaccurate. By visual inspection of mass spectra and FTs, we decided to use an accuracy of 50 ppm. So, mass accuracy is one order of magnitude worse than for the Orbitrap data. Most MS^2 spectra in this dataset were recorded in ramp mode, with collision energy varying from 5–60 eV. Some compounds were additionally measured at a fixed energy of 20 eV. In these cases we merged the spectra, but disabled the scoring of collision energies. Among others, the dataset contains flavonoids, with and without sugar moieties, saccharides, and nucleotides.

The QSTAR dataset was measured on an API QSTAR QTOF instrument by Applied Biosystems with mass accuracy 20 ppm. This dataset was measured at the Leibniz Institute of Plant Biochemistry (Halle, Germany) by Christoph Böttcher. It contains 44 compounds, most of them amino acids and phenolic choline esters, plus four biogenic amines and one carboxylic acid. MS^2 spectra were measured at three to five collision energies. Only four compounds were measured at a single collision energy. Experimental details for the QSTAR dataset can be found in [17].

We merged peak lists for product ion spectra acquired from the same precursor at different collision energies. For that, peaks from different product ion spectra with less than 50 mDa distance were considered to represent the same fragment ion. This relatively large mass window was found to improve the mass accuracy of the data by averaging over peaks from several measurements. Such peaks were combined into a single peak: The mass of the resulting peak is the weighted mean of peak masses, where weights were chosen as signal intensities in the product ion spectra. The intensity of the resulting peaks is simply the sum of intensities of the peaks in the product ion spectra. Intensities were not scaled, since this would compromise comparison of peak intensities from product ion spectra measured at different collision energies.

2 Identifying molecular formulas

In [17], molecular formulas were correctly identified in all cases for the QSTAR dataset and 21 compounds from the Orbitrap dataset. For another 30 compounds from the Orbitrap dataset used in this study, isotope patterns were measured. In 26 of 30 cases, we identified the correct molecular formula as described below. We found mass accuracy to be insufficient to identify the molecular formulas of the two anthocyanins with masses above 1000 Da. For two compounds (tyrosine and sphingosine), the correct molecular formula is in second place. In case of tyrosine, the isotope pattern intensities are inaccurate, whereas for sphingosine too few fragment peaks were recorded.

To limit memory usage, we slightly modify the method from [17] for determining the molecular formula. First, isotope patterns are scored as described in [2]. Then, fragmentation trees are calculated for the 20 best candidate molecular formulas only, and their score is calculated as described in the next section. We stress that a score from the hetero-to-carbon ratio of a molecular formula is added as a prior to the fragmentation tree score. The logarithmized value of the gaussian density function with mean 0.59 and SD 0.56 is used as prior score [3]. Fragmentation tree scores and isotope pattern scores are combined as described in [17], and molecular formula candidates are sorted with respect to the combined score.

For the MassBank dataset, no isotope pattern information is available, so we cannot identify molecular formulas from the experimental data.

It must be understood that even in cases were we cannot unambiguously determine the molecular formula from the data, it is possible to use the FT alignment setup described in this paper: In case of doubt about the molecular formula of an unknown, we can use the trees of several molecular formula hypotheses as queries or clustering input.

3 Computing fragmentation trees

We assume that the correct molecular formula of each compound is known: Such formulas can be determined without user interaction from high quality MS data. In [17], molecular formulas were correctly identified in all cases for the QSTAR dataset and a subset of the Orbitrap dataset used here. The results for another part of the Orbitrap dataset are described in the previous section.

For each compound, we calculate a hypothetical FT from the tandem MS data, as described in [3, 17]. FTs are computed solely from the experimental MS data, optimizing a scoring function. First, a fragmentation graph is build, where vertices correspond to molecular formulas that are within the mass accuracy of some peak, and that are sub-formulas of the compound ion molecular formula. Vertices of the graph are colored, and molecular formulas corresponding to the same peak receive the same color. We draw a directed edge (arc) between a pair of vertices if the second molecular formula is a sub-formula of the first.

We then *weight* vertices and edges of the fragmentation graph, based on the likelihood that a certain vertex or edge is “true”. Further details can be found in [3, 17]. For vertices, we use log odds to differentiate between the model (the peak is truly a fragment with the proposed molecular formula) and the background (the peak is noise):

– We use the mass difference between the measured peak and the molecular formula, and assume mass differences to be normally-distributed [10, 23]. The basic score of the vertex

loss name	loss formula	loss name	loss formula
Water	H_2O	Deoxyhexose equivalent	$C_6H_{10}O_4$
Methane	CH_4	Hexose equivalent	$C_6H_{10}O_5$
Ethene	C_2H_4	Hexuronic equivalent acid	$C_6H_8O_6$
Ethine	C_2H_2	Ammonia	NH_3
Butene	C_4H_8	Methylamine	CH_5N
Pentene	C_5H_8	Methylimine	CH_3N
Benzene	C_6H_6	Trimethylamine	C_3H_9N
Formaldehyde	CH_2O	Cyanic Acid	CHNO
Methanol	CH_4O	Urea	CH_4N_2O
Carbon monoxide	CO	Phosphonic acid	H_3PO_3
Formic acid	CH_2O_2	Phosphoric acid	H_3PO_4
Carbon dioxide	CO_2	Metaphosphoric acid	HPO_3
Acetic acid	$C_2H_4O_2$	Dihydrogen vinyl phosphate	$C_2H_5O_4P$
Ketene	C_2H_2O	Hydrogen sulfide	H_2S
Propionic acid	$C_3H_6O_2$	Sulfur	S
Malonic acid	$C_3H_4O_4$	Sulfur dioxide	SO_2
Malonic anhydride	$C_3H_2O_3$	Sulfur trioxide	SO_3
Pentose equivalent	$C_5H_8O_4$	Sulfuric acid	H_2SO_4

Supplementary Table 1: The *common losses* used in our calculations. If an entry from this table or a combination thereof occurs in a hypothetical fragmentation step, the score of this step is significantly increased.

is computed via the logarithmized Gaussian probability density function of the mass difference, with SD 1/3 of the instrument's mass accuracy.
- We then add λ times the peak intensity to the vertex score, with $\lambda = 0.04$. This is the negative log likelihood that the peak is a noise peak, assuming an exponential distribution of peak intensities.
- For the Orbitrap and QSTAR datasets, we use default parameters $\alpha = 0.1$ and $\beta = 0.8$ for collision energy scoring, see [3] for details. We found that these parameters have only small impact on FT computation, so we leave out further details.

Next, we score the edges of the fragmentation graph:

- We use a list of *common losses* that one expects to see in a tandem MS experiment, see Supplementary Table 1. This table was modified from Table 2 in [17] by including methanol (CH_4O). Combinations of up to three losses from this table are rewarded by $\log_{10}(\gamma/n)$, where n is the number of combined common losses. We use $\gamma = 10$ (+1) for the Orbitrap and the MassBank dataset, and $\gamma = 1000$ (+3) for the QSTAR dataset.
- Different from [17] we penalize for *implausible losses* that were repeatedly annotated "wrong" by MS experts, see Supplementary Table 2. If a loss *equals* a implausible loss, we penalize it by adding $\log_{10}(10^{-3}) = -3$ to its score.
- Similarly, losses containing only nitrogen or only carbon are penalized by $\log_{10}(10^{-4}) = -4$.
- Also different from [17] we allow radicals as fragments. We penalize a radical loss with $\log_{10}(10^{-3}) = -3$, unless it is one of the common radical losses from Supplementary Table 3. In that case, the score is not modified.
- To avoid star-like FTs where all fragments branch from the root, we penalize large losses by $\log_{10}(\frac{1-\text{mass loss}}{\text{parent mass}})$.

"loss name"	loss formula
"Dicarbon monoxide"	C_2O
"Tetracarbon monoxide"	C_4O
"Unsaturated cyclopropane"	C_3H_2
"Unsaturated cyclopentane"	C_5H_2
"Unsaturated cycloheptane"	C_7H_2

Supplementary Table 2: The *implausible losses* used in our calculations. If an entry from this table occurs in a hypothetical fragmentation step, the score of this step is significantly decreased. We believe that such losses should only very rarely (if ever) occur in a FT, so we penalize their appearance. We do no completely forbid them, as this conflicts the idea of an optimization-based method. It turns out that none of the implausible losses appears in any FT computed for this study.

loss name	loss formula
Atomar hydrogen	$H^{\cdot}$
Oxygen radical	$O^{\cdot}$
Hydroxy radical	$\cdot OH$
Methyl radical	$\cdot CH_3$
Methoxy radical	$CH_3O^{\cdot}$
Propyl radical	$\cdot C_3H_7$
tert-Butyl radical	$\cdot C_4H_9$
Phenoxy radical	$C_6H_5O^{\cdot}$

Supplementary Table 3: The *radical losses* used in our calculations. If an entry from this table occurs in a hypothetical fragmentation step, this is not penalized. Other radical losses are not forbidden, but the score of the corresponding step is significantly decreased.

The weight of every vertex is pulled to each incoming edge, so that the resulting graph is solely edge-weighted. To find the hypothetical FT, we search for a colorful subtree inside the fragmentation graph that has maximum weight. Calculations were carried out using an exact method, resulting in score-optimal fragmentation trees. We have attached all FTs computed from all datasets in Supplementary Figures 10–12.

Some compounds did not fragment significantly, resulting in hypothetical FTs with an insufficient number of losses. Especially amino acids and carboxylic acids have mostly less than three losses. This is due to current instruments limited mass range at 50 thomson, too high for small amino acids like glycine and alanine.

The quality of fragmentation trees has already been evaluated by experts [17]. For the datasets used in [17], 78.96 % of the losses were assigned as "correct", 13.37 % as "unsure", and 7.67 % as "wrong". Fragmentation tree results improved using the extended scoring described above, including a penalty for implausible losses (Supplementary Table 2).

4 Scoring alignments

Since we base our FT alignment on losses and fragments, we need a scoring function to evaluate pairs of losses, as well as pairs of fragments. In our scoring we distinguish three main cases for two losses nl_1 and nl_2. Those cases are a match $nl_1 = nl_2$, a mismatch $nl_1 \neq nl_2$, or an insertion/deletion (indel) where either $nl_1 = \lambda$ or $nl_2 = \lambda$ is a gap symbol. A summary of scores can be found in Supplementary Table 4. In detail, we define:

	Event	Score
losses	Basic match score	+5
	Modification for each non-hydrogen atom	+1
	Basic mismatch score	-2
	Modification for each non-hydrogen atom	-0.5
fragments	Basic match score	+5
	Modification for each non-hydrogen atom	+1
	Basic mismatch score	-3
	Modification for each non-hydrogen atom	±0
	Insertion/deletion score	±0
	Merging losses modification	±0

Supplementary Table 4: Scoring neutral losses and fragments.

- For a *match*, we assign a positive score. This score depends on the size of the losses, since agreement between larger losses is more significant than between smaller ones. We set $\delta(nl, nl) := 5 + \#atoms$ where $\#atoms$ is the number of non-hydrogen atoms in the loss nl (that is, all carbon and hetero atoms).
- For a *mismatch* we assign a negative score, that increases when the losses get more dissimilar. We set $\delta(nl_1, nl_2) := -5 - \#diff$ where $\#diff$ is the number of non-hydrogen atoms in the symmetric difference between the two losses. As an example, $nl_1 = C_2H_3O_2$ and $nl_2 = C_4H_4O_1N_1$ differ in two carbon, one oxygen, and one nitrogen atoms, a total of four non-hydrogen atoms, so $\delta(C_2H_3O_2, C_4H_4O_1N_1) = -5 - 4 = -9$.
- For an *insertion/deletion* we set $\delta(nl_1, \lambda) = \delta(\lambda, nl_2) = 0$, as deleting nodes from the alignment implicitly reduces the score that can be reached.
- Finally, we will allow two subsequent losses to be *merged* in one of the tree. Here, we set $\delta_{\text{merge}} := \pm 0$. We do not penalize merged losses, as merging losses implicitly reduces the score that can be reached by the alignment.

Scoring of fragment pairs is somewhat similar. For two fragments f_1 and f_2 we again distinguish between match $f_1 = f_2$ and mismatch $f_1 \neq f_2$. To correctly compare trees measured in negative and positive mode, we "neutralize" the fragment ion formulas by adding or subtracting a hydrogen atom.

- For a *match*, we assign a positive score depending on the size of the fragment. We set $\delta(f, f) := 5 + \#atoms$ where $\#atoms$ is the number of non-hydrogen atoms in the fragment f (that is, all carbon and hetero atoms).
- For a *mismatch* we assign a negative score not depending on the symmetric difference between the two fragments. We set $\delta(f_1, f_2) := -3$ for $f_1 \neq f_2$. In this way, we allow for matching losses even when the corresponding fragments show no similarity.

Recall that some compounds in the Orbitrap dataset are isotopically labeled with deuterium. When comparing molecular formulas of losses or fragments in the alignment, we treat deuterium as hydrogen. As an example, losses H_2O and HDO would receive a score of +6.

To avoid overfitting, we have deliberately kept the proposed scoring very simple. At a later stage, when more datasets become available, optimization of the scoring scheme may further improve the quality of alignments.

5 Aligning fragmentation trees

Whereas efficient, polynomial-time algorithms exist for the alignment of ordered trees, the alignment of unordered trees is computationally hard, namely MAX SNP-hard [11]. Still, there exists an algorithm for computing exact solutions to this problem, that has reasonable running time in practice. The reason for this is that FTs usually have comparatively small out-degree: Fragments rarely have more than, say, five daughter fragments. We can limit the inevitable exponential part of the running time to this out-degree. Jiang *et al.* [11] proposed an exact algorithm based on dynamic programming to compute global alignments of unordered trees. Here, we modify this algorithm for our purpose of aligning FTs.

We use dynamic programming to compute the maximal score $S(T_1,T_2)$ of a local alignment between two trees T_1,T_2. Let $N(v)$ denote the children of any node v in T_1 or T_2. In the following, let u be a node of T_1, and v a node of T_2. Let $D[u,v]$ be the maximal score of a local alignment of two subtrees of T_1,T_2, where the subtree of T_1 is rooted in u, and the subtree of T_2 is rooted in v. For $A \subseteq N(u)$ and $B \subseteq N(v)$ we define $D_{u,v}[A,B]$ to be the score of an optimal local alignment with subtree rooted in u and v, respectively, such that *at most* the children A of u and B of v are used in the alignment. Note that all children A of u and B of v can be used, but also, any subset is allowed, including the empty set. Clearly, we have $D_{u,v}[A,\emptyset] = D_{u,v}[\emptyset,B] = 0$ for all A,B. Now, $D[u,v] = D_{u,v}[N(u),N(v)]$ holds.

We initialize $D_{u,v}[A,B] = 0$ for $A = \emptyset$ or $B = \emptyset$. In the recurrence, we distinguish three cases, namely *match* (including mismatches), *deletion*, or *insertion*, where the latter two are symmetric to each other. For non-empty sets $A \subseteq N(u)$ and $B \subseteq N(v)$ we get

$$
\begin{aligned}
D_{u,v}[A,B] &= \max\Big\{0, match_{u,v}[A,B], delete_{u,v}[A,B], insert_{u,v}[A,B]\Big\} \\
match_{u,v}[A,B] &:= \max_{a\in A, b\in B}\Big\{D[a,b] + D_{u,v}\big[A-\{a\},B-\{b\}\big] + \delta(ua,vb)\Big\} \\
delete_{u,v}[A,B] &:= \max_{a\in A, B'\subseteq B}\Big\{D_{a,v}\big[N(a),B'\big] + D_{u,v}[A-\{a\},B-B'] + \delta(ua,\lambda)\Big\} \\
insert_{u,v}[A,B] &:= \max_{A'\subseteq A, b\in B}\Big\{D_{u,b}\big[A',N(b)\big] + D_{u,v}[A-A',B-\{b\}] + \delta(\lambda,vb)\Big\}
\end{aligned}
$$

where $\delta(ua,vb)$ denotes the score of the losses attached to arcs ua and vb, and $\delta(ua,\lambda),\delta(\lambda,vb)$ accordingly. Finally, we compute the maximal score of a local alignment of T_1,T_2 as

$$S(T_1,T_2) = \max_{u\in T_1, v\in T_2} D[u,v].$$

Merging two losses in T_1 or T_2 requires additional care, as several losses may be joined simultaneously: Every node can choose to become a JOIN node, in which case it cannot participate in the matching itself, whereas all losses below the JOIN node are incremented by the loss above the join node. To compute the corresponding score, we have to iterate over all subsets of children of some node u in T_1 that we assume to be JOIN nodes, and match them optimally to some children of v in T_2. This can be achieved by dynamic programming similar to above, where we have to introduce a PREJOIN case where a node will become a JOIN node for its parent. A corresponding optimization is required for the case that a JOIN node is present in T_2. We leave out the technical details.

6 Normalization of scores and fingerprinting

Since the score of an alignment is highly dependent on the size of the trees, alignment scores have to be normalized: In the extreme case of an FT with only one vertex (the parent

molecule), the alignment score is zero against *all* other trees. To this end, we normalize by the score that a *perfect match* would obtain. Since we do local alignments, a perfect match means that the one tree is a subtree of the other one. The same score is obtained by aligning this subtree with itself, $S(T_i, T_i)$. So, we normalize the score by

$$S_0(T_1, T_2) = \frac{S(T_1, T_2)}{\left(\min\{S(T_1, T_1), S(T_2, T_2)\}\right)^c} \tag{1}$$

where $c \in [0, 1]$ is the normalization parameter. Here, $c = 1$ corresponds to a full normalization by the perfect match score, whereas $c = \frac{1}{2}$ corresponds to the square root of this value. We do not to choose the full score for normalization, since it is much more likely for a very small tree to be a subtree of another tree, than it is that a medium-size or large tree is a subtree of another tree. To this end, $c = 1$ favors small trees and discriminates against large trees, whereas no normalization ($c = 0$) favors large trees. In our study, we choose $c = \frac{1}{2}$.

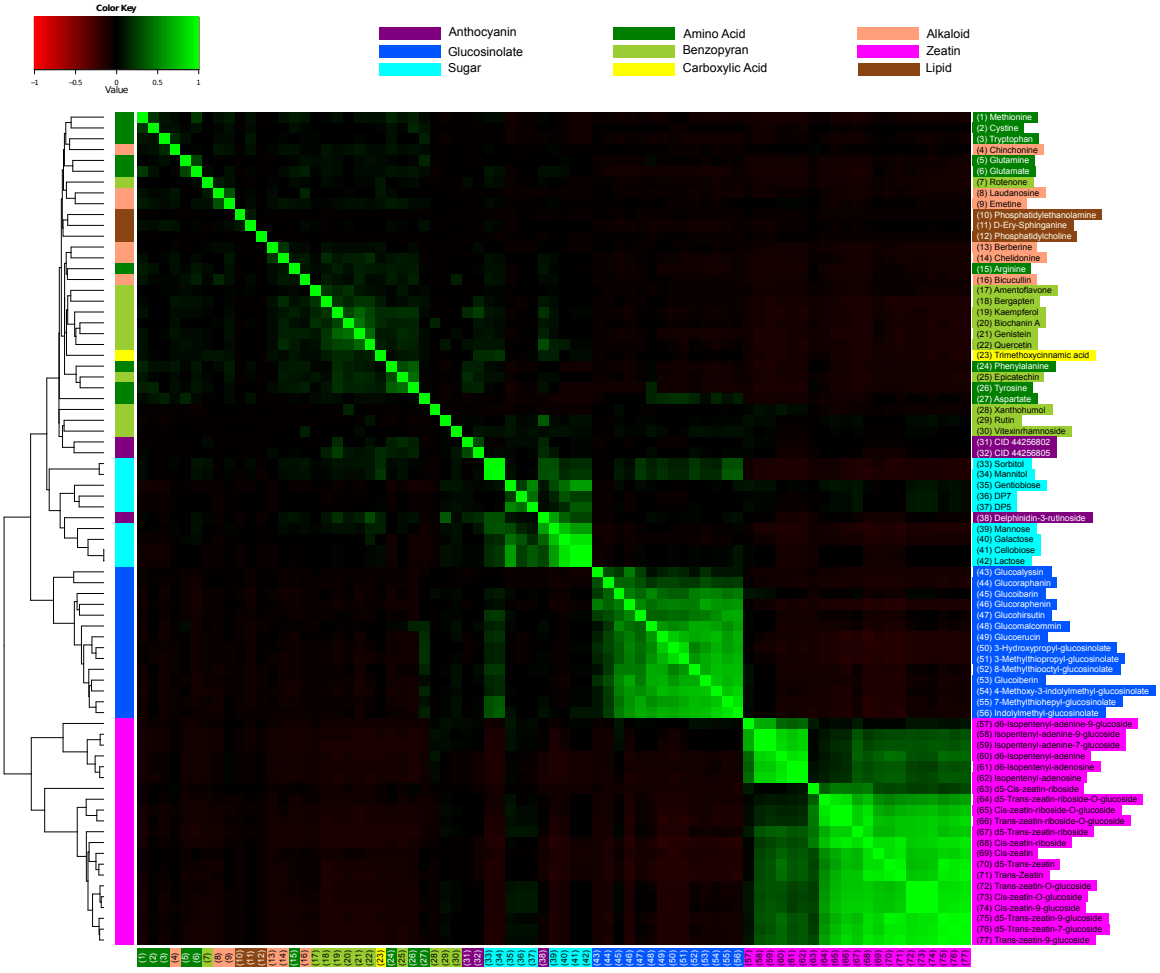
Instead of directly using normalized scores, we found that an additional re-evaluation of similarities is useful: When two compounds are structurally similar, they should show comparable FT similarities to *any* other compound. To this end, we use the scores of one compound against all others as its *fingerprint* or *feature vector*. We compare two compounds by comparing their fingerprints. This can be achieved using any classical methods for comparing feature vectors, such as Euclidean distance or Pearson correlation. In our study, we chose the Pearson correlation coefficient, see (2) in Section 8 below.

7 Clustering

We compute pairwise alignments of FTs for all compound pairs, as explained in Section 5. We normalize scores by perfect match score using $c = \frac{1}{2}$ in (1), and compute fingerprints of the compounds as described in Section 6. This results in a matrix of pairwise similarities. To this matrix, we apply hierarchical clustering or, more precisely, UPGMA (Unweighted Pair Group Method with Arithmetic Mean) agglomerative clustering [19]. Again, we stress that hierarchical clustering is probably not the best-suited method for clustering compounds based on FT similarity; rather, we have chosen this method as it is well-known, particularly in the context of analyzing gene expression data [5].

It is understood that for FTs with few losses, clustering results will become somewhat arbitrary: In the extreme case of a single neutral loss, similarity or dissimilarity to any other FT can easily be spurious. To this end, we limit clustering to FTs with a lower bound on the number of losses. Somewhat unexpectedly, we were able to set this lower bound as small as three losses, while still retaining a good quality of the clustering. Still and all, we have to exclude a number of compounds from our cluster analysis, see Table 1. We believe that this is not a shortcoming of our method, but rather the problem that certain compounds do not "fragment sufficiently" under tandem MS, resulting in mostly uninformative fragmentation spectra. As indicated in the Discussion section, this problem may be overcome by using multiple MS.

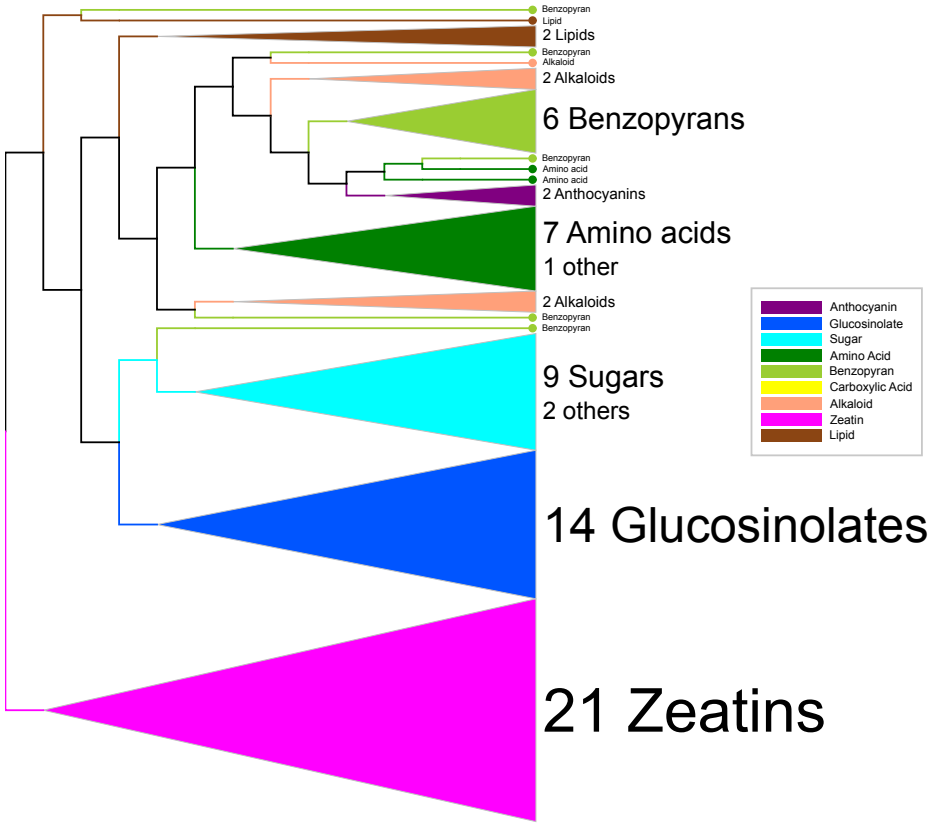
We first analyze the Orbitrap dataset. We discarded 20 compounds as the resulting FTs showed less than three losses. The Orbitrap dataset contains mostly zeatins (21 with 3+ losses), glucosinolates (14), benzopyrans (11), sugars (9), and amino acids (9). The heat map of the fingerprint similarity matrix is depicted in Supplementary Figure 1. The clustering is depicted in Figure 3 and Supplementary Figure 1. Finally, clustering with collapsed mostly-homogeneous clusters is depicted in Supplementary Figure 2. We observe



Supplementary Fig. 1: Heat map and hierarchical clustering for the Orbitrap dataset, compounds with 3+ losses.

that clusters are very homogeneous: There is a perfect glucosinolate cluster containing all 14 glucosinolates, a perfect zeatin cluster containing all 21 zeatins, and an almost perfect sugar cluster containing all nine sugars, plus one anthocyanin and one carboxylic acid. Furthermore, there is an almost perfect amino acid clusters containing seven of the nine amino acids plus one alkaloid. Similarly, there is a perfect benzopyran cluster containing six of the eleven benzopyrans.

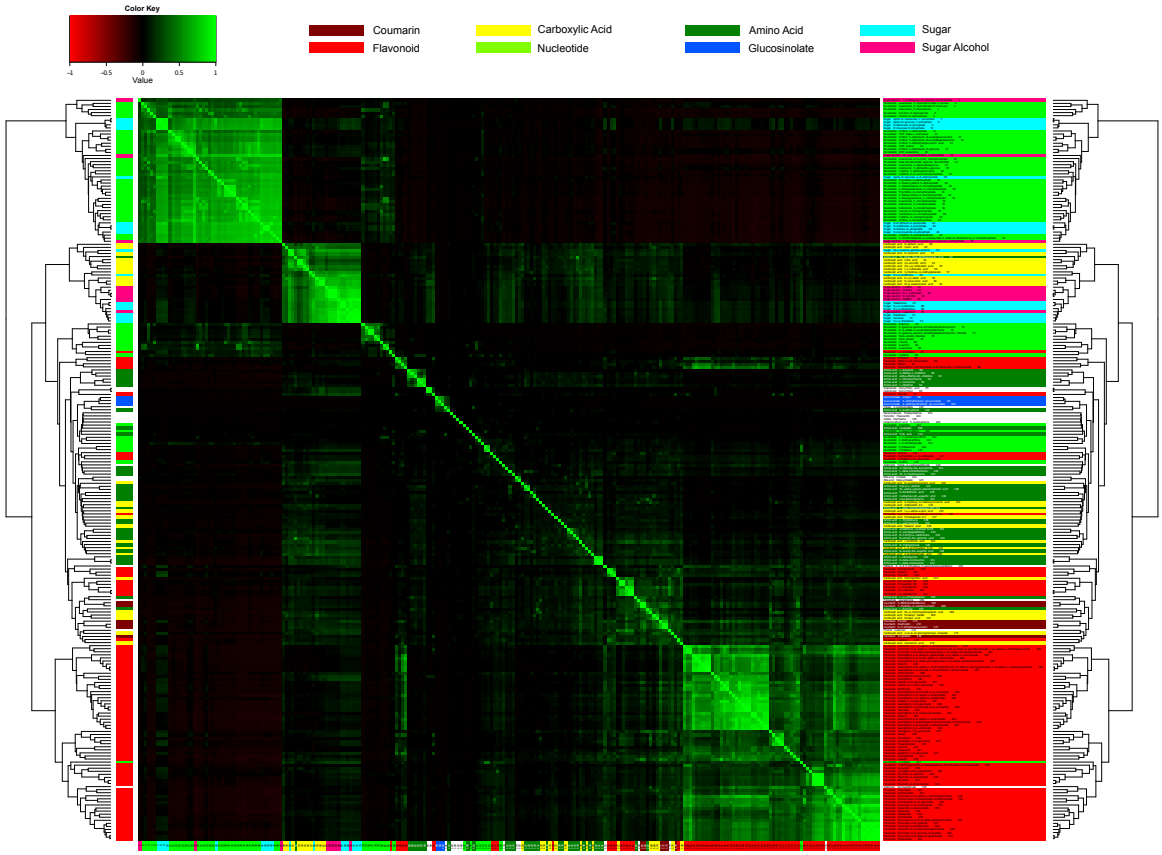
For the MassBank dataset, we had to discard 128 compounds with less than three losses. Here, we find a large group of flavonoids (81 with 3+ losses), nucleotides (54), amino acids (33), carboxylic acids (26), and sugars (17). The heat map of the similarity matrix plus the clustering is depicted in Supplementary Figure 3. Clustering with collapsed mostly-homogeneous clusters is depicted in Supplementary Figure 4. We observe an almost perfect cluster of 64 flavonoids containing only two non-flavonoid compounds. For amino acids we find five perfect clusters containing 22 of the 33 amino acids in total. Similarly, we find four carboxylic acid clusters containing ten carboxylic acids plus one other compound. For nucleotides there are seven small perfect clusters, containing 32 nucleotides in total, and a large cluster containing 16 nucleotides but also four sugars and two sugar alcohols.



Supplementary Fig. 2: Hierarchical clustering of the Orbitrap dataset (compounds with 3+ losses) where for better visualization, we have collapsed (mostly) homogeneous clusters.

Finally, we analyze the QSTAR dataset: This dataset contains biogenic amino acids and complex choline derivatives [3]. We observe a well partitioning of the compounds into amino acids, amines and cholines, see Figure 3 for heat map and hierarchical clustering.

To show the applicability of our method between measurements from different instruments, we performed a combined dataset clustering: We cluster all compounds from the Orbitrap, MassBank and QSTAR datasets for FTs with 5+ losses, leaving us with 157 compounds from the MassBank dataset, 65 compounds from the Orbitrap dataset, and 32 compounds from the QSTAR dataset. We report results in Figure 3. We observe a large amino acid cluster containing three amino acids from the MassBank, three amino acids from the Orbitrap and 17 amino acids from the QSTAR dataset. Furthermore, eight sugars from MassBank and eight sugars from Orbitrap form a large cluster with six sugar alcohols and five carboxylic acids from MassBank. The only remaining glucosinolate from MassBank forms a perfect cluster with the 13 remaining glucosinolates from Orbitrap. Finally, an almost perfect cluster of 27 nucleotides from MassBank forms a subcluster of the almost perfect zeatin cluster, containing 15 zeatins from Orbitrap and four nucleotides from MassBank. This demonstrates that the structures of the fragmentation trees are highly similar although/albeit the fundamental differences between Q-Tof and Orbitrap mass analyzers.



Supplementary Fig. 3: Heat map and hierarchical clustering for the MassBank dataset, compounds with 3+ losses.

8 Correlation with chemical similarity

As all of the compounds in our datasets are references with known molecular structure, we can estimate their structural similarity, termed *chemical similarity* in the following. This allows us to compare chemical similarities with our FT alignment-based similarities. This is meant as a proof-of-concept: In applications, we obviously do *not* know the molecular structure of the unknown query compound. But our results clearly show the correlation between these similarity values.

For measuring correlation, we use the well-known Pearson product-moment correlation coefficient r (*Pearson correlation coefficient* for short) that measures the linear dependence of two variables $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$:

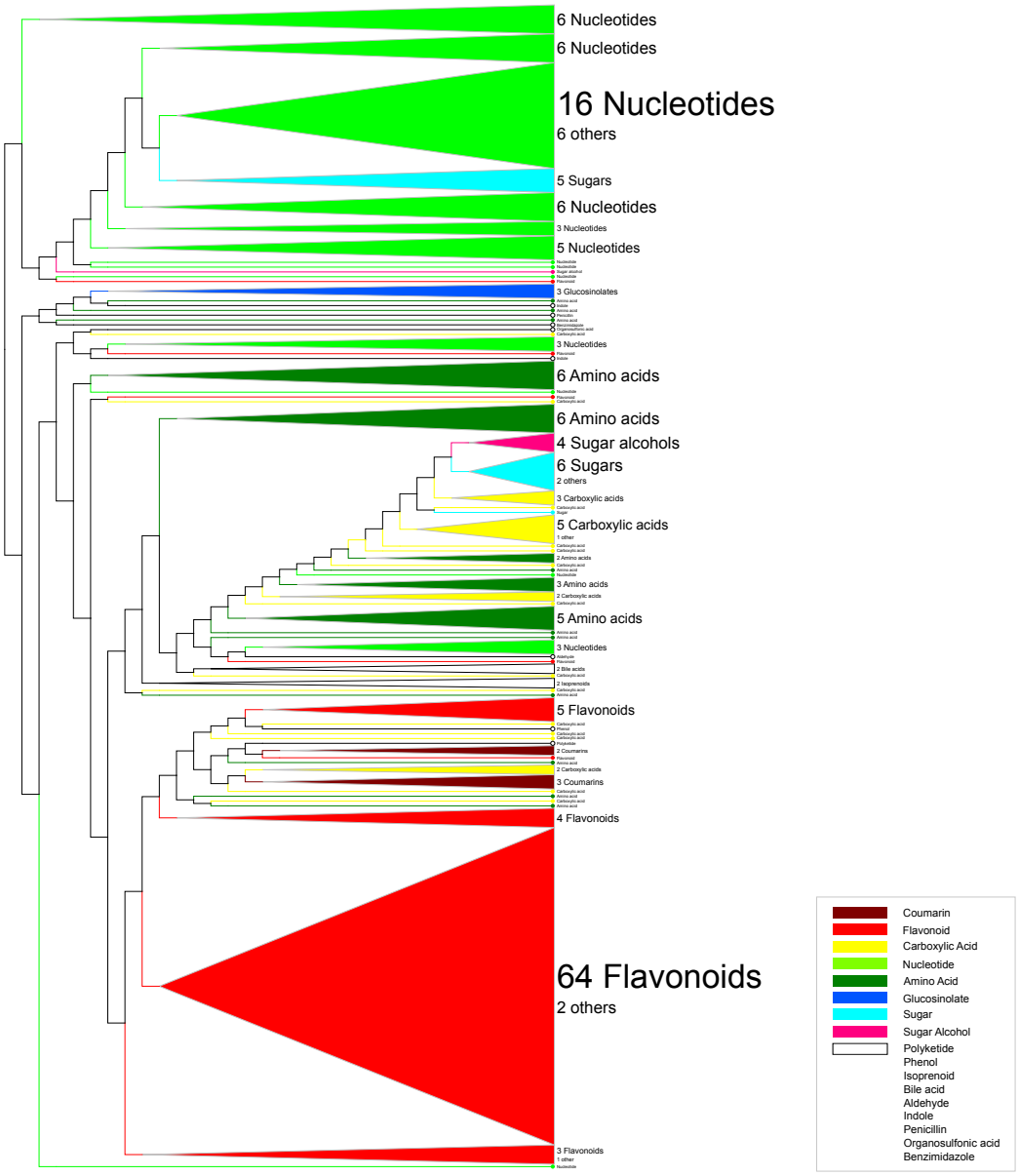
$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2}$$

with $-1 \le r \le +1$. Here, $\bar{X}$ denotes the mean of $X_1, \ldots, X_n$. We also compute the *Spearman correlation coefficient* ρ that is the Pearson correlation coefficient of the ranked variables. The values X_i, Y_i are each converted to ranks $x_i, y_i \in \{1, \ldots, n\}$, and

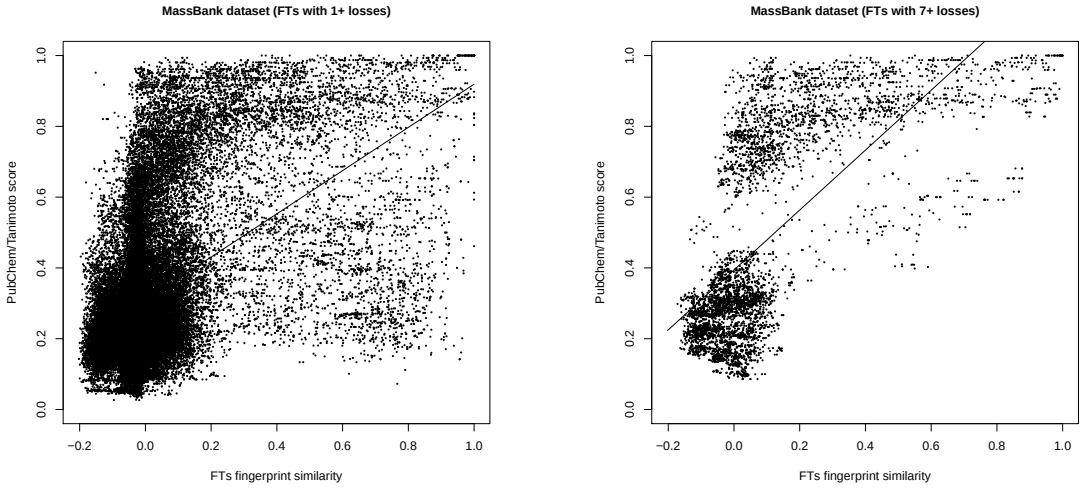
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}\left(x_i - \frac{n+1}{2}\right)\left(y_i - \frac{n+1}{2}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \frac{n+1}{2}\right)^2 \sum_{i=1}^{n}\left(y_i - \frac{n+1}{2}\right)^2}} \tag{3}$$

where again, $-1 \leq \rho \leq +1$. Ties can be broken by assigning fractional ranks. Computations of correlation coefficients were carried out using the program language R.

To judge the level of correlation between the two similarities, we stress that these are not two measurements where, say, by the laws of physics, one expects a linear dependence. This being said, we argue that any Pearson correlation coefficients $r > 0.5$ ($r^2 > 0.25$) can be regarded as strong correlation. This is even more so since two different chemical similarity scores based on comparing molecular (sub-)structures, namely PubChem/Tanimoto and another Tanimoto score that uses Molecular ACCess System (MACCS) fingerprints [6], show a Pearson correlation of less than $r = +0.82$, see below. Similarly, a Spearman correlation coefficient of $\rho > 0.5$ ($\rho^2 > 0.25$) indicates a strong but possibly non-linear correlation.



Supplementary Fig. 4: Hierarchical clustering of the MassBank dataset (compounds with 3+ losses) where for better visualization, we have collapsed (mostly) homogeneous clusters.



Supplementary Fig. 5: Correlation and regression line, MassBank dataset. FTs fingerprint similarity (x-axis) plotted against chemical similarity measured by PubChem/Tanimoto score (y-axis). Left: FTs with 1+ losses ($N = 58653$). Pearson correlation is $r = +0.50$ ($r^2 = 0.25$), Spearman correlation is $\rho = +0.43$ ($\rho^2 = 0.18$). Right: FTs with 7+ losses ($N = 5253$). Pearson correlation is $r = +0.68$ ($r^2 = 0.46$), Spearman correlation is $\rho = +0.71$ ($\rho^2 = 0.50$).

Again, we normalize FT alignment scores by perfect match score using $c = \frac{1}{2}$ in (1), and compute fingerprints of the compounds as described in Section 6. To show the effect of the fragmentation tree size on the correlation with chemical similarity, we differentiate between those compounds with FTs that have at least 1+, 3+, 5+, and 7+ losses, respectively. See Table 1 for the number of compounds remaining in the different datasets. For a dataset with n compounds, this results in $\binom{n}{2} = \frac{n(n-1)}{2}$ compound pairs where we can correlate the two similarity values. We stress that we do not measure the similarity of a compound against itself: Any method for comparing fragmentation patterns should be able to pick up the similarity of two *identical* patterns. Including such self-comparisons would result in even higher but possibly misleading correlation coefficients.

Many different similarity scores have been developed in chemoinformatics to compare molecular structures [13]. We concentrate on one of the most commonly used frameworks [1], namely binary fingerprint representations with Tanimoto similarity scores (Jaccard indices) [18]. We decided to use fingerprints of the PubChem database [22] as again, we argue that it is particularly widely used. We use the Chemistry Development Toolkit version 1.3.37 [21] for our computations. We stress that these computations were performed completely independent of FT alignment computations; computations of FT alignments were carried out without any knowledge of the chemical structures.

See Supplementary Table 5 for all correlation coefficients and the number of alignments from which the coefficients are computed. See Figure 4 for the correlation plot of the Orbitrap dataset, FTs with 3+ losses. See Supplementary Figure 5 for the correlation plots of the MassBank dataset, FTs with 1+ and 7+ losses, and Supplementary Figure 6 for the correlation plot of the QSTAR dataset, FTs with 1+ losses. Finally, see again Figure 4 for the correlation plot of the between-datasets analysis, FTs with 7+ losses.

Different methods for measuring chemical similarity will result in different similarities of the compounds. To this end, we have estimated the correlation of two different measures

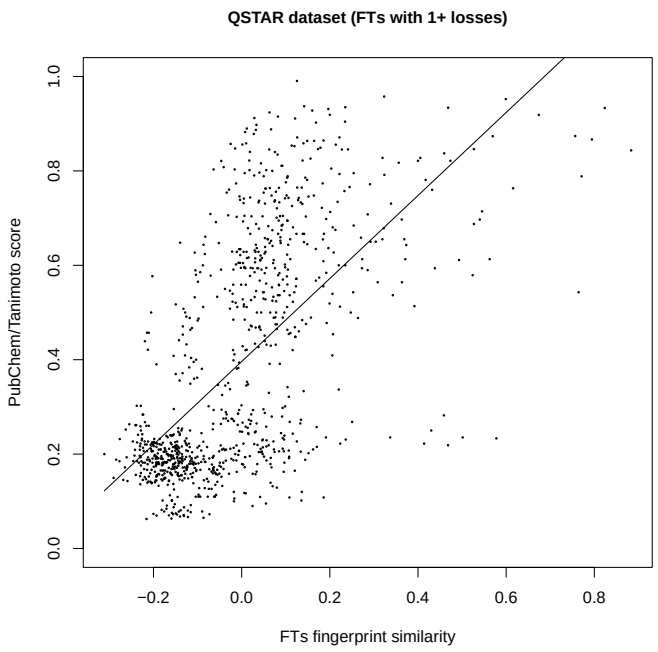
Dataset	correlation method	only compounds with 1+ losses	3+ losses	5+ losses	7+ losses
Orbitrap	Pearson r	0.65	0.67	0.64	0.58
	Pearson r^2	0.42	0.45	0.41	0.34
	Spearman ρ	0.45	0.47	0.48	0.51
	Spearman ρ^2	0.20	0.22	0.23	0.26
	no. compound pairs N	4278	2926	2080	1275
MassBank	Pearson r	0.50	0.60	0.67	0.68
	Pearson r^2	0.25	0.36	0.45	0.46
	Spearman ρ	0.43	0.52	0.64	0.71
	Spearman ρ^2	0.18	0.27	0.41	0.50
	no. compound pairs N	58653	29161	12246	5253
QSTAR	Pearson r	0.63	0.62	0.55	0.51
	Pearson r^2	0.40	0.38	0.30	0.26
	Spearman ρ	0.64	0.64	0.61	0.55
	Spearman ρ^2	0.41	0.41	0.37	0.30
	no. compound pairs N	946	903	496	378
Between-dataset	Pearson r	0.49	0.52	0.55	0.58
	Pearson r^2	0.24	0.27	0.30	0.34
	Spearman ρ	0.37	0.40	0.38	0.43
	Spearman ρ^2	0.14	0.16	0.14	0.18
	no. compound pairs N	51083	32351	17309	9565

Supplementary Table 5: Correlation of chemical similarity (PubChem/Tanimoto) with fragmentation tree similarity, for all datasets and different restrictions on the number of losses. For the between-dataset correlation, only compound pairs from different datasets are considered. We also report the number of alignments (compound pairs) N for every set.

of chemical similarity, namely the PubChem/Tanimoto score and the MACCS/Tanimoto score [6], the later being part of the Open Babel project [7]. The Pearson correlation of PubChem/Tanimoto and MACCS/Tanimoto scores is between $r = +0.74$ and $r = +0.81$ for the Orbitrap dataset, between $r = +0.79$ and $r = +0.82$ for the MassBank dataset, and between $r = +0.74$ and $r = +0.79$ for the QSTAR dataset. Analogously, the Spearman correlation is between $\rho = +0.66$ and $\rho = +0.70$ for the Orbitrap dataset, between $\rho = +0.70$ and $\rho = +0.82$ for the MassBank dataset, and between $\rho = +0.73$ and $\rho = +0.75$ for the QSTAR dataset. These values may be seen as *upper bounds* for the correlation that we can possibly reach between FT similarity and chemical similarity.

All three datasets show a good correlation ($r \geq 0.50$). We reach the best correlation ($r = +0.65$) for the Orbitrap dataset that contains many compound classes. For the QSTAR dataset comprised of only two major compound classes we still reach a very strong Pearson correlation of $r = +0.63$. But even for the MassBank dataset with mass accuracy much worse than 10 ppm there is a good correlation, which increases to very strong Spearman correlation $\rho = +0.71$ for FTs with 7+ neutral losses.

As shown in Supplementary Table 5, the correlation coefficients of the MassBank dataset increase by limiting the correlation analysis to FTs with more neutral losses. This may appear evident, since correlation with chemical similarity requires that information is present in the FTs. Nevertheless, the correlation coefficients of the QSTAR and the Orbitrap datasets decrease when limiting the analysis to bigger trees. Interestingly, also the correlation between the MACCS/Tanimoto scores and the PubChem/Tanimoto scores of these



Supplementary Fig. 6: Correlation and regression line, QSTAR dataset. FTs fingerprint similarity (x-axis) plotted against chemical similarity measured by PubChem/Tanimoto score (y-axis). Only FTs with 1+ losses (N = 946). Pearson correlation is r = +0.63 (r^2 = 0.40), Spearman correlation is ρ = +0.64 (ρ^2 = 0.41).

two datasets decreases from r = +0.79 to r = +0.74 for the QSTAR dataset, respectively from r = +0.81 to r = +0.74 for the Orbitrap dataset. We believe that the weaker correlation of FTs with more losses is an artifact of our data. Some compound classes fragment better than others, and limiting the compounds to bigger FTs implies limiting the compound subsets to less compound classes. For example, in the QSTAR dataset 13 of the 16 FTs with less than seven losses are cholines. Thus, the reduced subset consists of 64% amino acids. Possibly, a strong correlation within only one or few compound classes is more difficult, since FTs of one compound class are very similar and not sensitive enough to predict small differences between the structures.

To demonstrate that the strong correlation coefficients are not artifacts (measuring all compounds with one instrument and by one person), we performed a between-datasets analysis: Each compound from each dataset (Orbitrap, MassBank, QSTAR) is compared to each compound from the other two datasets. This is done to separate the intra-dataset correlation from the inter-dataset correlation. We reach Pearson correlation r = +0.49 (r^2 = +0.24) for the between-datasets analysis, and r = +0.58 (r^2 = +0.34) for FTs with 7+ losses. Our results indicate that the method is robust against differences in sample preparation, instruments, and raw data processing methods. This may allow us to search for compounds in “mixed” databases where we do not limit the search to reference compounds measured under similar conditions as the query compound, see the next section. In this way, we may considerably enlarge the set of reference compounds for identifying the unknown.

9 Fragmentation Tree Basic Local Alignment Search Tool

We noted above that the important point in database searching, is to differentiate between true and spurious hits. Obviously, one of the FTs has maximal similarity among all trees in the database, but this does not mean that this best hit is a good hit.

To assess the significance of hits, we generated a decoy database: For each FT in the target database, a FT in the decoy database is constructed. For a target tree tree with m edges, we randomly generate a decoy tree with m edges. Unfortunately, we have no statistical model of the structure of fragmentation trees; at the same time, we believe that the topology of FTs is extremely important for the alignment. To this end, we chose to generate *decoy fragmentation trees* from an independent dataset. We computed FTs for the fragmentation data from 102 compounds measured on a Micromass QTOF, published by Hill *et al.* [8]. Using compounds from an independent dataset has two advantages: On the one hand, these are true FTs, so decoy FTs are structurally "similar" to the true FTs. On the other hand, this is an independent dataset, so any similarity to true FTs must be fully at random. Using the Hill *et al.* dataset [8] has the additional advantage that resulting FTs are large, allowing us to compute subtrees more easily: To generate a random tree with m losses, we first discard all decoy trees with less than m edges. From the remaining, we randomly select one tree, where larger trees are chosen with higher probability: A tree with m' edges is chosen with weight $m'-m+1$. Starting with a random edge, we build a subtree from this tree by randomly adding incident edges to the subtree, until the subtree has size m edges. The root of the decoy tree is assigned the same molecular formula as the root of the target tree. We then label the edges and remaining nodes of the decoy tree: We randomly choose a loss from the target database, respecting multiplicities. So, whereas the structure of the tree and the succession of losses is random, the losses of a decoy fragmentation tree have the same "occurrence pattern" as those in the target database. The label for the target node of this edge is defined by subtracting the chosen loss from the label of its source node. In case the resulting molecular formula is invalid (the loss is not a sub-formula of the source node molecular formula), a new loss is selected. If no loss that would result in a valid formula exists, the whole tree is discarded, and the tree generation is restarted from scratch.

From this construction, we may assume that spurious hits in the target database and hits in the decoy database are equally likely: The decoy FTs are similar to true FTs with respect to size, tree topology, losses, and molecular formula of the parent compound. We also assume that hits in the decoy database are never "true" hits: It is extremely unlikely to construct a tree which, by chance, is also an element of the target database, or is the FT that we are actually searching for.

We align our sample FT to every tree in the combined database, containing both target and decoy FTs, and sort the results with respect to score (fingerprint similarity). We report hits from the true database only. Assume we are given a *False Discovery Rate* (FDR) threshold t, such as $t = 30\,\%$. If the TOP $n_\mathrm{T}+n_\mathrm{D}$ in the combined search contains n_T hits from the target database and n_D hits from the decoy database, then we calculate a FDR of $n_\mathrm{D}/n_\mathrm{T}$ for this list. We search for the largest set of top hits with FDR $n_\mathrm{D}/n_\mathrm{T} \leq t$. For each hit, we compute the *q-value* as the smallest FDR under which this hit is reported in the output.

In case we search for a FT in the database where we did not exclude this FT, our method recovered the correct FT in all cases. More precisely, the similarity of a FT to itself, is highest among all FTs in the dataset. Finding a "known compound" in a database is not a complicated task, and could be also done using methods based on spectral comparison. But we report this result here to show that our method will also "find the knowns", not only the unknowns.

We want to evaluate our method for those cases where the compound is *not* found in the database. To this end, we pursue a *leave-one-out* evaluation: For each compound, we deliberately delete the corresponding FT from the database before searching for it. We then compute an alignment score against all remaining compounds (both targets and decoys) in our dataset. As usual, these values are normalized by perfect match score with exponent 0.5 and used as fingerprints. Pearson correlation between the fingerprints is calculated and used as final fingerprint similarity score. We sort compounds with respect to fingerprint similarity, and estimate the FDR as described above.

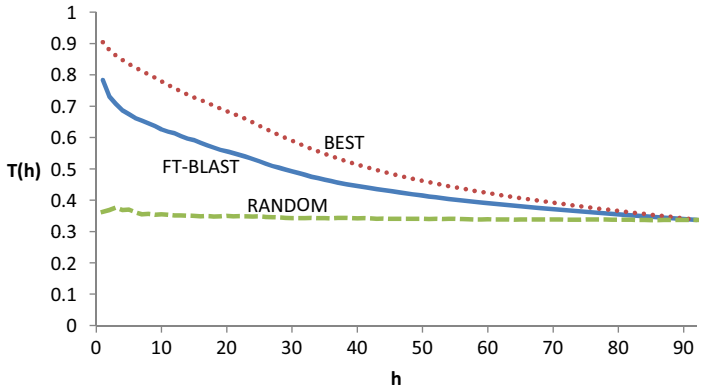
In Table 2 we report search results for the Orbitrap dataset with FDR threshold $t = 30\%$. One can see that the search results of glucosinolates, sugars and zeatins contain almost exclusively compounds of the respective group. Some benzopyrans receive several hits from their own and similar groups, whereas for other benzopyrans, no hits are found. Possibly, the corresponding spectra are of lower quality, or the chemical similarity to other benzopyrans is weak. Only few hits were found for the alkaloids. We attribute this to the fact that we have relatively few reference compounds available for the diverse class of alkaloids. We find almost no hits for amino acids, carboxylic acids, and lipids. Here, FTs were often too small to identify any hits.

To report the average Tanimoto structural similarity score of the hits returned by FT-BLAST, we calculated the Tanimoto score of the query compound and the hitlist entry. We then averaged either over all hits with an FDR below the threshold of 30% for the FT-BLAST approach, the five best scoring hits disregarding the FDR for the TOP 5 approach, or only those hits both within the FDR threshold and the TOP 5 for the combined approach. Now we average over all 93 queries (Orbitrap FTs with 1+ losses) to reach the final values of 0.76 for FT-BLAST, 0.67 for TOP 5, 0.78 for the combined approach. The TOP 5 approach is identical to Demuth *et al.* [4], the others are only adepted to the fact that an FDR estimation is available. Of course, this analysis is performed on the *leave-one-out* results.

Identical to Demuth *et al.* [4] we analyzed the Tanimoto scores $T(h)$ of the first h hits with h ranging from one to the number of compounds. Again, we did not use the FDR estimation but considered all scores obtained by a *leave-one-out* analysis. We then averaged over all compounds (Fig. 7). As Demuth *et al.* we compared these results with pseudo hitlists containing randomly ordered compounds (minimum value) and compounds arranged in descending order in accordance with the Tanimoto scores (upper limit). The average Tanimoto scores of our hitlists decrease from 0.78 ($h = 1$) to 0.34 ($h = 92$). The upper limit is between 0.90 ($h = 1$) and 0.34 ($h = 92$), and the minimum value is about 0.34 for all h. All three values converge to 0.34 as this is the average Tanimoto score of all pairwise different compounds. Compared to Figure 1 in [4], the correlation values of FT-Blast are considerably higher.

10 Poppy samples

Surface extracts of *P. nudicaule* were made using methanol: 1% acetic acid 2:1 mixture. The following organs of the plant were processed in different samples: petals, stamen with and without base, and stem. All extracts were directly infused using a Nanomate Triversa system (Advion, Ithaca, NY) on a Nanomate nanoelectrospray chip and analyzed on an Orbitrap XL (Thermo Fisher Scientific, Bremen, Germany). The instrument operated at 100 000 resolution and settings for tandem mass spectra acquisition as above. Measurements were conducted using both positive and negative mode. Precursor ions were manually selected based on ion intensities and fragmented using HCD with stepped collision energies of 0, 5, 10, 15, 20, 25,



Supplementary Fig. 7: Average Tanimoto scores $T(h)$ between query structures and the first h structures from hitlists obtained by FT-Blast without using FDR estimation (FT-BLAST), pseudo hitlists containing the database structures with maximum Tanimoto score to query structure (BEST) and randomly selected pseudo hitlists (RANDOM). All three analyses were performed on the Orbitrap dataset.

30, 40, and 50 arbitrary units. The data contained 489 non-empty fragmentation spectra of 89 potential compounds.

First, we tried to determine the molecular formulas of the unknown compounds, compare to Section 2 and [17]. To do so, it is necessary to measure both an isotope pattern and fragmentation pattern of the unknown compound. Unfortunately, isotope patterns had to be extracted from MS1 survey scans and were often of insufficient quality: In many cases, only the monoisotopic and the $M+1$ isotope could be detected as an extensive overlap of isotope peaks with peaks from other compounds occurs in the very rich direct-infusion MS spectra. In the mass range of up to a thousand dalton, this is usually insufficient to determine the molecular formula of an unknown compound from its isotope pattern [2]. To this end, we conservatively selected 29 poppy compounds where fragmentation tree analysis and isotope pattern analysis agreed upon the molecular formula of the unknown: For these compounds, the TOP 1 molecular formula of the combined analysis is among the TOP 5 molecular formulas of the isotope pattern analysis, and among the TOP 5 molecular formulas of the fragmentation pattern analysis (see Sec. 2).

For each of the 29 poppy compounds, we calculated fragmentation trees as described in Section 3. Afterwards, we performed an all-against-all alignment using the poppy FTs plus the Orbitrap FTs, and the corresponding decoy FTs. Scores were normalized and fingerprint similarities were calculated as described in Section 6. We then searched for the unknown compounds in the database of knowns (Orbitrap) using FT-BLAST described in Section 9. The FDR was again 30%. Results of this analysis are shown in Table 2.

We identified eight compounds in the sample by manual analysis of the spectra. FT-BLAST identified glutamine, arginine and quercetin by returning the respective references from the Orbitrap dataset as first hit. For the hexose (179 Da) galactose and mannose are the first hits. The unknown is most likely glucose, which was not in our reference, so FT-BLAST suggests other hexoses. Four other compounds were manually identified as alkaloids. The 328 Da feature is corytuberine, the 330 Da compound is reticuline. We consider the 370 Da feature as hydrogenated and hydroxylated palmatine. The 386 Da unknown is again hydrogenated and hydroxylated palmatine, but additionally with an methyl-group and a broken double bond. Unfortunately, our reference dataset only contained few alkaloids. Our

list of search results always contains the alkaloid laudanosine, which is most similar to the manual identifications. In case of corytuberine, chelidonine is always among the TOP3. These two alkaloids are extremely similar. The non-alkaloid hits are also reasonable: Phenylalanine is the biosynthetic precursor of these alkaloids. Benzopyrans and hydroxylated alkaloids only differ by the fact that the oxygen is not in the ring system but attached to it as hydroxy group, and anisic acid (the carboxylic acid occurring in all hit lists) is again very similar to phenylalanine.

We clustered the unknowns together with the reference measurements from Orbitrap, again using fingerprint similarities. We used all FTs with at least one loss to include as many reference compounds as possible. We computed all-against-all alignments for all compounds from the combined dataset poppy unknowns plus Orbitrap. We used hierarchical clustering as described in Section 7. Supplementary Figure 8 shows the clustering of the unknown compounds from poppy together with the Orbitrap reference dataset. All manually identified unknowns are grouped into their respective cluster. On top of the figure one can see the alkaloid cluster with four reference alkaloids and the four manually identified "unknowns". The 400 Da compound probably is also an alkaloid. Since it is located at the border of the cluster, more reference alkaloids are required for a reliable classification. Since the unknown at 229 Da falls into the amino acid cluster, we consider it at least strongly related with amino acids. The 277 Da molecule is probably a sugar, or contains a sugar moiety. With the limited reference data, it is not possible to assign a group to the 438 and 537 Da compounds, but we may assume that they are neither related to zeatins nor to glucosinolates, as no unknown falls into these well-separated clusters. Manual interpretation also failed to identify the compounds, NMR analysis is currently being performed. Additionally, our analysis correctly shows that a contamination with mass 338 Da, measured during a blank column run, is similar to the lipids. Database search and manual validation identified it as erucamide (PubChem CID 5365371), an additive originating from the plastic ware used for sample collection.

Results from the FT-BLAST and clustering analysis should be seen as strong hints towards a compound class. This can point towards unknowns of interest and simplify a downstream analysis, e.g. using NMR.

11 Peak counting score

Above, we have found a very strong correlation between FT similarity and chemical similarity. But how much of this correlation is due to the use of FTs, and what correlation can be reached with a "classical" shared peaks count? To this end, we correlate the normalized shared peaks count with chemical similarity. Given two fragmentation spectra, we count the number of peaks present in both spectra, respecting the mass accuracy of the measurement, then normalize this score. This score and variants thereof have been proposed repeatedly in the literature for searching tandem mass spectra of small compounds. For a fair comparison, we use the same subsets of compounds (with 1+, 3+, 5+, and 7+ losses) as above.

We tested different variants of the shared peak counting score. First, beside counting only similar peaks, also similar *parent losses* (mass differences to the parent peak) were counted. We tried various combinations of scoring peak masses and/or loss masses. Second, also considered the mass differences between two peaks, where two peaks with a lower mass difference receive a higher score. We tested a log likelihood-based scoring, based on the observation that mass differences in a well-calibrated mass spectrum are normally-distributed [2, 23]. Third, we include the intensities and masses of the matching peaks by

Supplementary Fig. 8: Clustering of the poppy and the Orbitrap datasets, FTs with 1+ losses. Colored compounds are known references. Many unknown compounds form a cluster together with several alkaloids (top of the figure). Other unknowns end up in amino acid or sugar clusters. The poppy sample most likely contained no glucosinolates and zeatins, as no unknowns can be found among these clusters.

scoring two matching peaks with $peakmass^3\sqrt{peakintensity}$ as suggested by Stein and Scott [12, 20]. The second and the third attempt did not improve the correlation with the chemical similarity score. The first attempt improved the correlation coefficient of the QSTAR dataset with a peak-loss-score of 0, but the overall performance was still best for the ordinary shared peak counting score. In the end, we normalized the shared peak counting score similar to the normalization of the FT alignment score by perfect match score using $c = 1.0$, and compute the fingerprints of the compounds as described in Section 6. Among all possibilities, we found that this normalization reached the best correlation with the chemical similarity score. Hence, it should be understood that while we report ordinary peak counting score results below, we were unable to reach consistently better results with any of the numerous variations of the peak counting score that we inspected.

It is noteworthy that correlations for the peak counting score (Supplementary Table 6) are very high, somewhat different from what has been reported in the literature. In fact, numerous variations of the peak counting score have been developed to cope with its

limitations, but these are often targeted at correlating raw spectra, not peak lists [12]. Pavlic *et al.* [16] and Oberacher *et al.* [14, 15] found that the unmodified peak counting score was inadequate for searching tandem MS databases. Also, counting shared peak is prone to artifact signals, see Figure 5 in [14]. It is therefore possible that the high correlations we reach for the peak counting score, are somewhat artificial.

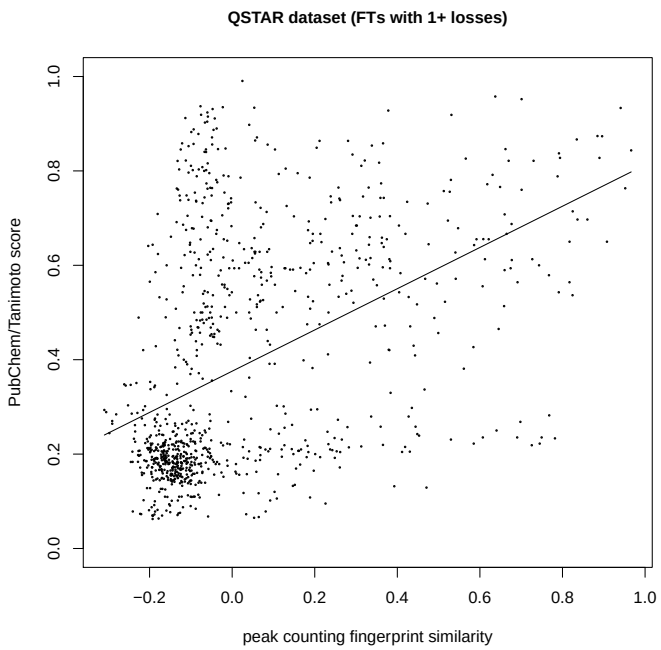
Still, comparing Supplementary Tables 5 and 6 we see that the correlation of the peak counting scores with chemical similarity (Tanimoto/PubChem) is — in all cases but two — weaker than for the tree alignment scores. It must be understood that, since correlation coefficients are rather high for the peak counting score, even small increases are significant improvements. This is particularly so as we have noted in Section 8 that even two different measures of chemical similarity (both Tanimoto scores) show a Pearson correlation of less than $r = +0.82$ on any of the data subsets. A particular large increase is observed for the QSTAR dataset, see Supplementary Table 6 and compare to Supplementary Table 5. Noteworthy is the large increase in Pearson correlation when analyzing the between-dataset: Whereas the peak counting score reaches a Pearson correlation coefficient of only $r = +0.38$ ($r^2 = 0.14$), Pearson correlation for the tree alignment fingerprint score is $r = +0.49$ ($r^2 = 0.24$). We believe this to be of particular importance, since it indicates the power of our tree alignment method to build up a database for identifying unknown metabolites measured on different instruments and with different settings.

		only compounds with			
Dataset	correlation method	1+ losses	3+ losses	5+ losses	7+ losses
Orbitrap	Pearson r	0.58	0.61	0.59	0.54
	Pearson r^2	0.34	0.37	0.35	0.29
	Spearman ρ	0.39	0.43	0.45	0.52
	Spearman ρ^2	0.15	0.18	0.20	0.27
MassBank	Pearson r	0.43	0.53	0.62	0.67
	Pearson r^2	0.18	0.28	0.38	0.45
	Spearman ρ	0.34	0.41	0.53	0.66
	Spearman ρ^2	0.12	0.17	0.28	0.44
QSTAR	Pearson r	0.45	0.44	0.45	0.43
	Pearson r^2	0.20	0.19	0.20	0.18
	Spearman ρ	0.51	0.50	0.45	0.42
	Spearman ρ^2	0.26	0.25	0.20	0.18
Between-dataset	Pearson r	0.38	0.42	0.48	0.52
	Pearson r^2	0.14	0.18	0.23	0.27
	Spearman ρ	0.33	0.36	0.39	0.42
	Spearman ρ^2	0.11	0.13	0.15	0.18

Supplementary Table 6: Correlation of chemical similarity (PubChem/Tanimoto) with the shared peak count, for all datasets and different restrictions on the number of losses. See Supplementary Table 5 for the number of compound pairs N.

References Supplementary Methods

1. P. Baldi and R. W. Benz. BLASTing small molecules–statistics and extreme statistics of chemical similarity scores. *Bioinformatics*, 24(13):i357–i365, 2008.
2. S. Böcker, M. Letzel, Z. Lipták, and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.



Supplementary Fig. 9: Correlation and regression line, QSTAR dataset: Shared peak counting fingerprint similarity (x-axis) plotted against PubChem/Tanimoto score (y-axis). To make results comparable with our above evaluation, we discarded all compounds from the Orbitrap dataset that resulted in FTs without any losses. Pearson correlation is $r = +0.45$ ($r^2 = 0.20$), Spearman correlation is $\rho = +0.51$ ($\rho^2 = 0.26$). Compare to Suppl. Figure 6.

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. P. D'haeseleer. How does gene expression clustering work? *Nat. Biotechnol.*, 23(12):1499–1501, 2005.
6. J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, 2002.
7. R. Guha, M. T. Howard, G. R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E. L. Willighagen. The Blue Obelisk: Interoperability in chemical informatics. *J. Chem. Inf. Model.*, 46(3):991–998, 2006.
8. 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.
9. H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, 45(7):703–714, 2010.
10. N. Jaitly, M. E. Monroe, V. A. Petyuk, T. R. W. Clauss, J. N. Adkins, and R. D. Smith. Robust algorithm for alignment of liquid chromatography-mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.*, 78(21):7397–7409, 2006.
11. T. Jiang, L. Wang, and K. Zhang. Alignment of trees: an alternative to tree edit. *Theor. Comput. Sci.*, 143(1):137–148, 1995.
12. I. Koo, X. Zhang, and S. Kim. Wavelet- and fourier-transform-based spectrum similarity approaches to compound identification in gas chromatography/mass spectrometry. *Anal. Chem.*, 83(14):5631–5638, 2011.
13. A. R. Leach and V. J. Gillet. *An Introduction to Chemoinformatics*. Springer, Berlin, Dordrecht, The Netherlands, 2005.

14. 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.
15. 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.
16. M. Pavlic, K. Libiseller, and H. Oberacher. Combined use of ESI-QqTOF-MS and ESI-QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. *Anal. Bioanal. Chem.*, 386(1):69–82, 2006.
17. F. Rasche, A. Svatoš, R. K. Maddula, C. Böttcher, and S. Böcker. Computing fragmentation trees from tandem mass spectrometry data. *Anal. Chem.*, 83:1243–1251, 2011.
18. D. J. Rogers and T. T. Tanimoto. A computer program for classifying plants. *Science*, 132(3434):1115–1118, 1960.
19. R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.
20. 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.
21. C. Steinbeck, C. Hoppe, S. Kuhn, M. Floris, R. Guha, and E. L. Willighagen. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des.*, 12(17):2111–2120, 2006.
22. Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 37(Web Server issue):W623–W633, 2009.
23. R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol. Cell. Proteomics.*, 6(3):377–381, 2007.

group	compound	PubChem ID	molecular formula	ion	monoisotopic mass	frag.method	collision energies	annotated NLs
Alkaloid	Berberine	2353	C20H18NO4+	[M+H]+	336.124	CID	35, 45	6
Alkaloid	Bicuculline	10237	C20H17NO6	[M+H]+	367.106	CID	35	25
Alkaloid	Chelidonine	10147	C20H19NO5	[M+H]+	353.126	CID	35, 45	12
Alkaloid	Cinchonine	8350	C19H22N2O	[M+H]+	294.173	CID	35, 45, 55	66
Alkaloid	Emetine	10219	C29H40N2O4	[M+H]+	480.299	CID	35, 45	62
Alkaloid	Harmane	5281404	C12H10N2	[M+H]+	182.084	CID	35, 45, 55	1
Alkaloid	Laudanosin	15548	C21H27NO4	[M+H]+	357.194	CID	35, 45, 55, 70	9
Amino acid	Alanine	602	C3H7NO2	[M-H]-	89.048	CID	5-90	0
Amino acid	Arginine	232	C6H14N4O2	[M+H]+	174.112	CID	5-80	7
Amino acid	Asparagine	236	C4H8N2O3	[M+H]+	132.053	CID	5-75	0
Amino acid	Aspartate	424	C4H7NO4	[M-H]-	133.038	CID	5-90	4
Amino acid	Cysteine	594	C3H7NO2S	[M-H]-	121.02	CID	5-90, 150	0
Amino acid	Cystine	595	C6H12N2O4S2	[M+H]+	240.024	CID	5-45	11
Amino acid	Glutamate	611	C5H9NO4	[M+H]+	147.053	CID	5-60	4
Amino acid	Glutamine	738	C5H10N2O3	[M-H]-	146.069	CID	5-90	5
Amino acid	Glycine	750	C2H5NO2	[M-H]-	75.032	HCD	5- 95	0
Amino acid	Isoleucine	791	C6H13NO2	[M+H]+	131.095	CID	5-60	2
Amino acid	Leucine	857	C6H13NO2	[M+H]+	131.095	CID	5-50	2
Amino acid	Methionine	876	C5H11NO2S	[M+H]+	149.051	CID	5-55	6
Amino acid	Phenylalanine	994	C9H11NO2	[M+H]+	165.079	CID	5-45	7
Amino acid	Proline	614	C5H9NO2	[M+H]+	115.063	CID	5-90	1
Amino acid	Serine	617	C3H7NO3	[M+H]+	105.043	HCD	5-75	2
Amino acid	Threonine	205	C4H9NO3	[M-H]-	119.058	CID	5-95, 9	2
Amino acid	Tryptophan	1148	C11H12N2O2	[M-H]-	204.09	HCD	5-95	6
Amino acid	Tyrosine	1153	C9H11NO3	[M+H]+	181.074	CID	5-45	7
Amino acid	Valine	1182	C5H11NO2	[M+H]+	117.079	CID	5-90	1
Anthocyanin	CID44256802	44256802	C47H55O27+	[M+H]+	1051.293	CID	5-45	9
Anthocyanin	CID44256805	44256805	C58H65O31+	[M+H]+	1257.351	HCD	5-45	18
Anthocyanin	Delphinidin-3-rutinoside	5492231	C27H31O16+	[M+H]+	611.161	HCD	5-45	18
Benzopyran	Armentoflavone	5281600	C30H18O10	[M+H]+	538.09	CID	35, 45, 55, 70	15
Benzopyran	Bergapten	2355	C12H8O4	[M+H]+	216.042	CID	35, 45, 55, 70	10
Benzopyran	BiochaninA	5280373	C16H12O5	[M+H]+	284.068	CID	35, 45, 55, 70	19
Benzopyran	Epicatechin	72276	C15H14O6	[M+H]+	290.079	CID	35, 45, 55, 70	8
Benzopyran	Genistein	5280961	C15H10O5	[M+H]+	270.053	CID	35, 45, 55	17
Benzopyran	Kaempferol	5280863	C15H10O6	[M+H]+	286.048	CID	35, 45, 55	26
Benzopyran	Quercetin	5280343	C15H10O7	[M+H]+	302.043	CID	35, 45, 55	23
Benzopyran	Rotenone	6758	C23H22O6	[M+H]+	394.142	CID	35, 45, 55, 70	8
Benzopyran	Rutin	5280805	C27H30O16	[M+H]+	610.153	CID	35, 45, 55, 70	9
Benzopyran	Vitexinrhamnoside	5282151	C27H30O14	[M+H]+	578.164	CID	35, 45, 55, 70	13
Benzopyran	Xanthohumol	639665	C21H22O5	[M+H]+	354.147	CID	35, 45, 55, 70	3
Carboxylic acid	Anisicacid	11370	C8H8O3	[M+H]+	152.047	CID	35, 45, 55, 70	1
Carboxylic acid	Indole-3-carboxylicAcid	69867	C9H7NO2	[M+H]+	161.048	CID	35, 45, 55, 70	2
Carboxylic acid	TrimethoxycinnamicAcid	735755	C12H14O5	[M+H]+	238.084	CID	35, 45, 55, 70	16
Glucosinolate	3-Hydroxypropyl-Glucosinolate	25245521	C10H17NO10S2	[M-H]-	375.029	HCD	5-90	9
Glucosinolate	3-Methylthiopropyl-Glucosinolate	25244538	C11H19NO9S3	[M-H]-	405.022	HCD	5-90	13
Glucosinolate	4-Methoxy-3-indolylmethyl glucosinolate	656562	C17H20N2O10S2	[M-H]-	476.056	HCD	5-90	19
Glucosinolate	7-Methylthioheptyl glucosinolate	44237368	C15H27NO9S3	[M-H]-	461.085	HCD	5-90	18
Glucosinolate	8-Methylthiooctyl glucosinolate	44237373	C16H29NO9S3	[M-H]-	475.1	HCD	5, 15-55, 65-90	21
Glucosinolate	Glucoalyssin	656523	C13H25NO10S3	[M-H]-	451.064	HCD	5, 15-50, 60	4
Glucosinolate	Glucoerucin	656538	C12H21NO9S3	[M-H]-	419.038	HCD	5-90	19
Glucosinolate	Glucohirsutin	44237257	C16H29NO10S3	[M-H]-	491.095	HCD	5-90	24
Glucosinolate	Glucoibarin	44237203	C15H27NO10S3	[M-H]-	477.08	HCD	5-90	28
Glucosinolate	Glucoiberin	9548621	C11H19NO10S3	[M-H]-	421.017	HCD	55-90	30
Glucosinolate	Glucomalcommin	25244201	C17H21NO11S2	[M-H]-	479.056	HCD	5-90	25
Glucosinolate	Glucoraphanin	9548633	C12H21NO10S3	[M-H]-	435.033	HCD	5-90	8
Glucosinolate	Glucoraphenin	6443008	C12H21NO11S3	[M-H]-	451.028	HCD	5-90	16
Glucosinolate	Indolylmethyl glucosinolate	25244590	C16H18N2O9S2	[M-H]-	446.045	HCD	5-90	22
Lipid	DErySphinganine	91486	C18H39NO2	[M-H]-	301.298	CID	25	12
Lipid	DErySphingosine	5280335	C18H37NO2	[M+H]+	299.282	CID	10	1
Lipid	Phosphatidylcholine	129900	C25H54NO6P	[M+H]+	495.369	HCD	30	3
Lipid	Phosphatidylethanolamine	46891780	C39H74NO8P	[M-H]-	715.515	CID	20	6
Sugar	Cellobiose	294	C12H22O11	[M+H]+	342.116	HCD	4	10
Sugar	DP5		C30H52O26	[M+Na]+	828.275	HCD	45	16
Sugar	DP7		C42H72O36	[M+H]+	1152.38	HCD	12	17
Sugar	Fucose	17106	C6H12O5	[M+Na]+	164.068	CID	46	2
Sugar	Galactose	6036	C6H12O6	[M+NH4]+	180.063	HCD	12	4
Sugar	Gentiobiose	441422	C12H22O11	[M+Na]+	342.116	CID	20	6
Sugar	Lactose	6134	C12H22O11	[M+H]+	342.116	HCD	4	10
Sugar	Mannitol	6251	C6H14O6	[M+H]+	182.079	HCD	20	12
Sugar	Mannose	18950	C6H12O6	[M+H]+	180.063	CID	15	6
Sugar	Rhamnose	19233	C6H12O5	[M+Na]+	164.068	CID	46	2
Sugar	Sorbitol	5780	C6H14O6	[M+H]+	182.079	CID	20	14
Sugar	Trehalose	7427	C12H22O11	[M+Na]+	342.116	CID	20	2
Zeatin	Cis-Zeatin	449093	C10H13N5O	[M+H]+	219.112	CID	44	7
Zeatin	Cis-Zeatin-9-glucoside	9842892	C16H23N5O6	[M+H]+	381.165	CID	17	5
Zeatin	Cis-Zeatin-o-glucoside	25244165	C16H23N5O6	[M+H]+	381.165	CID	19	6
Zeatin	Cis-Zeatin-riboside	6440982	C15H21N5O5	[M+H]+	351.154	CID	11	4
Zeatin	Cis-Zeatin-riboside-O-glucoside	11713250	C21H31N5O10	[M+H]+	513.207	CID	20	4
Zeatin	D5-Cis-Zeatin-riboside	6440982	C15H21N5O5	[M+H]+	351.154	CID	15	15
Zeatin	D5-Trans-Zeatin	449093	C10D5H8N5O	[M+H]+	224.143	CID	15	8
Zeatin	D5-Trans-Zeatin-7-glucoside		C16D5H18N5O6	[M+H]+	386.196	CID	14	8
Zeatin	D5-Trans-Zeatin-9-glucoside	9842892	C16D5H18N5O6	[M+H]+	386.196	CID	14	10
Zeatin	D5-Trans-Zeatin-riboside	6440982	C15H21N5O5	[M+H]+	351.154	CID	13	8
Zeatin	D5-Trans-Zeatin-riboside-o-glucoside	11713250	C21H31N5O10	[M+H]+	513.207	CID	23	15
Zeatin	D6-isopentenyl-Adenine		C10D6H7N5	[M+H]+	209.155	CID	27	4
Zeatin	D6-isopentenyl-Adenine-7-glucoside	330023	C16D6H17N5O5	[M+H]+	371.208	CID	30	1
Zeatin	D6-isopentenyl-Adenine-9-glucoside	23197432	C16D6H17N5O5	[M+H]+	371.208	CID	15	6
Zeatin	D6-isopentenyl-Adenosine	24405	C15D6H15N5O4	[M+H]+	341.197	CID	22	4
Zeatin	Isopentenyl-Adenine		C10H13N5	[M+H]+	203.117	CID	35	2
Zeatin	Isopentenyl-Adenine-7-glucoside	330023	C16H23N5O5	[M+H]+	365.17	CID	14	4
Zeatin	Isopentenyl-Adenine-9-glucoside	23197432	C16H23N5O5	[M+H]+	365.17	CID	14	5
Zeatin	Isopentenyl-Adenosine	24405	C15H21N5O4	[M+H]+	335.159	CID	13	3
Zeatin	Trans-Zeatin	449093	C10H13N5O	[M+H]+	219.112	CID	47	6
Zeatin	Trans-Zeatin-9-glucoside	9842892	C16H23N5O6	[M+H]+	381.165	CID	28	5
Zeatin	Trans-Zeatin-o-glucoside	25244165	C16H23N5O6	[M+H]+	381.165	CID	28	9
Zeatin	Trans-Zeatin-riboside	6440982	C15H21N5O5	[M+H]+	351.154	CID	24	1
Zeatin	Trans-Zeatin-riboside-O-glucoside	11713250	C21H31N5O10	[M+H]+	513.207	CID	12	5

Supplementary Table 7: Compound list for the Orbitrap dataset: Compound class, compound name, PubChem ID, molecular formula, ion type, monoisotopic mass (Da), fragmentation technique, collision energies, and number of annotated losses (edges) in hypothetical FTs. Collision energies are given in electron volt for CID and arbitrary units for HCD fragmentation. If a range is given, we used a step size of 5 units within this range. Compounds with less than three (seven) annotated losses are colored red (yellow).

group	compound	PubChem ID	molecular formula	monoisotopic mass	collision energies	annotated NLs
Aldehyde	1-Methoxy-3-carbaldehyde	398554	C10H9NO2	175.063	Ramp 5-60	2
Aldehyde	4-Hydroxy-3-methoxycinnamaldehyde	5280536	C10H10O3	178.063	Ramp 5-60	2
Aldehyde	Indole-3-acetaldehyde	800	C10H9NO	159.068	Ramp 5-60	2
Aldehyde	Indole-3-carboxyaldehyde	10256	C9H7NO	145.053	30, Ramp 5-60	6
Aldehyde	Syringaldehyde	8655	C9H10O4	182.058	Ramp 5-60	3
Amino acid	1-Aminocyclopropane-1-carboxylic_acid	535	C4H7NO2	101.048	Ramp 5-60	0
Amino acid	2-Aminoisobutyric_acid	6119	C4H9NO2	103.063	Ramp 5-60	0
Amino acid	3-Hydroxy-DL-kynurenine	89	C10H12N2O4	224.08	Ramp 5-60	6
Amino acid	3-Methyl-L-histidine	64969	C7H11N3O2	169.085	Ramp 5-60	3
Amino acid	5-Aminovaleric_acid	138	C5H11NO2	117.079	Ramp 5-60	0
Amino acid	Alpha-Methyl-DL-histidine	4396761	C7H11N3O2	169.085	Ramp 5-60	7
Amino acid	Alpha-Methyl-DL-serine	439656	C4H9NO3	119.058	Ramp 5-60	1
Amino acid	Carbamoyl-DL-aspartic_acid	93072	C5H8N2O5	176.043	Ramp 5-60	3
Amino acid	Creatine	586	C4H9N3O2	131.069	Ramp 5-60	1
Amino acid	Cystathionine	834	C7H14N2O4S	222.067	Ramp 5-60	2
Amino acid	D-Alloisoleucine	94206	C6H13NO2	131.095	Ramp 5-60	0
Amino acid	D-beta-homophenylalanine	102530	C10H13NO2	179.095	Ramp 5-60	2
Amino acid	D-beta-homoserine	779	C4H9NO3	119.058	Ramp 5-60	4
Amino acid	Delta-Aminolevulinic_acid	137	C5H9NO3	131.058	Ramp 5-60	1
Amino acid	DL-2-Aminobutyric_acid	80283	C4H9NO2	103.063	Ramp 5-60	0
Amino acid	DL-5-Hydroxylysine	1029	C6H14N2O3	162.1	Ramp 5-60	3
Amino acid	DL-alpha-epsilon-Diaminopimelic_acid	865	C7H14N2O4	190.095	Ramp 5-60	4
Amino acid	DL-threo-beta-Methylaspartic_acid	852	C5H9NO4	147.053	Ramp 5-60	3
Amino acid	D-Pantothenic_acid	6613	C9H17NO5	219.111	Ramp 5-60	4
Amino acid	Folic_acid	6037	C19H19N7O6	441.14	Ramp 5-60	4
Amino acid	Glutathione_(oxidized_form)	65359	C20H32N6O12S2	612.152	Ramp 5-60	13
Amino acid	Glycocyamine	763	C3H7N3O2	117.054	Ramp 5-60	1
Amino acid	Glycyl-L-proline	3013625	C7H12N2O3	172.085	Ramp 5-60	3
Amino acid	Gly-Gly	11163	C4H8N2O3	132.053	Ramp 5-60	2
Amino acid	L-(-)-Phenylalanine	6140	C9H11NO2	165.079	Ramp 5-60	4
Amino acid	L(+)-Arginine	6322	C6H14N4O2	174.112	Ramp 5-60	1
Amino acid	L-(+)-Lysine	5962	C6H14N2O2	146.106	Ramp 5-60	0
Amino acid	L-2-Aminobutyric_acid	80283	C4H9NO2	103.063	Ramp 5-60	0
Amino acid	L-allo-threonine	99289	C4H9NO3	119.058	Ramp 5-60	1
Amino acid	L-Anserine	112072	C10H16N4O3	240.122	Ramp 5-60	3
Amino acid	L-Arginine	6322	C6H14N4O2	174.112	Ramp 5-60	1
Amino acid	L-beta-Homoisoleucine	16211048	C7H15NO2	145.11	Ramp 5-60	0
Amino acid	L-beta-homoleucine	2761525	C7H15NO2	145.11	Ramp 5-60	0
Amino acid	L-beta-homolysine	2761529	C7H16N2O2	160.121	Ramp 5-60	0
Amino acid	L-beta-homomethionine	5706673	C6H13NO2S	163.067	Ramp 5-60	2
Amino acid	L-beta-Homophenylalanine	2761537	C10H13NO2	179.095	Ramp 5-60	2
Amino acid	L-beta-homoproline	2761541	C6H11NO2	129.079	Ramp 5-60	0
Amino acid	L-beta-homoserine	1502076	C4H9NO3	119.058	Ramp 5-60	4
Amino acid	L-beta-homothreonine	5706676	C5H11NO3	133.074	Ramp 5-60	3
Amino acid	L-beta-homotryptophan	2761550	C12H14N2O2	218.106	Ramp 5-60	3
Amino acid	L-beta-homotyrosine	2761554	C10H13NO3	195.09	Ramp 5-60	2
Amino acid	L-beta-homovaline	2761558	C6H13NO2	131.095	Ramp 5-60	1
Amino acid	L-Carnosine	439224	C9H14N4O3	226.107	Ramp 5-60	5
Amino acid	L-Citrulline	9750	C6H13N3O3	175.096	Ramp 5-60	1
Amino acid	L-Ethionine	25674	C6H13NO2S	163.067	Ramp 5-60	1
Amino acid	Leucylleucyltyrosine	88513	C21H33N3O5	407.242	Ramp 5-60	6
Amino acid	Leupeptin	439527	C20H38N6O4	426.295	Ramp 5-60	4
Amino acid	L-Glutamic_acid	33032	C5H9NO4	147.053	Ramp 5-60	2
Amino acid	L-Histidine	6274	C6H9N3O2	155.069	Ramp 5-60	8
Amino acid	L-Homocarnosine	89235	C10H16N4O3	240.122	Ramp 5-60	3
Amino acid	L-Homoserine	12647	C4H9NO3	119.058	Ramp 5-60	3
Amino acid	L-Leucine	6106	C6H13NO2	131.095	Ramp 5-60	1
Amino acid	L-Methionine_sulfone	445282	C5H11NO4S	181.041	Ramp 5-60	2
Amino acid	L-Norleucine	21236	C6H13NO2	131.095	Ramp 5-60	1
Amino acid	L-Norvaline	65098	C5H11NO2	117.079	Ramp 5-60	0
Amino acid	L-Proline	145742	C5H9NO2	115.063	Ramp 5-60	0
Amino acid	L-saccharopine	160556	C11H20N2O6	276.132	Ramp 5-60	4
Amino acid	L-Threonine	6288	C4H9NO3	119.058	Ramp 5-60	1
Amino acid	L-Tryptophane	6305	C11H12N2O2	204.09	Ramp 5-60	4
Amino acid	L-Tyrosine	6057	C9H11NO3	181.074	Ramp 5-60	4
Amino acid	L-Valine	6287	C5H11NO2	117.079	Ramp 5-60	0
Amino acid	N-Acetyl-DL-aspartic_acid	65065	C6H9NO5	175.048	Ramp 5-60	6
Amino acid	N-Acetyl-DL-glutamic_acid	70914	C7H11NO5	189.064	Ramp 5-60	6
Amino acid	N-acetyl-DL-serine	352294	C5H9NO4	147.053	Ramp 5-60	2
Amino acid	N-Acetylglycine	10972	C4H7NO3	117.043	Ramp 5-60	2
Amino acid	N-alpha-Acetyl-L-ornithine	439232	C7H14N2O3	174.1	Ramp 5-60	2
Amino acid	N-Formyl-L-methionine	439750	C6H11NO3S	177.046	Ramp 5-60	3
Amino acid	N-N-Dimethylglycine	673	C4H9NO2	103.063	Ramp 5-60	0
Amino acid	N-Tigloylglycine	6441567	C7H11NO3	157.074	Ramp 5-60	4
Amino acid	O-Phospho-L-serine	68841	C3H8NO6P	185.009	Ramp 5-60	2
Amino acid	S-Adenosyl-L-homocysteine	439155	C14H20N6O5S	384.122	Ramp 5-60	1
Amino acid	S-Lactoylglutathione	440018	C13H21N3O8S	379.105	Ramp 5-60	18
Amino acid	S-Sulfocysteine	115015	C3H7NO5S2	200.977	Ramp 5-60	6
Benzimidazole	Thiabendazole	5430	C10H7N3S	201.036	Ramp 5-60	3
Bile acid	Cholate	221493	C24H40O5	408.288	30, Ramp 5-60	10
Bile acid	Deoxycholate	440355	C24H40O4	392.293	30, Ramp 5-60	6
Capsaicinoid	Capsaicin	1548943	C18H27NO3	305.199	Ramp 5-60	1
Capsaicinoid	Dihydrocapsaicin	107982	C18H29NO3	307.215	Ramp 5-60	1
Carboxylic acid	(-)-Citramalic_acid	439766	C5H8O5	148.037	Ramp 5-60	4
Carboxylic acid	(-)Shikimic_acid	8742	C7H10O5	174.053	Ramp 5-60	7
Carboxylic acid	(+-)-Alpha-Lipoic_acid	864	C8H14O2S2	206.044	Ramp 5-60	5
Carboxylic acid	(R)-(-)-mandelic_acid	11914	C8H8O3	152.047	Ramp 5-60	1
Carboxylic acid	(S)-(+)-Citramailc_acid	441696	C5H8O5	148.037	Ramp 5-60	3
Carboxylic acid	16-Hydroxyhexadecanoic_acid	10466	C16H32O3	272.235	Ramp 5-60	0
Carboxylic acid	1-O-b-D-glucopyranosyl_sinapate	5280406	C17H22O10	386.121	Ramp 5-60	10
Carboxylic acid	2-5-Dihydroxy_benzoic_acid	3469	C7H6O4	154.027	Ramp 5-60	2
Carboxylic acid	2-Aminoethylphosphonic_acid	339	C2H8NO3P	125.024	Ramp 5-60	2
Carboxylic acid	2-Hydroxyisobutyric_acid	11671	C4H8O3	104.047	30, Ramp 5-60	2
Carboxylic acid	2-Hydroxyisocaproic_acid	439960	C6H12O3	132.079	Ramp 5-60	2
Carboxylic acid	2-Isopropylmalic_acid	5280523	C7H12O5	176.068	Ramp 5-60	5
Carboxylic acid	2-Methylglutaric_Acid	12046	C6H10O4	146.058	Ramp 5-60	2
Carboxylic acid	2-Oxobutyrate	58	C4H6O3	102.032	Ramp 5-60	0
Carboxylic acid	2-Oxovaleric_acid	74563	C5H8O3	116.047	Ramp 5-60	0
Carboxylic acid	3-4-Dihydroxybenzoic_acid	72	C7H6O4	154.027	Ramp 5-60	2
Carboxylic acid	3-Guanidinopropionic_acid	67701	C4H9N3O2	131.069	Ramp 5-60	1
Carboxylic acid	3-Hydroxy-3-methylglutarate	1662	C6H10O5	162.053	Ramp 5-60	3
Carboxylic acid	3-Hydroxymandelic_acid	86957	C8H8O4	168.042	Ramp 5-60	2

Supplementary Table 8: Compound list for the MassBank dataset: Compound class, compound name, PubChem ID, molecular formula, monoisotopic mass (Da), collision energies (eV), and number of annotated losses (edges) in hypothetical FTs. The ion type of all compounds is $[M+H]^+$. Compounds with less than three (seven) annotated losses are colored red (yellow).

Carboxylic acid	3-Indoleacetic_acid	802	C10H9NO2	175.063	Ramp 5-60	2
Carboxylic acid	4-Coumaric_acid	637542	C9H8O3	164.047	30, Ramp 5-60	2
Carboxylic acid	4-Hydroxy-3-methoxycinnamic_acid	445858	C10H10O4	194.058	Ramp 5-60	3
Carboxylic acid	4-Hydroxy-benzoate	135	C7H6O3	138.032	Ramp 5-60	1
Carboxylic acid	6-Hydroxynicotinic_Acid	72924	C6H5NO3	139.027	Ramp 5-60	1
Carboxylic acid	Anthranilic_acid	227	C7H7NO2	137.048	Ramp 5-60	1
Carboxylic acid	Caffeic_acid	689043	C9H8O4	180.042	Ramp 5-60	2
Carboxylic acid	Cis-Aconitic_Acid	643757	C6H6O6	174.016	Ramp 5-60	3
Carboxylic acid	Citraconic_Acid	643798	C5H6O4	130.027	Ramp 5-60	1
Carboxylic acid	Citric_acid	311	C6H8O7	192.027	Ramp 5-60	6
Carboxylic acid	D-(-)-Quinic_acid	6508	C7H12O6	192.063	Ramp 5-60	1
Carboxylic acid	D(+)-Galacturonic_acid	439215	C6H10O7	194.043	Ramp 5-60	11
Carboxylic acid	D-(+)-Glyceric_acid	439194	C3H6O4	106.027	Ramp 5-60	2
Carboxylic acid	D-(+)-Malic_acid	92824	C4H6O5	134.022	Ramp 5-60	4
Carboxylic acid	D-Gluconic_acid	10690	C6H12O7	196.058	Ramp 5-60	8
Carboxylic acid	D-Glucuronic_acid	94715	C6H10O7	194.043	Ramp 5-60	10
Carboxylic acid	DL-2-Hydroxyvaleric_acid	98009	C5H10O3	118.063	Ramp 5-60	1
Carboxylic acid	DL-3-4-Dihydroxymandelic_acid	85782	C8H8O5	184.037	Ramp 5-60	2
Carboxylic acid	DL-3-Aminoisobutyric_acid	64956	C4H9NO2	103.063	Ramp 5-60	1
Carboxylic acid	DL-4-Hydroxy-3-methoxymandelic_acid	1245	C9H10O5	198.053	Ramp 5-60	1
Carboxylic acid	DL-beta-Aminobutyric_acid	2761506	C4H9NO2	103.063	Ramp 5-60	0
Carboxylic acid	DL-beta-Hydroxybutyric_acid	441	C4H8O3	104.047	Ramp 5-60	1
Carboxylic acid	DL-Glyceric_acid	439194	C3H6O4	106.027	Ramp 5-60	2
Carboxylic acid	DL-Lactic_acid	107689	C3H6O3	90.032	Ramp 5-60	0
Carboxylic acid	DL-mandelic_acid	1292	C8H8O3	152.047	Ramp 5-60	1
Carboxylic acid	DL-p-Hydroxyphenyllactic_acid	9378	C9H10O4	182.058	Ramp 5-60	5
Carboxylic acid	DL-Pipecolinic_acid	439227	C6H11NO2	129.079	Ramp 5-60	0
Carboxylic acid	D-tartaric_acid	439655	C4H6O6	150.016	Ramp 5-60	4
Carboxylic acid	Gamma-Linolenic_acid	5280933	C18H30O2	278.225	Ramp 5-60	1
Carboxylic acid	Gibberellin_A4	443457	C19H24O5	332.162	Ramp 5-60	8
Carboxylic acid	Glutaric_acid	743	C5H8O4	132.042	Ramp 5-60	2
Carboxylic acid	Homogentisic_acid	780	C8H8O4	168.042	Ramp 5-60	3
Carboxylic acid	Indole-3-carboxylic_acid	69867	C9H7NO2	161.048	Ramp 5-60	1
Carboxylic acid	Isoguvacine	3765	C6H9NO2	127.063	Ramp 5-60	1
Carboxylic acid	Isonicotinic_acid	5922	C6H5NO2	123.032	Ramp 5-60	1
Carboxylic acid	Itaconic_acid	811	C5H6O4	130.027	Ramp 5-60	1
Carboxylic acid	Kynurenic_acid	3845	C10H7NO3	189.043	Ramp 5-60	1
Carboxylic acid	L(+)-Tartaric_acid	444305	C4H6O6	150.016	Ramp 5-60	2
Carboxylic acid	L-2-Aminoadipic_Acid	92136	C6H11NO4	161.069	Ramp 5-60	3
Carboxylic acid	L-Pyroglutamic_acid	7405	C5H7NO3	129.043	Ramp 5-60	0
Carboxylic acid	Maleic_acid	444266	C4H4O4	116.011	Ramp 5-60	1
Carboxylic acid	Mesaconic_acid	638129	C5H6O4	130.027	Ramp 5-60	1
Carboxylic acid	Methylsuccinic_acid	10349	C5H8O4	132.042	30, Ramp 5-60	1
Carboxylic acid	Mucic_acid	3037582	C6H10O8	210.038	Ramp 5-60	5
Carboxylic acid	N-acetylneuraminic_acid	439197	C11H19NO9	309.106	Ramp 5-60	3
Carboxylic acid	Nicotinic_Acid	938	C6H5NO2	123.032	Ramp 5-60	1
Carboxylic acid	Orotic_acid	967	C5H4N2O4	156.017	Ramp 5-60	1
Carboxylic acid	Phosphoenolpyruvic_Acid	1005	C3H5O6P	167.982	Ramp 5-60	1
Carboxylic acid	Prostaglandin_E1	5280723	C20H34O5	354.241	Ramp 5-60	6
Carboxylic acid	Rosmarinic_acid	639655	C18H16O8	360.085	Ramp 5-60	8
Carboxylic acid	Sebacic_acid	5192	C10H18O4	202.121	Ramp 5-60	3
Carboxylic acid	Sinapic_acid	637775	C11H12O5	224.068	Ramp 5-60	10
Carboxylic acid	Sinapoyl_malate	11953815	C15H16O9	340.079	Ramp 5-60	12
Carboxylic acid	Succinic_acid	1110	C4H6O4	118.027	Ramp 5-60	2
Carboxylic acid	Trans-4-Hydroxy-L-proline	5810	C5H9NO3	131.058	Ramp 5-60	2
Carboxylic acid	Trans-Cinnamic_acid	444539	C9H8O2	148.052	Ramp 5-60	1
Carboxylic acid	Urocanic_acid	736715	C6H6N2O2	138.043	Ramp 5-60	1
Coumarin	4-Methylumbelliferone	5280567	C10H8O3	176.047	Ramp 5-60	5
Coumarin	6-7-Dihydroxycoumarin	5281416	C9H6O4	178.027	30, Ramp 5-60	19
Coumarin	7-Hydroxy-4-methylcoumarin	5280567	C10H8O3	176.047	30, Ramp 5-60	10
Coumarin	Daphnetin	5280569	C9H6O4	178.027	30, Ramp 5-60	12
Coumarin	Esculin	5281417	C15H16O9	340.079	Ramp 5-60	4
Coumarin	Scopoletin	5280460	C10H8O4	192.042	Ramp 5-60	4
Ethanolamine	O-Phosphorylethanolamine	1015	C2H8NO4P	141.019	Ramp 5-60	1
Flavonoid	(-)-Epicatechin	72276	C15H14O6	290.079	Ramp 5-60	25
Flavonoid	(-)-Riboflavin	493570	C17H20N4O6	376.138	Ramp 5-60	4
Flavonoid	(+)-Catechin	9064	C15H14O6	290.079	Ramp 5-60	13
Flavonoid	(+)-Epicatechin	182232	C15H14O6	290.079	Ramp 5-60	13
Flavonoid	7-Methylquercetin-3-Galactoside-6-Rhamnoside-3-Rhamnoside	44259338	C34H42O20	770.227	30, Ramp 5-60	4
Flavonoid	Apigenin	5280443	C15H10O5	270.053	Ramp 5-60	2
Flavonoid	Apigenin-7-O-glucoside	5280704	C21H20O10	432.106	Ramp 5-60	7
Flavonoid	Baicalin	64982	C21H18O11	446.085	Ramp 5-60	3
Flavonoid	Daidzein	5281708	C15H10O4	254.058	30, Ramp 5-60	18
Flavonoid	Daidzin	107971	C21H20O9	416.111	Ramp 5-60	10
Flavonoid	Datiscin	5883291	C27H30O15	594.158	30, Ramp 5-60	14
Flavonoid	Eriodictyol	440735	C15H12O6	288.063	Ramp 5-60	5
Flavonoid	Eriodictyol-7-O-glucoside	5319853	C21H22O11	450.116	Ramp 5-60	7
Flavonoid	Flavanomarein	101781	C21H22O11	450.116	Ramp 5-60	4
Flavonoid	Formononetin	5280378	C16H12O4	268.074	Ramp 5-60	7
Flavonoid	Fortunellin	5317385	C28H32O14	592.179	Ramp 5-60	2
Flavonoid	Gossypin	5281621	C21H20O13	480.09	Ramp 5-60	7
Flavonoid	Hesperidin	10621	C28H34O15	610.19	Ramp 5-60	5
Flavonoid	Homoorientin	114776	C21H20O11	448.101	Ramp 5-60	13
Flavonoid	Hyperoside	5281643	C21H20O12	464.095	Ramp 5-60	8
Flavonoid	Isorhamnetin	5281654	C16H12O7	316.058	Ramp 5-60	3
Flavonoid	Isorhamnetin-3-Galactoside-6-Rhamnoside	44259338	C28H32O16	624.169	30, Ramp 5-60	8
Flavonoid	Isorhamnetin-3-O-glucoside	5318645	C22H22O12	478.111	30, Ramp 5-60	13
Flavonoid	Isorhamnetin-3-O-rutinoside	5481663	C28H32O16	624.169	30, Ramp 5-60	8
Flavonoid	Kaempferide	5281666	C16H12O6	300.063	Ramp 5-60	10
Flavonoid	Kaempferol	5280863	C15H10O6	286.048	Ramp 5-60	3
Flavonoid	Kaempferol-3-7-O-bis-alpha-L-rhamnoside	5323562	C27H30O14	578.164	30, Ramp 5-60	10
Flavonoid	Kaempferol-3-Galactoside-6-Rhamnoside-3-Rhamnoside	5281693	C33H40O19	740.216	30, Ramp 5-60	4
Flavonoid	Kaempferol-3-Glucoside-2-p-coumaroyl	25245527	C30H26O13	594.137	Ramp 5-60	6
Flavonoid	Kaempferol-3-Glucoside-2-Rhamnoside-7-Rhamnoside	25202803	C33H40O19	740.216	30, Ramp 5-60	7
Flavonoid	Kaempferol-3-Glucoside-3-Rhamnoside	25202803	C27H30O15	594.158	Ramp 5-60	4
Flavonoid	Kaempferol-3-Glucoside-6-p-coumaroyl	5320686	C30H26O13	594.137	30, Ramp 5-60	11
Flavonoid	Kaempferol-3-Glucuronide	5318759	C21H18O12	462.08	Ramp 5-60	3
Flavonoid	Kaempferol-3-O-alpha-L-arabinoside	5481882	C20H18O10	418.09	Ramp 5-60	7
Flavonoid	Kaempferol-3-O-alpha-L-rhamnopyranosyl(1-2)-beta-D-glucopyranoside-7-O-alpha-L-rhamnopyranoside	44258837	C33H40O19	740.216	30, Ramp 5-60	8
Flavonoid	Kaempferol-3-O-alpha-L-rhamnoside	5316673	C21H20O10	432.106	Ramp 5-60	9
Flavonoid	Kaempferol-3-O-beta-D-galactoside-7-O-alpha-L-rhamnoside	5281693	C27H30O15	594.158	30, Ramp 5-60	13
Flavonoid	Kaempferol-3-O-beta-glucopyranosyl-7-O-alpha-rhamnopyranoside	25203808	C27H30O15	594.158	30, Ramp 5-60	11
Flavonoid	Kaempferol-3-O-glucoside	5282102	C21H20O11	448.101	30, Ramp 5-60	13

Supplementary Table 8: Compound list for the MassBank dataset (continued)

Flavonoid	Kaempferol-3-O-rutinoside	5318767	C27H30O15	594.158	30, Ramp 5-60	6
Flavonoid	Kaempferol-3-Rhamnoside-4-Rhamnoside-7-Rhamnoside	44259005	C33H40O18	724.221	Ramp 5-60	6
Flavonoid	Kaempferol-7-O-alpha-L-rhamnoside	5316673	C21H20O10	432.106	30, Ramp 5-60	28
Flavonoid	Kaempferol-7-O-neohesperidoside	5483905	C27H30O15	594.158	30, Ramp 5-60	3
Flavonoid	Linarin	5317025	C28H32O14	592.179	Ramp 5-60	2
Flavonoid	Luteolin	5280445	C15H10O6	286.048	30, Ramp 5-60	19
Flavonoid	Luteolin-3-7-di-O-glucoside	5490298	C27H30O16	610.153	Ramp 5-60	3
Flavonoid	Luteolin-4-O-glucoside	5319116	C21H20O11	448.101	Ramp 5-60	6
Flavonoid	Luteolin-7-O-glucoside	5280637	C21H20O11	448.101	Ramp 5-60	8
Flavonoid	Marein	6441269	C21H22O11	450.116	Ramp 5-60	8
Flavonoid	Maritimein	6450184	C21H20O11	448.101	Ramp 5-60	3
Flavonoid	Myricetin-3-Galactoside	5491408	C21H20O13	480.09	Ramp 5-60	11
Flavonoid	Myricetin-3-Rhamnoside	5281673	C21H20O12	464.095	Ramp 5-60	12
Flavonoid	Myricetin-3-Xyloside	5281673	C20H18O12	450.08	Ramp 5-60	9
Flavonoid	Myricitrin	5281673	C21H20O12	464.095	Ramp 5-60	11
Flavonoid	Naringenin-7-O-glucoside	92794	C21H22O10	434.121	Ramp 5-60	7
Flavonoid	Neodiosmin	44258230	C28H32O15	608.174	Ramp 5-60	2
Flavonoid	Ononin	442813	C22H22O9	430.126	30, Ramp 5-60	5
Flavonoid	Peltatoside	5484066	C26H28O16	596.138	30, Ramp 5-60	18
Flavonoid	Poncirin	442456	C28H34O14	594.195	30, Ramp 5-60	5
Flavonoid	Procyanidin_B1	11250133	C30H26O12	578.142	Ramp 5-60	15
Flavonoid	Procyanidin_B2	122738	C30H26O12	578.142	Ramp 5-60	16
Flavonoid	Puerarin	5281807	C21H20O9	416.111	Ramp 5-60	6
Flavonoid	Quercetin	5280343	C15H10O7	302.043	Ramp 5-60	8
Flavonoid	Quercetin-3-(6-malonyl)-Glucoside	5282159	C24H22O15	550.096	Ramp 5-60	8
Flavonoid	Quercetin-3-4-O-di-beta-glucopyranoside	5320835	C27H30O17	626.148	30, Ramp 5-60	9
Flavonoid	Quercetin-3-7-O-alpha-L-dirhamnopyranoside	44259217	C27H30O15	594.158	30, Ramp 5-60	10
Flavonoid	Quercetin-3-Arabinoside	5481224	C20H18O11	434.085	Ramp 5-60	8
Flavonoid	Quercetin-3-D-xyloside	5320863	C20H18O11	434.085	Ramp 5-60	9
Flavonoid	Quercetin-3-Glucuronide	5274585	C21H18O13	478.075	Ramp 5-60	8
Flavonoid	Quercetin-3-O-alpha-L-rhamnopyranoside	5280459	C21H20O11	448.101	Ramp 5-60	12
Flavonoid	Quercetin-3-O-alpha-L-rhamnopyranosyl(1-2)-beta-D-glucopyranoside-7-O-alpha-L-rhamnopyranoside	5489459	C33H40O20	756.211	30, Ramp 5-60	8
Flavonoid	Quercetin-3-O-beta-D-galactoside	5281643	C21H20O12	464.095	Ramp 5-60	9
Flavonoid	Quercetin-3-O-beta-glucopyranoside	5280804	C21H20O12	464.095	Ramp 5-60	6
Flavonoid	Quercetin-3-O-beta-glucopyranosyl-7-O-alpha-rhamnopyranoside	5280805	C27H30O16	610.153	30, Ramp 5-60	11
Flavonoid	Quercetin-3-O-glucose-6-acetate	5280804	C23H22O13	506.106	Ramp 5-60	8
Flavonoid	Quercetin-7-O-rhamnoside	5748601	C21H20O11	448.101	Ramp 5-60	9
Flavonoid	Rhamnetin	5281691	C16H12O7	316.058	Ramp 5-60	6
Flavonoid	Rhoifolin	5282150	C27H30O14	578.164	30, Ramp 5-60	3
Flavonoid	Robinin	5281693	C33H40O19	740.216	30, Ramp 5-60	7
Flavonoid	Spiraeoside	5320844	C21H20O12	464.095	Ramp 5-60	6
Flavonoid	Syringetin-3-O-galactoside	5321576	C23H24O13	508.122	30, Ramp 5-60	17
Flavonoid	Syringetin-3-O-glucoside	5321577	C23H24O13	508.122	Ramp 5-60	14
Flavonoid	Tiliroside	5320686	C30H26O13	594.137	30, Ramp 5-60	9
Flavonoid	Vitexin	5280441	C21H20O10	432.106	Ramp 5-60	4
Flavonoid	Vitexin-2-O-rhamnoside	5282151	C27H30O14	578.164	Ramp 5-60	5
Glucosinolate	4-Methylsulfinylbutyl_glucosinolate	9548634	C12H23NO10S3	437.048	Ramp 5-60	6
Glucosinolate	4-Methylthiobutyl_glucosinolate	9548895	C12H23NO9S3	421.053	Ramp 5-60	4
Glucosinolate	Sinigrin	6911854	C10H17NO9S2	359.034	Ramp 5-60	4
Indole	3-Indoxylsulfate	10258	C8H7NO4S	213.01	Ramp 5-60	3
Indole	Harmaline	5280951	C13H14N2O	214.111	Ramp 5-60	4
Isoprenoid	Glycyrrhizic_acid	14982	C42H62O16	822.404	30, Ramp 5-60	3
Isoprenoid	Glycyrrhizin	14982	C42H62O16	822.404	30, Ramp 5-60	3
Nucleotide	1-3-Dimethylurate	70346	C7H8N4O3	196.06	30, Ramp 5-60	10
Nucleotide	1-7-Dimethylxanthine	4687	C7H8N4O2	180.065	Ramp 5-60	5
Nucleotide	2-Deoxyadenosine-5-monophosphate	12599	C10H14N5O6P	331.068	Ramp 5-60	5
Nucleotide	2-Deoxycytidine	13711	C9H13N3O4	227.091	Ramp 5-60	3
Nucleotide	2-Deoxycytidine-5-diphosphate	150855	C9H15N3O10P2	387.023	Ramp 5-60	6
Nucleotide	2-Deoxyguanosine_5-monophosphate	65059	C10H14N5O7P	347.063	Ramp 5-60	4
Nucleotide	2-Deoxyguanosine-5-diphosphate	439220	C10H15N5O10P2	427.029	Ramp 5-60	2
Nucleotide	2-Deoxyinosine-5-monophosphate	91531	C10H13N4O7P	332.052	Ramp 5-60	6
Nucleotide	2-Deoxyuridine-5-monophosphate	65063	C9H13N2O8P	308.041	Ramp 5-60	5
Nucleotide	3-Hydroxypyridine	7971	C5H5NO	95.037	30, Ramp 5-60	1
Nucleotide	3-Methylxanthine	70639	C6H6N4O2	166.049	Ramp 5-60	5
Nucleotide	4-Pyridoxate	6723	C8H9NO4	183.053	Ramp 5-60	2
Nucleotide	5-Aminoimidazole-4-carboxamide-1-beta-D-ribofuranosyl_5-monophosphate	65110	C9H15N4O8P	338.063	Ramp 5-60	3
Nucleotide	5-Deoxy-5-Methylthioadenosine	439176	C11H15N5O3S	297.09	Ramp 5-60	2
Nucleotide	6-(Gamma-gamma-Dimethylallylamino)purine	92180	C10H13N5	203.117	Ramp 5-60	6
Nucleotide	6-(Gamma-gamma-Dimethylallylamino)purine_riboside	24405	C15H21N5O4	335.159	Ramp 5-60	4
Nucleotide	Adenine	190	C5H5N5	135.054	30, Ramp 5-60	4
Nucleotide	Adenosine	60961	C10H13N5O4	267.097	Ramp 5-60	2
Nucleotide	Adenosine_3-monophosphate	41211	C10H14N5O7P	347.063	Ramp 5-60	4
Nucleotide	Adenosine_5-diphosphate	6022	C10H15N5O10P2	427.029	Ramp 5-60	3
Nucleotide	Adenosine_5-diphospho-glucose	16500	C16H25N5O15P2	589.082	Ramp 5-60	9
Nucleotide	Adenosine_5-monophosphate	6083	C10H14N5O7P	347.063	Ramp 5-60	3
Nucleotide	Beta-Nicotinamide_adenine_dinucleotide	5893	C21H27N7O14P2	663.109	Ramp 5-60	10
Nucleotide	Cytidine	6175	C9H13N3O5	243.086	Ramp 5-60	4
Nucleotide	Cytidine_5-diphosphocholine	13804	C14H26N4O11P2	488.107	Ramp 5-60	8
Nucleotide	Cytidine-3-5-cyclicmonophosphate	19236	C9H12N3O7P	305.041	Ramp 5-60	6
Nucleotide	Cytidine-3-monophosphate	66535	C9H14N3O8P	323.052	Ramp 5-60	4
Nucleotide	Cytidine-5-diphosphate	6132	C9H15N3O11P2	403.018	Ramp 5-60	5
Nucleotide	Cytidine-5-monophosphate	6131	C9H14N3O8P	323.052	Ramp 5-60	3
Nucleotide	Guanine	764	C5H5N5O	151.049	Ramp 5-60	3
Nucleotide	Guanosine	6802	C10H13N5O5	283.092	Ramp 5-60	3
Nucleotide	Guanosine_5-diphosphate-D-mannose	18396	C16H25N5O16P2	605.077	Ramp 5-60	6
Nucleotide	Guanosine_5-diphospho-beta-L-fucose	10918995	C16H25N5O15P2	589.082	Ramp 5-60	9
Nucleotide	Guanosine_5-diphosphoglucose	439225	C16H25N5O16P2	605.077	Ramp 5-60	7
Nucleotide	Guanosine_5-monophosphate	6804	C10H14N5O8P	363.058	Ramp 5-60	5
Nucleotide	Guanosine-3-5-cyclic_monophosphate	24316	C10H12N5O7P	345.047	Ramp 5-60	6
Nucleotide	Inosine	6021	C10H12N4O5	268.081	Ramp 5-60	3
Nucleotide	Inosine-5-diphosphate	6831	C10H14N4O11P2	428.013	Ramp 5-60	7
Nucleotide	Inosine-5-monophosphate	8582	C10H13N4O8P	348.047	Ramp 5-60	6
Nucleotide	N-6-(delta-2-Isopentenyl)adenosine	24405	C15H21N5O4	335.159	Ramp 5-60	5
Nucleotide	Oxypurinol	4644	C5H4N4O2	152.033	Ramp 5-60	1
Nucleotide	Pyridoxal	1050	C8H9NO3	167.058	Ramp 5-60	3
Nucleotide	Pyridoxal_5-phosphate	1051	C8H10NO6P	247.025	Ramp 5-60	2
Nucleotide	Pyridoxamine	1052	C8H12N2O2	168.09	Ramp 5-60	9
Nucleotide	Pyridoxine	1054	C8H11NO3	169.074	Ramp 5-60	7
Nucleotide	Thiamine	1130	C12H17N4OS	265.112	Ramp 5-60	5
Nucleotide	Thymidine-5-diphosphate	164628	C10H16N2O11P2	402.023	Ramp 5-60	8
Nucleotide	Thymidine-5-monophosphate	9700	C10H15N2O8P	322.057	Ramp 5-60	5
Nucleotide	Thymine	1135	C5H6N2O2	126.043	Ramp 5-60	0
Nucleotide	Trans-Zeatin	449093	C10H13N5O	219.112	Ramp 5-60	8

Supplementary Table 8: Compound list for the MassBank dataset (continued)

Nucleotide	Trans-Zeatin-riboside	6440982	C15H21N5O5	351.154	Ramp 5-60	6
Nucleotide	UDP-beta-L-rhamnose	23724469	C15H24N2O16P2	550.06	Ramp 5-60	13
Nucleotide	UDP-Galactose	23724458	C15H24N2O17P2	566.055	Ramp 5-60	13
Nucleotide	UDP-xylose	23724459	C14H22N2O16P2	536.044	Ramp 5-60	15
Nucleotide	Uracil	1174	C4H4N2O2	112.027	Ramp 5-60	0
Nucleotide	Uridine	6029	C9H12N2O6	244.07	Ramp 5-60	5
Nucleotide	Uridine_5-diphosphate	6031	C9H14N2O12P2	404.002	Ramp 5-60	5
Nucleotide	Uridine_5-diphospho-D-glucose	8629	C15H24N2O17P2	566.055	Ramp 5-60	13
Nucleotide	Uridine_5-diphosphoglucuronic_acid	17473	C15H22N2O18P2	580.034	Ramp 5-60	14
Nucleotide	Uridine_5-diphospho-N-acetylgalactosamine	23724461	C17H27N3O17P2	607.082	30, Ramp 5-60	17
Nucleotide	Uridine_5-diphospho-N-acetylglucosamine	445675	C17H27N3O17P2	607.082	30, Ramp 5-60	17
Nucleotide	Uridine_5-monophosphate	6030	C9H13N2O9P	324.036	Ramp 5-60	4
Nucleotide	Xanthine	1188	C5H4N4O2	152.033	Ramp 5-60	1
Nucleotide	Xanthosine	64959	C10H12N4O6	284.076	Ramp 5-60	2
Nucleotide	Xanthosine-5-monophosphate	73323	C10H13N4O9P	364.042	Ramp 5-60	6
Organosulfonic acid	2-Mercaptoethanesulfonic_acid	598	C2H6O3S2	141.976	Ramp 5-60	2
Organosulfonic acid	Hypotaurine	107812	C2H7NO2S	109.02	Ramp 5-60	2
Organosulfonic acid	S-Sulforaphene	6433206	C6H9NOS2	175.013	Ramp 5-60	4
Penicillin	Piperacillin	6604563	C23H27N5O7S	517.163	Ramp 5-60	5
Phenol	4-Nitrophenol	980	C6H5NO3	139.027	Ramp 5-60	0
Phenol	4-Nitrophenyl_phosphate	378	C6H6NO6P	218.993	30, Ramp 5-60	1
Phenol	Catechol	289	C6H6O2	110.037	30, Ramp 5-60	3
Polyketide	Zearalenone	5281576	C18H22O5	318.147	Ramp 5-60	8
Stilbene	E-3-4-5-trihydroxy-3-glucopyranosylstilbene	5281712	C20H22O9	406.126	Ramp 5-60	5
Sugar	2-Deoxyribose-5-phosphate	439288	C5H11O7P	214.024	Ramp 5-60	2
Sugar	Alpha-D-(+)-mannose-1-phosphate	439279	C6H13O9P	260.03	Ramp 5-60	2
Sugar	Alpha-D-Galactose-1-phosphate	123912	C6H13O9P	260.03	Ramp 5-60	4
Sugar	Alpha-D-Glucose-1-6-diphosphate	82400	C6H14O12P2	339.996	Ramp 5-60	6
Sugar	Alpha-D-glucose-1-phosphate	439165	C6H13O9P	260.03	Ramp 5-60	4
Sugar	D(-)-Gulono-gamma-lactone	165105	C6H10O6	178.048	Ramp 5-60	9
Sugar	D-(+)-Cellotriose	440950	C18H32O16	504.169	Ramp 5-60	22
Sugar	D-(+)-Melezitose	92817	C18H32O16	504.169	Ramp 5-60	12
Sugar	D-(+)-Raffinose	439242	C18H32O16	504.169	Ramp 5-60	9
Sugar	D-(+)-Trehalose	7427	C12H22O11	342.116	Ramp 5-60	10
Sugar	D-Arabinose-5-phosphate	230	C5H11O8P	230.019	Ramp 5-60	3
Sugar	D-Erythrose-4-phosphate	697	C4H9O7P	200.009	Ramp 5-60	3
Sugar	D-Fructose-6-phosphate	439160	C6H13O9P	260.03	Ramp 5-60	2
Sugar	D-Glucosamine-6-phosphate	439217	C6H14NO8P	259.046	Ramp 5-60	3
Sugar	D-Glucose-6-phosphate	5958	C6H13O9P	260.03	Ramp 5-60	3
Sugar	D-Mannose-6-phosphate	65127	C6H13O9P	260.03	Ramp 5-60	4
Sugar	D-Ribose-5-phosphate	439167	C5H11O8P	230.019	Ramp 5-60	3
Sugar	D-Ribulose-5-phosphate	439184	C5H11O8P	230.019	Ramp 5-60	2
Sugar	L-(+)-Rhamnose	25310	C6H12O5	164.068	Ramp 5-60	0
Sugar	Maltotriose	439586	C18H32O16	504.169	Ramp 5-60	25
Sugar	Palatinose	439559	C12H22O11	342.116	Ramp 5-60	14
Sugar	Sucrose	5988	C12H22O11	342.116	Ramp 5-60	11
Sugar alcohol	1-2-Dilauroyl-sn-Glycero-3-Phosphate	9547171	C27H53O8P	536.348	Ramp 5-60	5
Sugar alcohol	1-Lauroyl-2-Hydroxy-sn-Glycero-3-Phosphocholine	460605	C20H42NO7P	439.27	Ramp 5-60	1
Sugar alcohol	1-Myristoyl-2-Hydroxy-sn-Glycero-3-Phosphate	9547180	C17H35O7P	382.212	Ramp 5-60	3
Sugar alcohol	D-(-)-Mannitol	6251	C6H14O6	182.079	Ramp 5-60	9
Sugar alcohol	DL-Glyceraldehyde_3-phosphate	729	C3H7O6P	169.998	Ramp 5-60	3
Sugar alcohol	D-Sorbitol	5780	C6H14O6	182.079	Ramp 5-60	9
Sugar alcohol	D-Sorbitol-6-phosphate	152306	C6H15O9P	262.045	Ramp 5-60	2
Sugar alcohol	Dulcitol	11850	C6H14O6	182.079	Ramp 5-60	11
Sugar alcohol	Galactinol	439451	C12H22O11	342.116	Ramp 5-60	14
Sugar alcohol	Glycerol-2-phosphate	2526	C3H9O6P	172.014	Ramp 5-60	2
Sugar alcohol	L-Iditol	5460044	C6H14O6	182.079	Ramp 5-60	5
Sugar alcohol	Maltitol	493591	C12H24O11	344.132	Ramp 5-60	10
Sugar alcohol	Rac-Glycerol_3-phosphoate	439162	C3H9O6P	172.014	Ramp 5-60	2
	2-Hydroxyphenylacetic_acid	11970	C8H8O3	152.047	Ramp 5-60	1
	Hinokitiol	3611	C10H12O2	164.084	30, Ramp 5-60	0
	Methyl_Salicylate	4133	C8H8O3	152.047	Ramp 5-60	1

Supplementary Table 8: Compound list for the MassBank dataset (continued)

group	compound	molecular formula	monoisotopic mass	collision energies	annotated NLs
Amine	Dopamine	C8H11NO2	153.079	10, 20, 30, 40, 50	19
Amine	Spermidine	C7H19N3	145.158	15, 25, 35, 45	17
Amine	Spermine	C10H26N4	202.216	15, 25, 35, 45	12
Amine	Tyramine	C8H12NO+	138.092	15, 20, 30, 40, 50	23
Amino acid	Alanine	C3H7NO2	89.048	10	1
Amino acid	Arginine	C6H14N4O2	174.112	20, 25, 30	15
Amino acid	Asparagine	C4H8N2O3	132.053	10, 15, 20, 30, 40	15
Amino acid	Aspartic acid	C4H7NO4	133.038	10, 15, 20, 30	8
Amino acid	Citrulline	C6H13N3O3	175.096	10, 15, 20, 25, 30	22
Amino acid	Cysteine	C3H8NO2S+	122.028	10, 15, 20, 30	7
Amino acid	Cystine	C6H12N2O4S2	240.024	10, 15, 20, 30, 40	40
Amino acid	Glutamic acid	C5H9NO4	147.053	10, 15, 20, 30	7
Amino acid	Glutamine	C5H10N2O3	146.069	10, 15, 20, 30	8
Amino acid	Histidine	C6H9N3O2	155.069	15, 25, 35, 45	18
Amino acid	Isoleucine	C6H13NO2	131.095	10, 15, 25, 40	18
Amino acid	Leucine	C6H13NO2	131.095	15, 25, 40	9
Amino acid	Lysine	C6H14N2O2	146.106	10, 15, 20, 30, 40	23
Amino acid	Methionine	C5H11NO2S	149.051	10, 15, 20, 30	10
Amino acid	Phenylalanine	C9H11NO2	165.079	15, 25, 40	15
Amino acid	Proline	C5H9NO2	115.063	10, 15, 55	7
Amino acid	Serine	C3H7NO3	105.043	10, 15, 20, 30	5
Amino acid	Threonine	C4H9NO3	119.058	10, 15, 20, 30	6
Amino acid	Tryptophane	C11H12N2O2	204.09	15, 25, 40, 55	38
Amino acid	Tyrosine	C9H11NO3	181.074	10, 15, 25, 30, 40	22
Amino acid	Valine	C5H11NO2	117.079	10, 25, 40, 55	15
Carboxylic acid	6-Aminocapronic acid	C6H13NO2	131.095	15, 20, 30, 40	29
Choline	3-(4-Hexosyloxyphenyl)propanoyl choline	C20H32NO8+	414.213	25, 40, 55	4
Choline	4-Coumaroyl choline	C14H20NO3+	250.144	15, 25, 40	4
Choline	4-Hexosylferuloyl choline	C21H32NO9+	442.208	15, 25, 40, 55	5
Choline	4-Hexosyloxybenzoyl choline	C18H28NO8+	386.181	15, 25, 40, 55, 90	5
Choline	4-Hexosyloxycinnamoyl choline	C20H30NO8+	412.197	25, 40, 55	4
Choline	4-Hexosylvanilloyl choline	C19H30NO9+	416.192	15, 25, 40, 55, 70	3
Choline	4-Hydroxybenzoyl choline	C12H18NO3+	224.129	15, 25, 40, 55	4
Choline	5-Hydroxyferuloyl choline	C15H22NO5+	296.15	15, 25, 40, 55	11
Choline	Acetyl choline	C7H16NO2+	146.118	20	3
Choline	Benzoyl choline	C12H18NO2+	208.134	15, 25, 40, 55	3
Choline	Cafeoyl choline	C14H20NO4+	266.139	15, 25, 40, 55	8
Choline	Choline with Arylglycerol-arylether backbone	C23H32NO8+	450.213	50	3
Choline	Cinnamoyl choline	C14H20NO2+	234.149	15, 25, 40, 55	3
Choline	Feruloyl choline	C15H22NO4+	280.155	15, 25, 40	7
Choline	Nicotinic acid choline ester	C11H17N2O2+	209.129	15, 25, 40, 55	3
Choline	Sinapoyl choline	C16H24NO5+	310.165	15, 25, 40	4
Choline	Syringoyl choline	C14H22NO5+	284.15	50	19
Choline	Vanilloyl choline	C13H20NO4+	254.139	15, 25, 40, 55	10

Supplementary Table 9: Compound list for the QSTAR dataset: Compound class, compound name, molecular formula, monoisotopic mass (Da), collision energies (eV), and number of annotated losses (edges) in hypothetical FTs. The ion type of all compounds is $[M\text{-}H]^-$. Compounds with less than three (seven) annotated losses are colored red (yellow).

Supplementary Fig. 10: All FTs for the Orbitrap dataset in separate file.

Supplementary Fig. 11: All FTs for the MassBank dataset in separate file.

Supplementary Fig. 12: All FTs for the QSTAR dataset in separate file.