Identifying the unknowns by aligning fragmentation trees

Florian Rasche¹, Kerstin Scheubert¹, Franziska Hufsky¹,⁴, Thomas Zichner², Marco Kai³, Aleš Svatůš³ & Sebastian Böcker¹

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

Mass spectrometry allows sensitive, automated and high-throughput analysis of small molecules such as metabolites. 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 fragmentation trees of the compounds. 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 also describe a tool for searching a database for compounds with fragmentation pattern similar to an unknown sample compound. We apply this tool to the classification of metabolites from Icelandic poppy. Our method allows fully automated computational identification of small molecules that cannot be found in any database, enabling high-throughput screening.

Metabolomics, the identification and quantification of all low molecular weight compounds in a cell or tissue, complements investigation of the genome, transcriptome, and proteome of plants, fungi and bacteria¹,². The vast majority of metabolites remain unknown, however¹. This is particularly so for plants and many bacteria that synthesize large numbers of secondary metabolites. We cannot normally deduce the structure of these metabolites from genome sequences, as is done with proteins. Secondary metabolites play an important role in natural product research³,⁴, and many top-selling drugs and antibiotics are derived from nature⁵.

Mass spectrometry (MS) is a key technology for detecting and identifying metabolites and other small molecules¹,²,⁶. It is orders of magnitude more sensitive than nuclear magnetic resonance (NMR). Several analytical techniques have been developed, most notably gas chromatography MS (GC-MS) and liquid chromatography MS (LC-MS). We can analyze thermally unstable metabolites using LC-coupled tandem MS, where the gentle ionization results in minimal fragmentation of the adduct ions formed. 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 proteomics⁷,⁸. Computational methods for analyzing tandem MS data of metabolites were developed as part of the DENDRAL project⁹. Unfortunately, the project failed to achieve its major objective of automated structure elucidation using MS data and research was discontinued. Forty-six years later, the computational analysis of such data is still in its infancy. This is one of the major technological hurdles in current metabolomics¹⁰. 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 data¹¹, as well as
searching in spectral libraries, are major problems. Apart from a few pioneering studies (e.g. ref.), there are few computational methods for the automated analysis of tandem MS data from small molecules. In contrast, there are methods for de novo sequencing of linear or cyclic non-ribosomal peptides, but these polymers are structurally much more restricted than the vast variety of metabolites known from plants or microbes.

For decades, MS experts have manually determined fragmentation pathways to explain tandem MS data and determine the molecular structure. In 2008, Böcker and Rasche 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 excellent quality. They can also be computed from multiple MS and GC-MS data. 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”.

We present an automated method for comparing the FTs of two compounds. Our method is based on local tree alignments, generalizing local sequence alignments. 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 biochemical properties of the compounds. Second, FT alignment scores and chemical similarity 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 present the fragmentation tree basic local alignment search tool, FT-BLAST for short, for searching in a database for compounds with FTs similar to the unknown sample compound. To filter out spurious hits, we present a statistical evaluation based on decoy database searching. 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 (Fig. 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. The method is not restricted to metabolomics but can be applied in many other areas where small molecules have to be identified.
Results

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 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. The masses of all compounds ranged from 75 Da to 1258 Da.

We assume that we know the correct molecular formula of each compound. Computing molecular formulas is possible through combining isotope and fragmentation pattern data. For the QSTAR dataset,

![Workflow diagram](image)

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.
Table 1: Datasets used in this study. The QSTAR dataset and 38 compounds from the Orbitrap dataset were used for evaluating FTs in ref. The MassBank dataset was downloaded from the MassBank database (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.

Identifying the correct molecular formula is possible in all cases. 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.

For each compound, we computed a hypothetical FT, annotating fragment peaks with molecular formulas and modeling fragmentation reactions through dependencies between fragment ions (Fig. 2; Methods section). FTs with fewer losses contain less information and were therefore excluded from our analysis in some cases (Table 1).

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. 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. 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 and, in our experience, FTs rarely
contain nodes of out-degree six or higher (example in Fig. 2). The average running time for each alignment was below 4 milliseconds on a laptop computer.

Alignment scores will clearly be large for large trees and small for small trees, so we normalized by the perfect match score. We aligned each of the two trees against itself and chose the smaller score. We refrained from using the similarity matrix directly. Instead, for each compound we viewed its similarity matrix column as a fingerprint, as is done with gene expression data. We computed the Pearson correlation for any two fingerprints, and processed the resulting fingerprint similarities.

**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\(^{25}\) to the FT fingerprint similarities (Fig. 3 and Supplementary Fig. 1–4). 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 (Fig. 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.** 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\(^{26,27}\) measure of chemical similarity because it is the most widely used. 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 even stronger correlations). We computed the Pearson correlation between FT fingerprint similarity and the PubChem/Tanimoto similarity score (Supplementary Table 5). For the Orbitrap dataset, the Pearson correlation was \(r = +0.65 \ (r^2 = 0.42)\); this correlation increased slightly for FTs with 3+ losses (Fig. 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).

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 (Fig. 4).

**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. 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

![Diagram](image_url)

**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.
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. 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 (Table 2). Many hit lists contained compounds mostly from the same class or with high chemical similarity; other hit lists were short or empty. Only a 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.

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 measured their tandem MS spectra at several collision energies. We identified molecular formulas following a published method. 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 clustered poppy unknowns together with the Orbitrap dataset (Supplementary Fig. 7). 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, 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 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 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.

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 often of excellent quality, but further improvements are possible with better scoring and when more data becomes available. We found that a precise experimental protocol 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\(^\text{21}\). 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 in the method more expert knowledge on characteristic losses and ions.

After 45 years of research\(^\text{9}\), FT alignments may be the key step towards “identifying the unknowns”,

---

\(^\text{9}\) For 45 years of research, FT alignments may be the key step towards “identifying the unknowns”.

\(^\text{21}\) For 45 years of research, FT alignments may be the key step towards “identifying the unknowns”.

enabling the automated analysis of large-scale compound screens. Our method opens a way to a fast classification of metabolites, limiting work spent on “uninteresting” molecules. Areas of application include natural product discovery, identifying derivatisation, searching for signaling molecules, biomarkers\textsuperscript{29}, novel drugs or ones that are illegal, or other “interesting” organic compounds. In future, the method may support systems biology improving our ability to infer metabolic and biosynthetic pathways and networks from tandem MS data. Ultimately, this may allow us to overcome the limits of the “known universe of organic chemistry”\textsuperscript{30}.

\textbf{Addendum}

\textbf{Acknowledgements}  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) for providing the MassBank data, Ravi Kumar Maddula for measuring some of the compounds in the Orbitrap dataset, and Miroslav Strnad (Palackeho University Olomouc, Czech Republic) for supplying the zeatins. Financial support from the Max Planck Society is acknowledged.

\textbf{Author contributions}  SB, FR and TZ developed the initial method. FH and KS further improved the technique. AS contributed chemical knowledge to the method. AS and MK performed the measurements and manual analyses. SB, FH, FR, KS, and AS jointly wrote the manuscript.

\textbf{Availability}  The method will be made available online at \url{http://bio.informatik.uni-jena.de/ftblast/}.

\textbf{Competing Interests}  SB, FH, FR, KS and TZ declare that they have competing financial interest. Value of two patents may be affected by publication.

\textbf{Correspondence}  Correspondence should be addressed to Sebastian Böcker (email: sebastian.boecker@uni-jena.de).

\textbf{Methods}

\textbf{Experimental section.}  Part of the Orbitrap dataset was obtained from a published source\textsuperscript{20}. Other compounds originated from our stock, purchased or donated by M. Strnad (Palackeho University, Olomouc, Czech Republic). Fragmentation spectra were measured using Collision Induced Dissociation (CID) or High-energy Collision Dissociation (HCD, for Orbitrap) fragmentation using parameters described in the Supplementary section. Peak picking was carried out by the vendor software. The MassBank dataset was downloaded from the MassBank database\textsuperscript{23} with accession numbers PR100001 to PR101056. We discarded compounds with unfragmented 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\textsuperscript{20}. Peak lists at different collision energies were merged using a window of 50 mDa.

\textbf{Fragmentation trees and molecular formulas.}  For Orbitrap and QSTAR data, we identified molecular formulas following a published method\textsuperscript{20}. For each compound, we computed a FT as described in ref.\textsuperscript{20} using 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\textsubscript{2}O were given a bonus (Supplementary Table 1). In contrast to the published method\textsuperscript{20} 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. 9,10,11).

**Aligning fragmentation trees.** We computed the optimal local alignment of a pair of FTs using dynamic programming. 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\textsuperscript{24} 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. We normalized similarities by perfect 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 used columns of the resulting similarity matrix as fingerprints (or feature vectors), corresponding to similarities of the respective tree to all other trees in the entire dataset. The *fingerprint similarity* of two compounds is the Pearson coefficient of the corresponding two columns in the original similarity matrix. All algorithms were implemented in Java.

**Clustering.** For clustering, computations were limited to FTs with 3+ losses, or 5+ losses for the combined dataset. We applied hierarchical clustering (Unweighted Pair Group Method with Arithmetic Mean, UPGMA) to the fingerprint similarity matrix using EPoS\textsuperscript{31}. Mostly homogeneous clusters were collapsed based on visual inspection.

**Correlation with chemical similarity.** For each compound pair we estimated chemical similarity and FT fingerprint similarity. We used the Chemistry Development Toolkit\textsuperscript{32} (version 1.3.37) for PubChem/Tanimoto scores\textsuperscript{26,27}, and Open Babel\textsuperscript{33} (version 2.3.0) for MACCS/Tanimoto scores. We estimated Pearson and Spearman correlation coefficients for all datasets and restrictions using the programming language R. Similarly, for between-datasets 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. We found correlation to be weaker than for the FT fingerprint similarity (Supplementary Fig. 8 and Supplementary Table 6).

**FT-BLAST.** We created a database of decoy fragmentation trees\textsuperscript{34} by using the backbones of real fragmentation trees from another dataset\textsuperscript{13}. We searched both databases simultaneously. Results were sorted with respect to FT fingerprint similarity. We restrict search results to a fixed false discovery rate.

**Poppy data.** Surface extracts from *P. nudicaule* flower organs were directly infused on Nanomate nanoelectrospray chips and measured on an Orbitrap XL (Thermo Fischer Scientific, Bremen, Germany). Measurements were conducted in both positive and negative mode. Precursor ions were manually selected based on ion intensities and expected masses obtained from literature, and HCD-fragmented. We used a published method\textsuperscript{20} 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.

References


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 \( \rho = +0.43 \) (\( \rho^2 = 0.18 \)).
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 chemical similarity above 0.85, and 63 compounds that are not considered similar to the query compound. In 56 cases (60%) 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).