nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-020-00234-6

Check for updates

Database-independent molecular formula annotation using Gibbs sampling through ZODIAC

Marcus Ludwig[1], Louis-Félix Nothias[2,3], Kai Dührkop[1], Irina Koester[2,4], Markus Fleischauer[1], Martin A. Hoffmann[1,5], Daniel Petras[2,3,4], Fernando Vargas[3,6], Mustafa Morsy[7], Lihini Aluwihare[4], Pieter C. Dorrestein[2,3] and Sebastian Böcker[1]✉

The confident high-throughput identification of small molecules is one of the most challenging tasks in mass spectrometry-based metabolomics. Annotating the molecular formula of a compound is the first step towards its structural elucidation. Yet even the annotation of molecular formulas remains highly challenging. This is particularly so for large compounds above 500 daltons, and for de novo annotations, for which we consider all chemically feasible formulas. Here we present ZODIAC, a network-based algorithm for the de novo annotation of molecular formulas. Uniquely, it enables fully automated and swift processing of complete experimental runs, providing high-quality, high-confidence molecular formula annotations. This allows us to annotate novel molecular formulas that are absent from even the largest public structure databases. Our method re-ranks molecular formula candidates by considering joint fragments and losses between fragmentation trees. We employ Bayesian statistics and Gibbs sampling. Thorough algorithm engineering ensures fast processing in practice. We evaluate ZODIAC on five datasets, producing results substantially (up to 16.5-fold) better than for several other methods, including SIRIUS, which is the state-of-the-art algorithm for molecular formula annotation at present. Finally, we report and verify several novel molecular formulas annotated by ZODIAC.

Metabolomics characterizes metabolites with high-throughput techniques. Recently, liquid chromatography mass spectrometry (LC-MS) has become widely adopted by the metabolomics community as a powerful and sensitive analytical platform. In untargeted LC-MS experiments, hundreds to thousands of metabolites can be detected from a single biological sample. Annotation of these metabolites remains highly challenging. Today, molecular formula annotation of mass spectrometry (MS) features is often performed by searching in some molecular structure database[1,2]; but this implies that only those molecular formulas can be assigned that are already present in the database chosen. Further, to achieve satisfactory performance, the databases used for this search are usually too small. Here we present ZODIAC, a method that can generate high-quality, high-confidence molecular formula annotations. ZODIAC can overcome the limits described above and discover molecular formulas.

Annotating the molecular formula often allows us to deduce important information about the likely structure of the compound; it improves the performance of in silico methods such as MetFrag[3] or CFM-ID[4] that search within molecular structure databases[5–7]; it is required for downstream computational analysis with, say, CANOPUS for compound class annotation[8]; and finally, annotation guides data interpretation based on atoms and unsaturation degree for full structure elucidation via nuclear magnetic resonance (NMR) or X-ray crystallography. Annotation of molecular formulas is far from trivial, especially if executed de novo, that is, without the use of a molecular structure database and without artificially restricting candidate elements. In de novo annotation, the number of candidate molecular formula grows rapidly with compound size and elements beyond C, H, N and O (CHNO). Heuristic constraints for 'permissible' molecular formulas will counter this growth[9], but can also prevent the annotation of true molecular formulas.

Further annotation of compounds is performed using tandem mass spectrometry (MS/MS); tandem mass spectra of individual compounds can be acquired in an automated fashion. Again, annotation of MS/MS spectra remains highly challenging and is often restricted to searching within a spectral library[10]; spectral libraries are usually tiny in comparison even to small molecular structure databases[11]. Consequently, only a fraction of acquired MS/MS spectra can be annotated by spectral library search[12–15]. In silico methods for searching in molecular structure databases[5,7] overcome the limits of spectral libraries but are restricted to searching molecular structure databases instead. As noted, such in silico methods also profit from knowing the molecular formula of the query compound.

Arguably the best-performing computational method for molecular formula annotation is SIRIUS 4 (ref. [7]), combining isotope pattern matching [7,9,16–22] and the analysis of MS/MS spectra[20,23–25]. SIRIUS reaches best-of-class performance for de novo molecular formula annotation; but even SIRIUS has problems annotating molecular formulas for compounds above 500 daltons (Da): Böcker and Dührkop[25] found that the percentage of correctly identified molecular formulas dropped substantially for larger masses.

An alternative approach to annotating molecular formulas for a complete LC-MS run uses Gibbs sampling and Bayesian statistics,

[1]Chair for Bioinformatics, Friedrich-Schiller-University, Jena, Germany. [2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [3]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [4]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. [5]International Max Planck Research School 'Exploration of Ecological Interactions with Molecular and Chemical Techniques', Max Planck Institute for Chemical Ecology, Jena, Germany. [6]Division of Biological Science, University of California San Diego, La Jolla, CA, USA. [7]Department of Biological and Environmental Sciences, University of West Alabama, Livingston, AL, USA. ✉e-mail: sebastian.boecker@uni-jena.de

utilizing co-occurrence of molecular formulas differing by a predefined set of biotransformations[26–29]. Implicitly, these approaches try to identify molecular structures (or their isomers) from a restricted structure database, and cannot annotate novel molecular formulas. Network visualization approaches that connect compounds by hypothetical biotransformations and common chemical functional groups have been demonstrated to facilitate manual molecular formula annotation[30]. Independently, network-based methods have been developed for structural elucidation and dereplication[14,31,32]. All of these approaches are based on the fact that compounds in an LC-MS run usually co-occur in a network of derivatives.

Results and discussion

Improved molecular formula annotation on five datasets. We present the ZODIAC algorithm (ZODIAC stands for ZODIAC: Organic compound Determination by Integral Assignment of elemental Compositions) for confident, database-independent molecular formula annotation using LC-MS/MS data. ZODIAC takes advantage of the fact that an organism produces related metabolites derived from multiple but limited biosynthetic pathways. ZODIAC builds upon SIRIUS and uses, say, the top 50 molecular formula annotations from SIRIUS as candidates for one compound. ZODIAC then re-ranks molecular formula candidates using Bayesian statistics. Prior probabilities are derived from fragmentation tree similarity, which supports reciprocal plausibility within an LC-MS/MS dataset. On the theoretical side, we have established that finding an optimal solution to the resulting computational problem is non-deterministic polynomial time (NP)-hard; therefore we resorted to Gibbs sampling. By using extensive algorithm engineering, we reduced Gibbs sampling running times to a practical level. To boost robustness, ZODIAC can integrate spectral library search hits. ZODIAC itself does not use isotope patterns, but these are part of the SIRIUS score that ZODIAC builds on. We show that ZODIAC improves molecular formula annotation on a diverse set of biological samples. Furthermore, ZODIAC scores allow us to rank molecular formula annotations by confidence. ZODIAC is not limited to molecular formulas from some structure database, allowing us to discover novel molecular formulas not present in any structural databases.

We evaluated ZODIAC on five diverse datasets representing samples from plants (the dendroides dataset uses *Euphorbia dendroides*, the tomato dataset uses *Solanum lycopersicum*), human plasma (NIST1950), marine microalgae (diatoms from three *Pseudo-nitzschia* species) and mice fecal sample (mice stool); see Extended Data Figs. 1 and 2. For each dataset, we evaluated ZODIAC against a ground truth established by manual validation (for dendroides, see Supplementary Table 1) or spectral library search (for all others, see Supplementary Table 2). Three of the datasets were chosen because we expect a reasonable number of hits using spectral library search, as required for the evaluation. Tomato is a model organism, NIST1950 is reference material, and the mice stool[33] data has been previously analysed manually. Dendroides data has been investigated previously[34] and allowed us to manually expand the ground truth. The discovery of novel molecular formulas in these well-studied datasets is less likely. In contrast, diatoms have been studied less comprehensively and compounds often contain uncommon elements, so we expect to find novel compounds in this dataset. Input mzML/mzXML files were processed with OpenMS[35] and low-quality MS/MS spectra were discarded; see Extended Data Fig. 3, Supplementary Tables 3 and 4 and the Methods section.

For all five datasets, we observe that ZODIAC outperforms SIRIUS (Fig. 1a) and all other publicly available methods (Extended Data Fig. 4), often substantially decreasing molecular formula annotation error rates. Improvements are most distinctive for the dendroides dataset, which contains many larger compounds: 75% of the ground truth compounds have a mass-to-charge ratio (m/z) of 605 or higher (Extended Data Fig. 2). Hence, this dataset is particularly challenging for molecular formula assignment. Out of the 201 ground truth compounds, the preprocessing assigned an incorrect adduct to three; for these, the correct molecular formula is not contained in the candidate list considered by ZODIAC. For one compound, the correct molecular formula was not ranked into the top 50. For the remaining 197 compounds, SIRIUS correctly annotated 49.75% (98), compared to 96.95% (191) for ZODIAC without anchors (library hits). This represents an 16.50-fold decrease in error rate. Error rates improve for compounds over the whole mass range (Fig. 1b).

On the NIST1950 and tomato datasets, SIRIUS already showed excellent performance, with more than 90% correctly annotated molecular formulas. For these, ZODIAC further decreased error rates, from 9.57% to 6.38% for NIST1950 and 4.81% to 1.48% for tomato. The diatoms dataset is rather complex, and compounds may contain halogens. For the diatoms dataset, SIRIUS reached an error rate of 12.90%, which is reduced two-fold to 6.45% by ZODIAC. Again, ZODIAC consistently improves error rates over most of the total mass range (Supplementary Fig. 1).

Mice stool MS/MS spectra were measured with a broader isolation window, and the dataset contains numerous chimeric and low-quality spectra, most of which were discarded before running ZODIAC. Consequently, ZODIAC has a much smaller network of interdependent compounds than for the other datasets. But even spectra that were not discarded often have substantially worse quality than spectra from other datasets: for example, these spectra often contain isotope peaks of fragments or are undetected chimeric spectra. Our evaluation shows that even for this extremely challenging dataset, ZODIAC improves annotation results, decreasing the error rate from 32.56% for SIRIUS to 18.60% for ZODIAC.

Some compounds in an LC-MS/MS run can result in high-scoring hits when searching in a MS/MS spectral library. ZODIAC's stochastic model allows us to integrate these hits as 'anchors', assuming that we can trust assigned molecular formulas to a high degree. We performed a 10-fold cross-validation to assess the improvement using anchors. We ensured structure-disjoint evaluation on the library hits, as multiple 'compounds' in the dataset may correspond to the same structure; see ref. [36] on the importance of structure-disjoint evaluation. We find that ZODIAC with anchors does not improve the error rate.

For four of the datasets (NIST1950, tomato, diatoms and mice stool), ground truth molecular formulas were established by library searching only. We tested whether there is a distinct difference between the cosine score of ZODIAC's correct and incorrect molecular formula assignments, but did not find such a difference (Supplementary Fig. 2).

We find that differentiating between adducts $[M + H]^+$ and $[M + Na]^+$ is sometimes challenging for ZODIAC[37]. This is observable for the dendroides dataset, where all six incorrect ZODIAC annotations show an incorrect adduct annotation, mistaking $[M + H]^+$ for $[M + Na]^+$ or vice versa. In all six cases, the molecular formula of the best ZODIAC hit and the ground truth differ by exactly two carbon minus two hydrogen atoms (21.984349 Da), with mass difference highly similar to that between $[M + H]^+$ and $[M + Na]^+$ (21.981944 Da). Sodium-ionized compounds can produce protonated fragments, making the interpretation of these spectra challenging. We reran ZODIAC on the dendroides dataset, assuming we knew the correct adduct for each reference compound. For all 201 compounds, the correct hit is contained in the SIRIUS top 50 candidate list. SIRIUS correctly annotated 66.17% (133) and ZODIAC 99.50% (200) of the compounds, corresponding to a 68-fold decrease of the error rate. The other four datasets contain fewer sodium adducts.

ZODIAC implicitly tries to estimate the probability that an annotated molecular formula is correct; as expected from the statistical

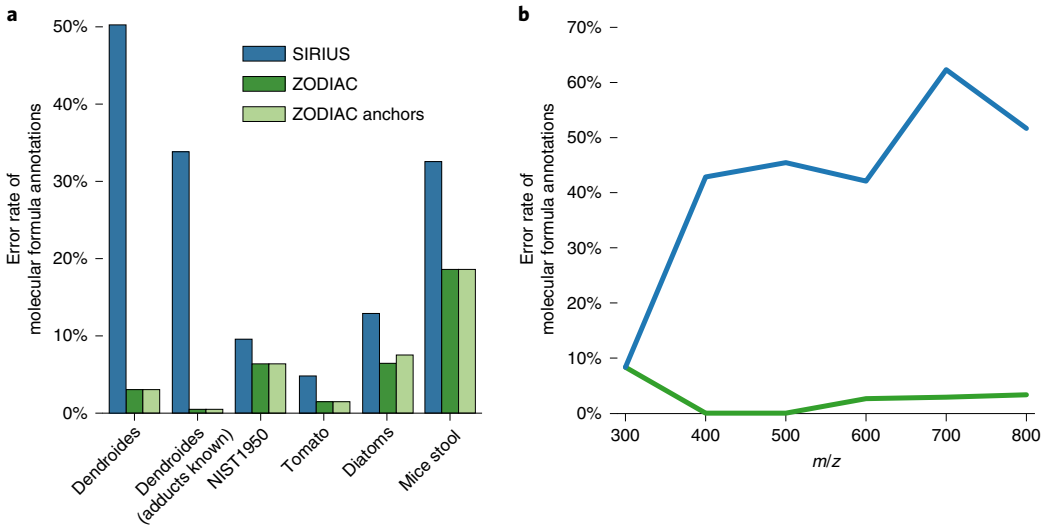


Fig. 1 | Molecular formula annotation error rates. a, Error rate on five datasets. The rate of incorrect molecular formula annotations is displayed for SIRIUS and ZODIAC, with and without anchors. 'Dendroides (adducts known)' is identical to 'Dendroides', but correct adduct assignments are provided to SIRIUS and ZODIAC. SIRIUS and ZODIAC both use MS1 and MS/MS data for molecular formula annotation. Between 43 and 270 compounds were processed per dataset. Evaluation is based on compounds for which SIRIUS ranks the correct molecular formula in the top 50, leaving out few ground truth compounds. For the exact number of compounds and other statistics such as compound masses (Extended Data Fig. 1). Compounds in the dendroides dataset are considerably larger than in the other datasets. ZODIAC reduces error rates on all datasets. Data quality of the mice stool dataset is substantially worse than for the other datasets, resulting in increased error rates; see text for details. See Extended Data Fig. 4 for results on all ground truth compounds and other methods. **b**, Error rates versus mass for the dendroides dataset. Error rates for SIRIUS and ZODIAC without anchors are binned by compound *m/z*; bins of width 100 are centred at *m/z* values of 300, 400, …, 800. See Supplementary Fig. 1 for the other datasets.

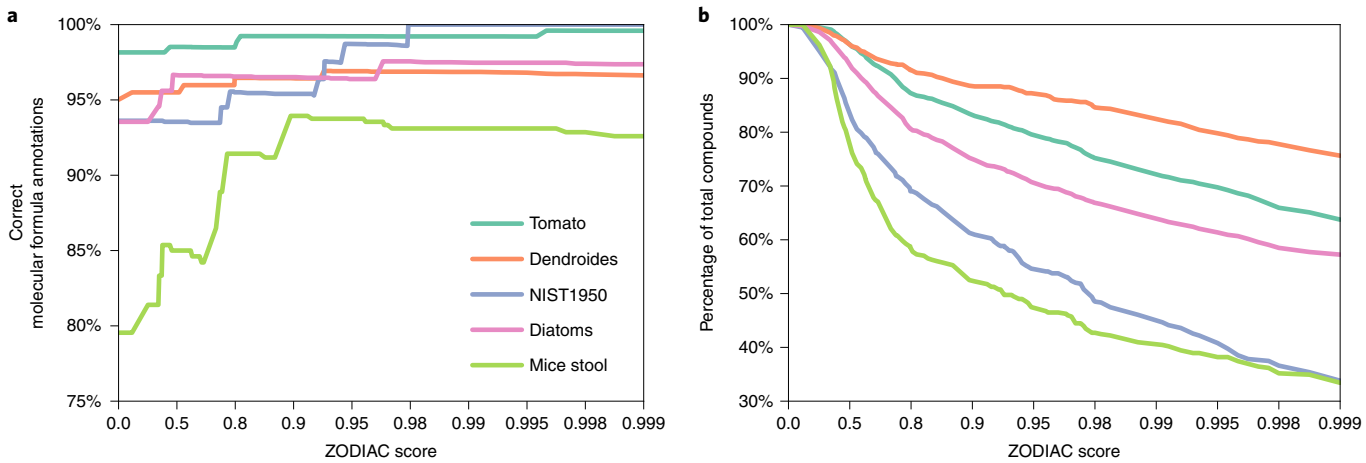


Fig. 2 | Percentage of correct annotations and number of compounds in relation to ZODIAC score. a, Percentage of correct molecular formula annotations for different ZODIAC score thresholds for five datasets. We sort compounds by ZODIAC score and calculate the rate of correct annotations for all compounds above the given thresholds. **b**, Percentage of total compounds with a ZODIAC score above different thresholds on five datasets. Here, we consider all compounds, with and without established ground truth. Note that scores on the *x*-axis are not equidistant. As noted, the data quality of the mice stool dataset is substantially worse than for the other datasets, resulting in decreased correct annotation rates.

theory, we find these estimates to be imprecise (Fig. 2). But the ZODIAC score can be used to differentiate between true and incorrect annotations: for each dataset, we sort molecular formula annotations by the ZODIAC score, and calculate the rate of correct annotations for any subset of top-scoring annotations. We find that high-scoring ZODIAC annotations are more likely to be correct (Fig. 2). For this evaluation, we also considered previously discarded compounds for which SIRIUS did not rank the correct molecular formula in the top 50; for these compounds, ZODIAC cannot find the correct molecular formula but, at best, the incorrect molecular formula should receive low ZODIAC scores. Selecting a ZODIAC score threshold of 0.9 results in more than 93.94% correct annotations while keeping 52.26% to 88.51% of the compounds of each dataset (Fig. 2). In comparison, spectral library search allowed us to annotate between 3.78% and 16.55% of a dataset; see Supplementary Table 1.

We manually investigated incorrect annotations with a ZODIAC score of 0.95 or higher, but did not find conclusive reasons for these high-confidence failures: in several cases, the fragmentation trees of both the ground truth molecular formula and of the top-ranked

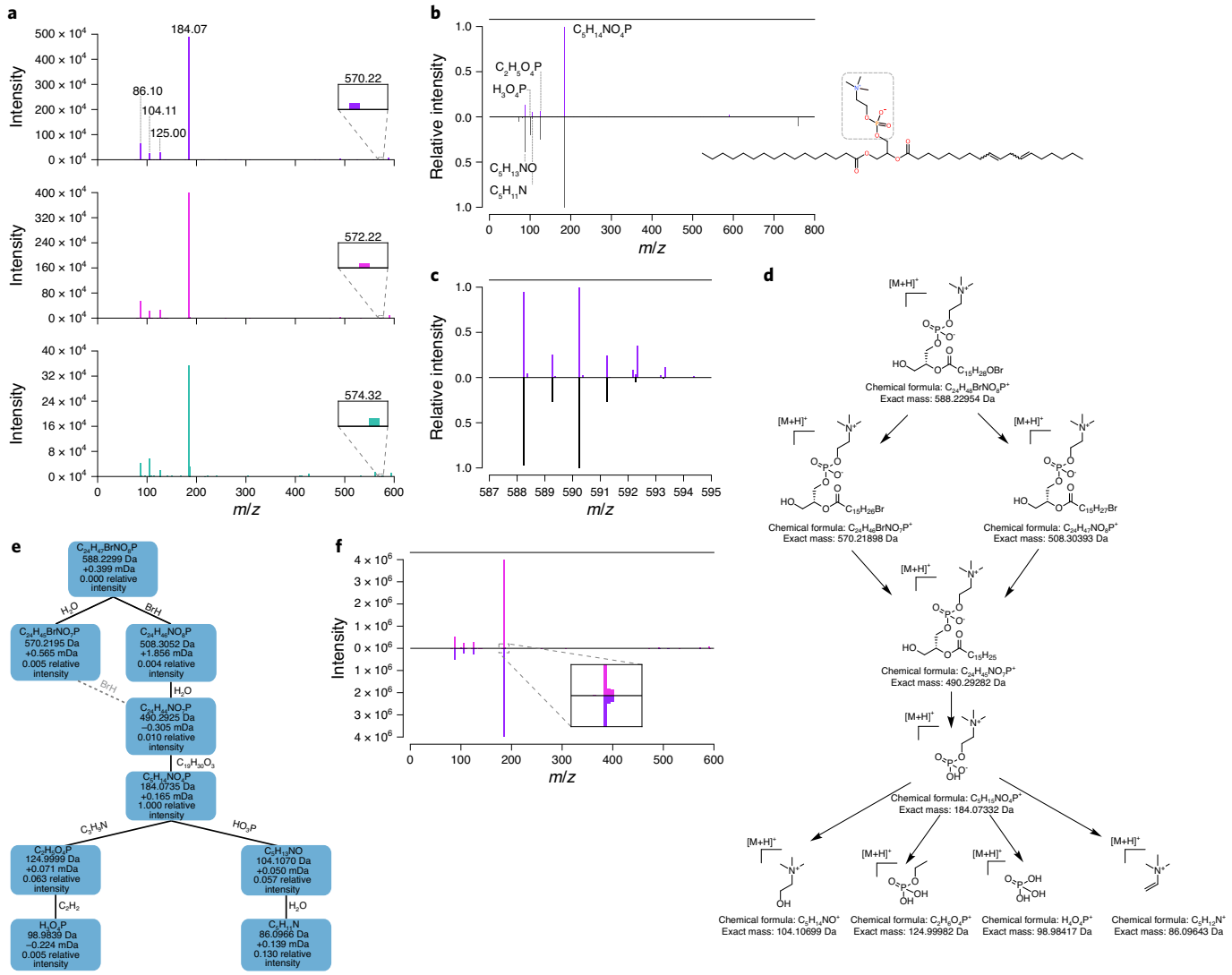


Fig. 3 | Annotation of a novel bromine-containing compound in the diatoms dataset. a, MS/MS spectra for *m/z* 588.230, 590.228 and 592.325, corresponding to the monoisotopic, the M+2 peak and the M+4 peak (top to bottom). The 'moving' peak at *m/z* 570, 572 and 574 corresponds to the same molecular formula but different isotopes. This annotated fragment molecular formula is based on the fragmentation tree in **e** and is the only one containing bromine in these MS/MS spectra. **b**, Partial match to 1-palmitoyl-2-linoleoyl-sn-glycero-3-phosphocholine in the NIST library. The mirror plot compares the MS/MS spectrum of the monoisotopic peak at *m/z* 588.230 (top) to the NIST library spectrum (bottom). The displayed molecular formulas were annotated using the fragmentation tree of the query compound (**e**), and are identically annotated in the NIST reference spectrum. The substructure of the NIST reference compound that corresponds to the annotated peaks is highlighted. **c**, Mirror plot of measured against simulated isotope pattern. The top part displays *m/z* 587 to 595 of the MS1 spectrum with retention time 503.97 s, measured prior to the MS/MS spectrum for precursor *m/z* 588.230. The bottom part is the simulated isotope pattern for $[C_{24}H_{47}BrNO_8P + H]^+$. **d**, Putative structure and fragmentation pathway of the novel compound. **e**, Fragmentation tree computed by SIRIUS. Nodes correspond to fragments, edges to neutral losses. Nodes are annotated with the (neutralized) molecular formula, peak *m/z*, mass deviation in millidaltons and relative intensity. The dashed grey line is not part of the computed fragmentation tree, but is part of a 'parallelogram' which cannot be resolved through a tree[25]. **f**, Mirror plot of measured (top) against simulated (bottom) MS/MS spectrum for precursor M+2.

ZODIAC candidate annotated most fragment peaks with identical molecular formulas, and differ only in the annotation of high-mass fragments. For two compounds in the mice stool dataset, observed mass deviations were larger than the assumed 10 parts per million (ppm): several high-intensity peaks cannot be explained as sub-formulas of the correct molecular formula using mass accuracy of 15 ppm. The corresponding library spectra exhibit similar mass errors, allowing library search nevertheless to identify the (presumably correct) molecular formula. One incorrect candidate in the diatoms dataset explained many low-intensity peaks, whereas the correct candidate did not.

Filters are commonly applied to exclude 'exotic' molecular formulas; this improves annotation performance unless the true molecular formula is 'exotic'. Most notably, Kind and Fiehn[9] introduced the Seven Golden Rules in 2007, which are frequently used in the metabolomics community to filter molecular formula candidates based on element frequencies, mass and isotope ratios. These rules were empirically established from molecular structure databases, but are nowadays sometimes used as though they represented a biological ground truth. We checked ZODIAC molecular formula annotations against the Seven Golden Rules; see Extended Data Fig. 5: We find that all but one molecular formula in the

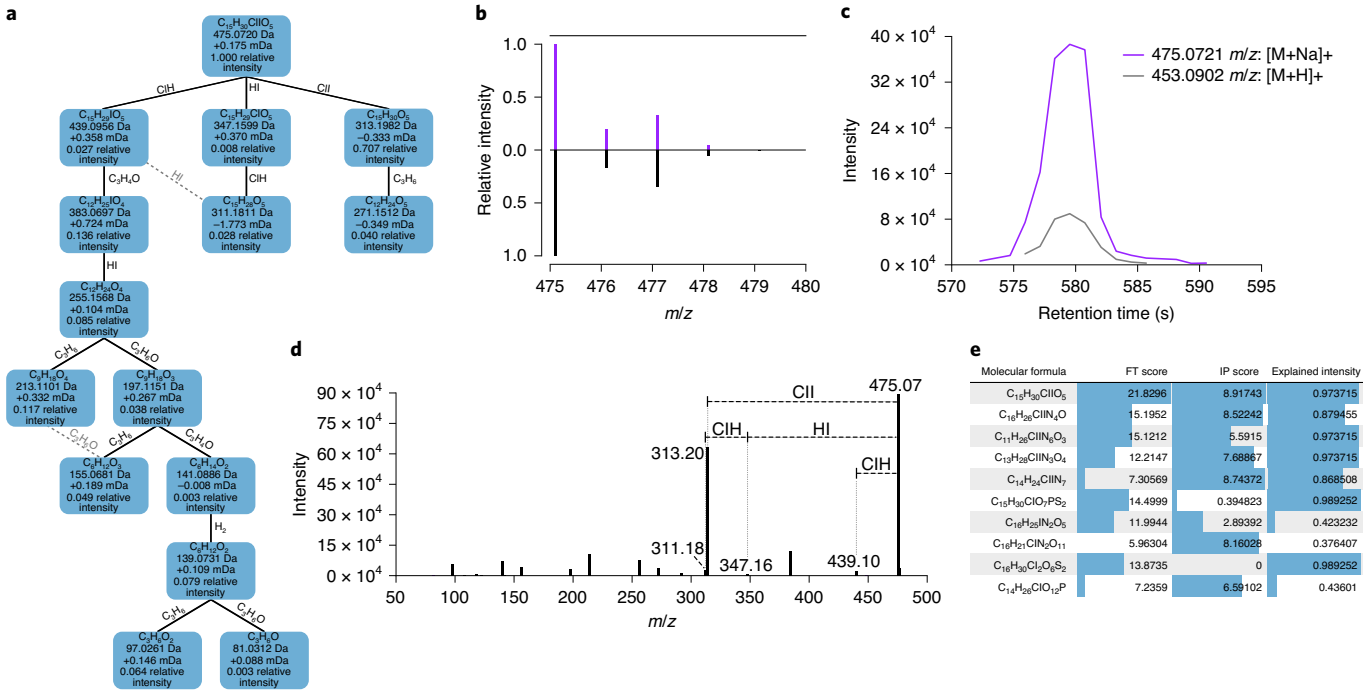


Fig. 4 | Annotation of a novel chlorine- and iodine-containing compound in the diatoms dataset. a, Fragmentation tree for $[C_{15}H_{30}ClIO_5 + Na]^+$ computed from the corresponding MS/MS spectrum in **d**. The dashed grey lines are not part of the computed fragmentation tree, but are part of 'parallelograms' that cannot be resolved through a tree[25]. **b**, Mirror plot of the measured and simulated isotope pattern. The top part displays the measured isotope pattern, extracted as described in the Methods section. The bottom part is the simulated isotope pattern for $[C_{15}H_{30}ClIO_5 + Na]^+$. **c**, Extracted ion chromatograms for *m/z* 475.0721 and *m/z* 453.0902 around retention time 579.16 s with a 5 ppm mass error window. The mass difference of 21.9819 suggests that *m/z* 475.0721 is the sodiated ion and *m/z* 453.0902 the protonated ion of the same compound. **d**, Merged MS/MS spectrum. The indicative neutral losses for chlorine and iodine as annotated by the fragmentation tree in **a** are displayed. **e**, Top ten SIRIUS molecular candidates. Displayed is the fragmentation tree (FT) score, isotope pattern (IP) score and the total explained intensity of fragments in the MS/MS spectrum for each candidate. Only the top two candidates have a ZODIAC score greater than zero.

dendroides dataset adhere to the Seven Golden Rules; this is not surprising because molecular formulas are mainly comprised of CHNO for this dataset. In contrast, diatoms and tomato samples contain many compounds with 'uncommon elements' (phosphorus, sulfur, halogens). For these datasets, numerous molecular formulas do not adhere to the Seven Golden Rules; this includes molecular formulas with high ZODIAC score. We argue that using heuristic filters to exclude 'exotic' molecular formulas will in practice often exclude valid molecular formulas and may ultimately result in fewer correct annotations.

We also evaluated how improved molecular formula annotation affects molecular structure annotation: we found that correct annotations of CSI:FingerID improved by between zero (NIST1950) and 6.38 (diatoms) percentage points, and by 2.80 percentage points in total for the four datasets with putative structure annotations established through spectral library search. The largest improvements in molecular formula annotation error rate were observed for the dendroides dataset; unfortunately, ground truth molecular formulas were established by expert annotation on the dendroides dataset, rendering this type of analysis impossible.

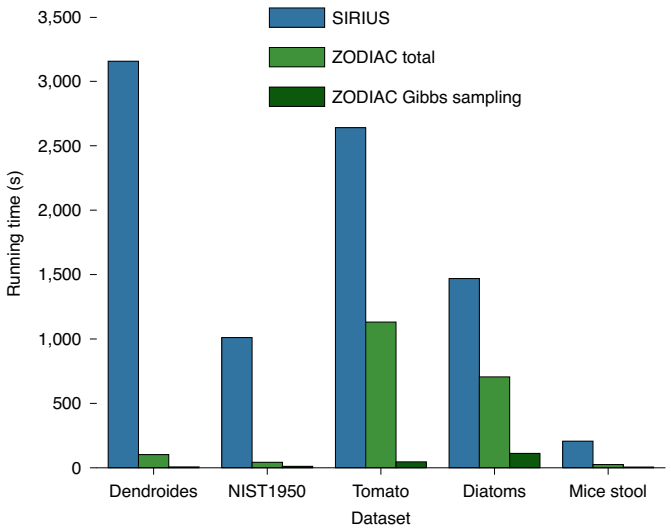


Fig. 5 | Running time comparison of SIRIUS and ZODIAC on five datasets. We run SIRIUS and ZODIAC on 2× Intel XEON E5-2698 with 40 cores in total. 'ZODIAC total' running time includes estimation of the edge score distribution, construction of the similarity graph and computation of ZODIAC scores via Gibbs sampling; the later running time is also given separately. ZODIAC requires SIRIUS results as input, and the total processing time is SIRIUS time plus ZODIAC time.

Novel molecular formulas. We now concentrate on novel molecular formulas discovered by ZODIAC in the five datasets; such discoveries are possible because ZODIAC does not rely on any molecular structure databases such as PubChem[2] and ChemSpider[38], but rather considers all chemically feasible formulas. For a molecular formula to be novel, we require it not to be present in PubChem and, in case it corresponds to a protonated ion, also that the molecular formula obtained by adding NH_3 is not contained either. The second

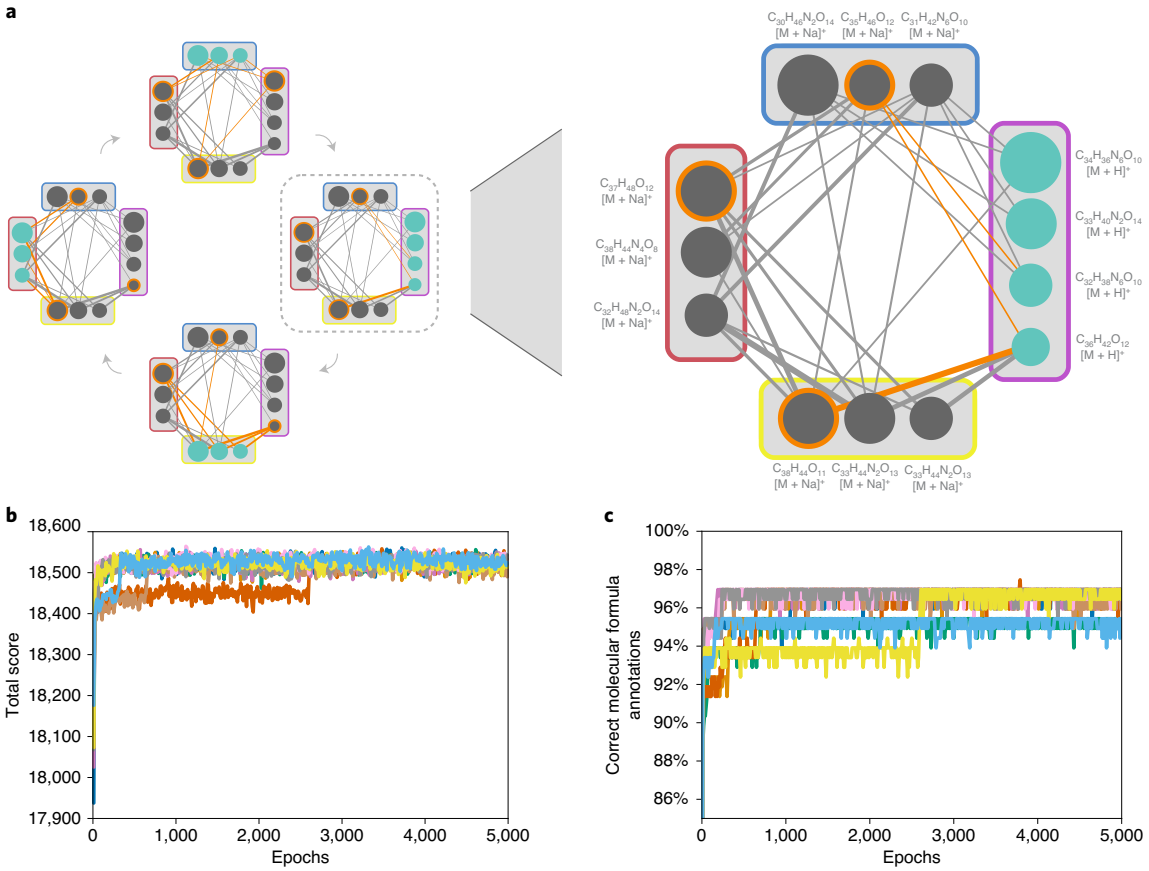


Fig. 6 | Gibbs sampling. a, Illustration of the Gibbs sampling process. The left panel shows that during each epoch compounds are iterated in random order and for each compound a new active molecular formula candidate is sampled based on prior probabilities and active candidates of all other compounds. The right panel shows the sampling step for one compound. This illustrated sub-network with four compounds is based on the dendroides dataset. Each circle corresponds to a molecular formula candidate. The size depicts the rank estimated by SIRIUS. Orange rings mark active candidates. Edge width depicts fragmentation tree-based similarity between candidates. The candidates that are sampled in this step are coloured cyan. **b**, **c**, The total assignment score of the molecular formula candidate network (**b**) and rate of correct annotations over the course of epochs (**c**) for Gibbs sampling on the dendroides dataset.

constraint ensures that we do not treat an undetected ammonium adduct as a novel molecular formula.

To cut down the number of reported formulas, we used strict filters for the ZODIAC score, the quality of the underlying MS/MS data, and the support by other molecular formulas in the dataset. We report molecular formula annotations from all five datasets with (1) a minimum ZODIAC score of 0.98, (2) at least 95% of the MS/MS spectrum intensity being explained by SIRIUS, and (3) at least one molecular formula of the compound connected to five or more compounds in the ZODIAC similarity network. The third criterion discards compounds for which ZODIAC's results are basically identical to SIRIUS's. This results in 15 novel molecular formulas in the tomato dataset, 15 in the diatoms dataset, one in the NIST1950 dataset and one in the mice stool dataset; see Extended Data Fig. 6 and Supplementary Table 5. Filtering less restrictively (ZODIAC score at least 0.95, at least 90% of the MS/MS spectrum intensity being explained by SIRIUS), we annotated 32 novel molecular formulas in the tomato dataset, 26 in the diatoms dataset, three in the NIST1950 dataset and one in the mice stool dataset. The diatoms and tomato datasets contain by far the largest number of novel molecular formulas. ZODIAC thus allows the user to select a few, potentially highly interesting compounds (molecular formulas) from a set of hundreds or thousands without the need for additional experiments or manual data analysis.

Manual evaluation of novel molecular formulas. We show that two top-scoring annotations from Extended Data Fig. 6 can be presumed to be correct. To do so, we manually validated the two novel molecular formulas $C_{24}H_{47}BrNO_8P$ and $C_{15}H_{30}ClIO_5$ in the diatoms dataset; see Figs. 3, 4, Supplementary Note 1 and Supplementary Figs. 3, 4, 5 and 6. This type of manual evaluation is possible for all novel molecular formulas from Extended Data Fig. 6 but we consider that this is beyond the scope of this manuscript.

Running times and stability. In practice, application of Gibbs sampling can be limited by high time demand for burn-in and for sampling a reasonable number of epochs. To avoid this problem, we used extensive algorithm engineering to reduce running times, as detailed in the Methods section. Running times were measured on a computer with 40 cores (2× Intel XEON 20 Core E5-2698). We used ten parallel chains, a burn-in of 1,000 epochs and sampling of 2,000 epochs. ZODIAC required between 1 min and 19 min per dataset, whereas SIRIUS required between 3 min and 53 min per dataset (Fig. 5). SIRIUS required the most time for the dendroides dataset, which contains many high-mass compounds. For the dendroides, NIST1950 and mice stool datasets, ZODIAC computation did not add much to the total running time, whereas for the tomato and diatoms datasets, ZODIAC accounts for roughly one-third of the total running time. In all cases, ZODIAC running time is governed by constructing the

similarity network of molecular formula candidates, whereas running the Gibbs sampler has a negligible impact. We did not evaluate our optimized Gibbs sampler against a naive version, but on the basis of theoretical considerations (discussed in the Methods section), we estimate that the achieved speedup is about 25-fold.

In practice, we can speed up the construction of the similarity network, which depends quadratically on the total number of candidates: we used the top 50 candidates for each compound; this conservative approach avoids the exclusion of correct molecular formulas, and also demonstrates the swiftness of our Gibbs sampling method. But running times can easily be reduced by considering fewer candidates, in particular for low-mass compounds where SIRIUS usually ranks the correct molecular formula much higher. Consequently, ZODIAC can be integrated into existing pipelines without substantial increase in running times.

With regard to stability and required number of epochs, we see that in the beginning both the total network score and the number of correct molecular formula annotations are increasing (Fig. 6). After 500 to 1,000 epochs, nine out of the ten Markov chains reach their different local optima. Estimating the most likely candidates from each chain individually results in 96.95% correct molecular formula annotation in seven of the ten cases. In practice, we run ten parallel Markov chains to allow for parallelization and to make sampling more robust.

Conclusion

We have presented the ZODIAC algorithm, a Gibbs-sampling-based approach for assigning molecular formulas in biological samples analysed by LC-MS/MS. Using ZODIAC, we observed substantial improvements of correct molecular formula annotations, in particular for large compounds; error rates decrease up to 16-fold. Furthermore, the ZODIAC score allows us to select the most confident annotations. In principle, in silico tools such as CFM-ID or MetFrag can also be used for molecular formula annotation, but are intrinsically limited to molecular formulas present in some molecular structure database; in contrast, ZODIAC allows for the de novo annotation of molecular formulas. This is not only of theoretical interest: we have confirmed two novel molecular formulas discovered by ZODIAC that were not contained in PubChem or ChemSpider.

The elaborate preprocessing executed here is no longer required to apply ZODIAC, because detection of features, isotope patterns and adducts as well as feature alignment are now part of SIRIUS 4.4. Furthermore, ZODIAC can import data from popular mass spectrometry frameworks such as MZmine[39], OpenMS[35], and the FBMN workflow[40]. ZODIAC is currently available only for LC-MS/MS data and not for gas chromatography MS.

We found that adduct annotations are very important for molecular formula assignment, because it is challenging to deduce this information from isotope pattern and MS/MS data. Hence, high-quality adduct annotations should be established during preprocessing; furthermore, the combination of positive and negative ionization mode can help with adduct determination. In contrast, we observed that anchors (library hits) have only a small effect on molecular formula annotations.

Searching an unknown compound with novel molecular formula in a structure database may result in an incorrect hit, and this will often go unnoticed. In contrast, a metabolite identification workflow that makes use of de novo annotation methods facilitates the identification of highly interesting new metabolites. ZODIAC constitutes a major step in the discovery and structural elucidation of novel metabolites, natural products and other molecules of biological interest.

Methods

Throughout this paper, the entities of interest consist of signals detected by mass spectrometry where one or more MS/MS spectra have been recorded by the instrument. It is understood that not all of these signals correspond to compounds in the biological sample; but clearly only those signals that do correspond are of interest for our analysis. It is also understood that we usually cannot ultimately decide whether a certain signal stems from the protonated molecule $[M + H]^+$ or, say, the protonated molecule with a water loss $[M - H_2O + H]^+$ or an ammonia adduct $[M + NH_3 + H]^+$. This is not a problem of our method but rather a general problem of mass spectrometry. For the sake of readability, we will nevertheless use the term 'compound' instead of 'hypothetical compound', 'feature', 'adduct' or 'ion'. In contrast, our methods decide for each compound whether it is protonated $[M + H]^+$, a sodium adduct $[M + Na]^+$ or a potassium adduct $[M + K]^+$; in evaluation, compounds that are assigned a wrong adduct are also assigned a wrong molecular formula and hence are always counted as misannotations.

Datasets. *Dendroides.* The extract and fractions of *Euphorbia dendroides* plants collected in Corsica were analysed by high-performance LC-MS/MS in data-dependent acquisition mode on a Orbitrap (LTQ-XL Orbitrap) in positive ionization mode. The same samples[34] were previously investigated and enabled the isolation and identification of novel diterpene esters, including some endowed with antiviral activities against chikungunya and human immunodeficiency virus[37]. These structurally complex and large molecules are characteristic of the *Euphorbia* genus, and the MS/MS spectra of the isolated diterpene esters were deposited in the Global Natural Products Social Molecular Networking (GNPS) library (CCMSLIB00000840316 to CCMSLIB00000840340).

NIST1950. A serial dilution of a plasma methanol extract prepared from the human plasma NIST reference material (SRM 1950)[41] was analysed in data-dependent acquisition mode by ultrahigh-performance LC (UHPLC)-MS/MS on the Q Exactive Orbitrap mass spectrometer in positive ionization mode. Previously, more than 322 compounds were identified by LC-MS in the SRM 1950 reference material[41]. Here, we considered only compounds that were annotated by MS/MS spectral matching with spectra from the NIST17 MS/MS library.

Tomato. Fresh tomato seedling samples (*Solanum lycopersicum*) were extracted in acidic aqueous methanol, offering a wide range of plant metabolites[42]. Samples were analysed in data-dependent acquisition mode by UHPLC-MS/MS on a Q Exactive Orbitrap mass spectrometer in positive ionization mode. Chromatographic separation was performed on mixed-mode C_{18} columns allowing weak anion/cation exchange, which results in the retention of both medium-polarity compounds and apolar compounds. We expect a large number of plant metabolite annotations by spectral library search, including phenylpropanoids and terpenoids.

Diatoms. This dataset consists of solid-phase (PPL, Agilent) extracts of the intra- and exo-metabolomes of a single diatom genus, a major group of marine microalgae. In total, five culture samples of the species *Pseudo-nitzschia subpacifica* (2), *Pseudo-nitzschia delicatissima* (2) and *Pseudo-nitzschia multiseries* (1) were grown in culture from environmental isolates. Cultures included each diatom species and its associated microbiome, resulting in a complex and diverse pool of metabolites. We expected the occurrence of compounds containing uncommon elements: marine microorganisms can contain halogenated organic compounds such as brominated molecules[43]; additionally, the growth medium contained uncommon elements such as selenium. The samples were analysed in data-dependent acquisition mode by UHPLC-MS/MS on a Q Exactive Orbitrap mass spectrometer in positive ionization mode. Metabolomics studies of marine algae are rare and so existing libraries are not expected to have many relevant entries, making this an interesting dataset for testing novel compound identification.

Mice stool. Quinn *et al.* used this dataset to examine the difference between germ free and colonized C57Bl/6J mice by metabolomics[33]. In particular, some molecules were observed only in colonized mice, including novel conjugated bile acids and other food-derived plant metabolites, which showed the role of the microbiome in their metabolization. Mice stool samples from a microbiome study were analysed by UHPLC-MS/MS in data-dependent acquisition mode mode on a maXis QTOF mass spectrometer in positive ionization mode. Owing to technical limitations, the maXis QTOF used for mass spectrometry employs a broader isolation window of (at least 4 m/z), and this makes identification by computational methods challenging because isotope peaks can be found in the MS/MS; it also strongly increases the chance of chimeric spectra (fragmentation spectra comprised of fragments from multiple compounds; see below).

Sample preparation and LC-MS/MS analysis. Mass spectrometry data is deposited on MassIVE (https://massive.ucsd.edu/) and MassIVE accession numbers are specified: MSV000080502 for dendroides, MSV000081364 for NIST1950, MSV000081463 for tomato, MSV000081731 for diatoms and MSV000079949 for mice stool. The exact set of analysed mzML/mzXML input files is listed in Supplementary Table 3.

Dendroides dataset. The latex of *Euphorbia dendroides* was collected and an ethyl acetate extract was prepared as described by Esposito *et al.*[34]. The extract

was then fractionated in 17 fractions that were subjected to mass spectrometry analysis. Subsequent purification led to the isolation and structural elucidation of thirteen diterpene esters characterized by extensive NMR spectroscopy and X-ray crystallography diffraction analysis.

Mass spectra were acquired between *m/z* 150 and *m/z* 1,000. In the full scan mode, the full-width at half-maximum mass resolution of the Orbitrap mass analyser was fixed at 30,000 for MS spectra and at 15,000 for MS/MS spectra. The data-dependent MS^n mode was used to monitor 1 to 3 most intense ions with an exclusion duration of 40 s after 8 repetitions. Instrumental parameters were set as follows: source voltage: 5 kV, lens 1 voltage: −15 V, capillary temperature: 275 °C, gate lens voltage: −35 V, capillary voltage: 25 V, tube lens voltage: 65 V. The collision-induced dissociation parameters were set as follows: collision energy at 30% of the maximum and an activation time of 30 ms. High-performance LC was performed with an HPLC Ultimate 3000 system (Dionex, Voisins-le-Bretonneux, France) consisting of a degasser, a quaternary pump, an autosampler, a column oven, and a photodiode array detector. Separation was achieved using an octadecyl column (Sunfire, 150 mm × 2.1 mm × 3.5 µm; Waters, Guyancourt, France), equipped with a guard column. Column oven temperature was set at 25 °C. Elution was conducted with a mobile phase consisting of water + 0.1% formic acid (A) and MeCN + 0.1% formic acid (B), following the gradient 5–95 % B in 40 min, then maintaining 100% B for 10 min at a flow rate of 250 µl min^{-1}. Injection volume was fixed at 10 µl.

NIST1950 dataset. The NIST SRM-1950 human plasma samples were prepared and extracted with 80% ethanol as proposed in the SRM 1950 paper[41].

The SRM1950 samples were analysed using an UHPLC system (Vanquish, Thermo) coupled to an Orbitrap mass spectrometer (Q Exactive, Thermo) fitted with a heated electrospray ionization (HESI-II, Thermo) probe. Chromatographic separation was accomplished using a Kinetex C18 1.3 µm, 100 Å, 2.1 mm × 50 mm column fitted with a C18 guard cartridge (Phenomenex) with a flow rate of 0.5 ml min^{-1}. 5 µl of extract was injected per sample/QC. The column compartment and autosampler were held at 40 °C and 4 °C, respectively, throughout all runs. Mobile phase composition was: LC-MS grade water with 0.1% formic acid (v/v) (A) and LC-MS grade acetonitrile with 0.1% formic acid (v/v) (B). The chromatographic elution gradient was: 0.0–2.0 min, 5% B; 2.0–10.0 min, 100% B; 10.0–12.0 min, 100% B; 12.0–12.5 min, 5% B; and 12.5–14.5 min, 5% B. Heated electrospray ionization parameters were: spray voltage, 3.5 kV; capillary temperature, 268.0 °C; sheath gas flow rate, 52.0 (arbitrary units); auxiliary gas flow rate, 14.0 (arbitrary units); auxiliary gas heater temperature, 433.0 °C; and S-lens RF, 60 (arbitrary units). MS data was acquired in positive mode using a data-dependent method with a resolution of 35,000 in MS and a resolution of 17,000 in MS/MS. An MS1 scan from 100 to 1,500 *m/z* was followed by an MS/MS scan, using collision-induced dissociation, of the five most abundant ions from the prior MS1 scan.

Tomato dataset. Tomato (*Solanum lycopersicum* var. Better Boy) seeds were surface sterilized in 1.0% (v/v) sodium hypochlorite for 15 min with moderate agitation, rinsed with 20 volumes of sterile distilled water five times, and sown in pots containing peat. Seeds were incubated in a growth chamber with 16 h light/8 h dark photoperiod and 25 ± 2 °C temperature. Two weeks after germination, roots were washed, dried and whole seedlings were frozen in liquid nitrogen and stored at −80 °C until further extraction. Three seedlings were pooled into 2 ml tubes and 1.2 ml of acidified aqueous methanol was added (75% methanol (v/v), 24.9% water (v/v), and 0.1% formic acid (v/v)) to obtain a wide range of plant metabolites[42]. The samples were then homogenized in a tissue-lyser (QIAGEN) at 25 Hz for 5 min, and then centrifuged at 15,000 rpm for 15 min. The supernatant was collected in 96-well plates and dried with a vacuum centrifuge. The samples were resuspended in 130 µl of 7/3 MeOH/H_2O containing 0.2 µM of amitriptyline (*m/z* 278.189, 542 s) as an internal standard. After the plates centrifugation at 2,000 rpm for 15 min at 4 °C, 100 µl of samples were transferred into a new 96-well plate for mass spectrometry analysis.

Samples were analysed with an UHPLC device (Vanquish, Thermo Scientific) coupled to a quadrupole-Orbitrap mass spectrometer (Q Exactive, Thermo Scientific). Chromatographic separation was performed in mixed-mode on a Scherzo SM-C18 (Imtakt, Torrance, USA) column allowing weak anion/cation exchange (250 × 2 mm, 3 µm) with a guard cartridge (Imtakt). The column was maintained at 40 °C. The mobile phases used were 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B), and flow rate was set to 0.5 ml min^{-1}. Chromatographic elution method was set as follows: 0.00–5.00 min, isocratic 2% B; 5.00–8.00 min, gradient 2–50% B; 8.00–13.00 min, gradient 50–100% B; 13.00–14.00 min, isocratic 100% B; 14.00–14.10 min, 100–2% B; 14.10–18.00 min, isocratic 2% B. The injection volume was set to 10 µl. The analyses were performed in electrospray ionization, operating either in positive or in negative ionization mode with a heated electrospray ionization source. In positive ionization mode, the following source parameters were used: spray voltage, +3,000 V; heater temperature, 370 °C; capillary temperature, 350 °C; S-lens RF, 55 (arbitrary units); sheath gas flow rate, 55 (arbitrary units); and auxiliary gas flow rate, 20 (arbitrary units). In negative ionization mode, the following source parameters were used: −3,000.0 V; heater temperature, 375 °C; capillary temperature, 350 °C; S-lens RF, 55 (arbitrary units); sheath gas flow rate, 55 (arbitrary units); and auxiliary gas flow rate, 20 (arbitrary units). The MS1 scans were acquired at a resolution of 35,000 (at *m/z* 200) for the 100–1,500 *m/z* range, and the MS/MS scans at a resolution of 17,500 from 0.48 to 16.0 min. The automatic gain control target and maximum injection time were set at 5 × 105 and 150 ms for MS1 and MS/MS scans. Up to four MS/MS scans in data-dependent mode were acquired for most abundant ions per duty cycle, with a starting value of *m/z* 70. Higher-energy collision-induced dissociation was performed with a normalized collision energy of 20 eV, 35 eV and 50 eV. The apex trigger mode was used (2–7 s) and the isotopes were excluded. The exclusion parameter for the data-dependent parameter was set to 9 s.

Diatoms dataset. Diatoms of the genus *Pseudo-nitzschia* were isolated from waters off the Scripps Pier and their associated microbiome were cultured in filtered Natural Seawater Medium, supplemented with inorganic nutrients ($NaNO_3$, NaH_2PO_4, Na_2SiO_3), vitamins and AQUIL trace metals[44]. Cultures were harvested in stationary growth phase and media and intracellular metabolites were solid-phase extracted using Bond Elute PPL resin (Agilent) to retain a wide range of non-polar to polar molecules[45,46].

The methanol extracts were analysed by UHPLC-MS/MS on a Q Exactive Orbitrap mass spectrometer in positive data-dependent acquisition mode[46]. In short, dried samples were re-dissolved in 100 µl methanol/formic acid (99:1, Fisher Scientific, San Diego, USA) of which 10 µl were injected into a Vanquish UHPLC system coupled to a Q-Exactive Orbitrap mass spectrometer (ThermoFisher Scientific, Bremen, Germany). For UHPLC separation, a reversed phase C18 porous core column (Kinetex C18, 150 × 2 mm, 1.8 µm particle size, 100 Å pore size, Phenomenex, Torrance, USA) was used. As mobile phase A H_2O + 0.1 formic acid and solvent B acetonitrile (ACN) + 0.1% formic acid was used. The flow rate was set to 0.5 ml min^{-1} and after injection, compounds were eluted with a gradient from 0–0.5 min, 5% B, 0.5–8 min 5–50% B, 8–10 min 50–99% B, followed by a 2 min washout phase at 99% B and a 3 min re-equilibration phase at 5% B. ESI parameters were set to 52 l min^{-1} sheath gas flow, 14 l min^{-1} auxiliary gas flow, 01 min^{-1} sweep gas flow and 400 °C auxiliary gas temperature. The spray voltage was 3.5 kV and the inlet capillary 320 °C. S-lens voltage was 50 V. MS scan range was 150–1,500 *m/z* with a resolution at *m/z* 200 ($R_{m/z200}$) of 70,000 with one micro-scan. The maximum ion injection time was set to 100 ms with an automatic gain control target of 1.0×10^6. Up to 5 MS/MS spectra per MS1 survey scan were recorded DDA mode with $R_{m/z200}$ of 17,500 and one micro-scan. The maximum ion injection time for MS/MS scans was 100 ms with an automatic gain control target of 3.0 x 10^5 ions and minimum 5% C-trap filling. The precursor isolation width was set to *m/z* 1. Normalized collision energy was set to a stepwise increase from 20% to 30% to 40% with default charge state $z = 1$. MS/MS scans were triggered at the apex of chromatographic peaks within 2 s to 15 s from their first occurrence. Dynamic exclusion of precursors was set to a duration of 5 s and precursors with unassigned charge states as well as isotope peaks were excluded from MS/MS acquisition.

Mice stool dataset. The mice stool samples were obtained and extracted in 70% ethanol as described by Quinn et al.[33].

The samples were analysed with an UHPLC device (UltraMate 3000 Dionex, Fisher Scientific, Waltham, MA, USA) coupled with a Bruker Daltonics MaXis quadrupole time-of-flight (QTOF) mass spectrometer (Bruker, Billerica, MA USA) as described in ref. [33]. In brief, the metabolites were separated using a Kinetex 2.6 µm C18 (30 × 2.10 mm) UHPLC column fitted with a guard column. The isolation width was dependent on *m/z* value, with a 4 *m/z* isolation for 50 *m/z* to 8 *m/z* at values of 1,000 or higher. Lower isolation width results in a drop in sensitivity.

Preprocessing. To ensure reproducibility, we provide a virtual machine comprising all steps of the preprocessing. The virtual machine incorporates the OpenMS sources, executables and parameter files and all data processing scripts written in Java and Python. The workflow described below is visualized in Extended Data Fig. 3.

OpenMS. We used OpenMS 2.4.0 (ref. [35]) to process the mzML/mzXML files. We performed minor modifications on the OpenMS source code by removing 12 lines and adding 78 lines. This allowed us to detect more isotope peaks, to match MS/MS to MS1 features based on the actual isolation window and to add functionality to the SIRIUSAdapter to directly output SIRIUS file format including retention time information. These numbers are based on the patch file (see 'Code availability'), and include lines with comments and blank lines; the modification of a line corresponds to the removal and insertion of a new line. The recent OpenMS version 2.5.0 includes modifications based on these changes.

Feature finding and clustering of isotopic mass traces was performed using the FeatureFinderMetabo module. Next, adducts were detected with the MetaboliteAdductDecharger module. Finally, spectra were exported to the SIRIUS-specific format using the SIRIUSAdapter module. OpenMS parameter files are provided as part of the virtual machine. Parameters were chosen by manual inspection; in particular, we used a small noise intensity threshold to increase chances that isotope peaks of a compound are picked.

Discarding features and MS/MS spectra. First, we excluded *m/z* features which eluted over a very long time during chromatography and did not produce desired mass traces in a limited time window, as such traces are considered chemical

noise. To do so, we binned MS1 peaks with a bin size of 0.006 *m/z*. Each MS1 was normalized by the most intense peak and only peaks with at least 0.01 relative intensity were considered. For each possible *m/z*, we count the number of MS1 that contain a peak in the corresponding bin. If more than 20% of MS1 contain a peak with this *m/z*, it was considered chemical noise. Because these spurious chemical noise features have rather high mass deviation we removed all MS1 features within 30 ppm of this *m/z*.

Second, we performed blank removal using blank samples from the corresponding datasets. Features within 15 ppm and 20 s of a blank feature were removed if intensities were lower than twice the blank feature intensity. We did not perform blank removal on the mice stool dataset, because this resulted in a low number of remaining compounds.

Third, we removed features from the beginning and end of the chromatography run and features with low relative or absolute intensity; and we removed MS/MS spectra which could not be assigned to an MS1 feature, MS/MS of a precursor peak with low absolute or relative intensity, and chimeric MS/MS. See Supplementary Table 4 for dataset-specific parameter values. Chimeric spectra contain fragments of multiple precursor ions; we detected chimeric spectra as follows. All peaks within the isolation window, excluding isotope peaks, were considered to contribute their intensity to the measured MS/MS. We estimated the relative intensity that the target precursor ion contributes to the MS/MS; if the target precursor ion contributed to less than 50% of the MS/MS intensity or if a second precursor ion contributed more than 33% of the target precursor ion intensity, the MS/MS was marked as chimeric and excluded. The isolation window width for the Orbitrap mass spectrometer used for the dendroides, NIST1950, tomato and diatoms is 1 Da; for the mice stool dataset analysed on a QTOF mass spectrometer an isolation window of 3 Da width and shifted by 1 Da to the right, centred at the +1 isotope peak, was assumed.

Filtering MS/MS spectra. In each MS/MS spectrum, we filtered peaks using an intensity threshold of two times the median noise intensity, see Supplementary Table 4. The median noise intensity of a dataset was estimated from peaks that had no molecular formula decomposition within a 40 ppm window considering elements CHNOP plus those elements predicted from the isotope pattern; see below. Isotope peaks were removed from MS/MS spectra of the mice stool dataset.

The SIRIUSAdapter OpenMS module combines MS/MS that are associated with the same MS1 feature. In addition, complete linkage hierarchical clustering was conducted to merge features over different LC-MS/MS runs. Features were merged using 15 ppm mass accuracy and a 15 s retention time window. Features with different adduct annotations or features from the same run were not merged. Feature similarity was computed by the cosine product of the MS/MS (see below), and the similarity threshold for clustering was set to 0.8. When multiple features were merged into a single feature, where each feature has an assigned isotope pattern, then the isotope pattern with the highest number of isotope peaks was kept. In cases where multiple isotope patterns had the same number of isotope peaks, the one with the most intense monoisotopic peak was kept. After merging, features were discarded if the summed MS/MS intensity was below a threshold; see Supplementary Table 4. Features with precursor mass above 850 Da are discarded: although ZODIAC is clearly capable of processing such features, we found that there are no spectral library hits above this mass that can be used for evaluation; see below. Only 2.72% of features across all datasets have *m/z* above 850 Da, so excluding these cannot have a substantial impact on result statistics.

Extending isotope patterns. OpenMS often misses low-intensity isotope peaks. To recover those peaks, we post-processed OpenMS results as follows: for each isotope pattern detected by OpenMS, we tried to extend it using isotope peaks from the corresponding MS1 spectra chosen by OpenMS. Isotope pattern peaks were picked using the SIRIUS 4 isotope pattern picking subroutine. If an additional isotope peak was present in at least 66% of the corresponding MS1, the peak was added to the isotope pattern. Subsequently, features with less than two isotope peaks are discarded.

Discarding low-quality merged MS/MS spectra. Even when considering all MS/MS spectra for some features, we sometimes have insufficient information for both spectral library search and molecular formula annotation; to this end, such 'low-quality features' were discarded. A feature is discarded if it produces fewer than five fragment peaks, estimated after merging peaks within 10 ppm or 0.0025 *m/z* from all corresponding MS/MS spectra; and if no fragmentation tree in the top 50 candidate list can explain at least 5 peaks accounting for at least 80% of total spectrum intensity; see the SIRIUS analysis below. Filtering 'low quality' features decreased the number of features for dendroides from 1,078 to 784, for the NIST1950 dataset from 568 to 400, for the tomato dataset from 3,583 to 2,584, for the diatoms dataset from 3,227 to 2,075 and for the mice stool dataset from 577 to 377.

For brevity, we will refer to the features detected by OpenMS as compounds; see above.

SIRIUS analysis and establishing a ground truth. SIRIUS 4 was run with the default alphabet of the elements CHNO, at most five phosphorus atoms, and one iodine atom; automatic element detection from the isotope pattern[47] was enabled for sulfur, chlorine, bromine, boron and selenium. For the dendroides, NIST1950 and tomato datasets we used 15 ppm maximum mass deviation for SIRIUS; for the diatoms and mice stool datasets we used 10 ppm. The SIRIUS default ring double bond equivalent (RDBE) value to filter molecular formula candidates was lowered from −0.5 to −1.0, to account for undetected ammonium adducts. Isotope patterns were not used to filter molecular formula candidates before computing fragmentation trees. Furthermore:

(1) If OpenMS provided an ionization adduct type (such as protonation, sodium adduct, potassium adduct) for a compound, only this ionization was used. We export the 50 best-scoring molecular formula candidates from SIRIUS.
(2) In cases where no ionization adduct type was provided by OpenMS, we selected one or more adducts from $[M + H]^+$, $[M + Na]^+$, and $[M + K]^+$ by searching for characteristic mass differences, using the MS1 that contained the most intense peak of the precursor ion. Peaks below 5% relative intensity were discarded for this decision. For each compound, we export the 50 best-scoring molecular formula candidates; we simultaneously ensure that for each considered ionization adduct type, at least ten candidates are considered.

We refer to the resulting candidate list as the 'top 50'.

To evaluate the performance of SIRIUS and ZODIAC, we had to annotate a subset of compounds with 'correct' molecular formulas, to serve as our ground truth. For this, we combined manual annotation and spectral library search, as follows. For the dendroides dataset, spectral library hits were obtained for the isolated molecules that had their reference MS/MS spectra added to the GNPS library. We used molecular networking and spectral library search in analogue mode[14], along with a set of known typical biotransformations, to annotate related diterpene esters. They differ mainly by their degree of acylation, and by the nature of acyl residues on the diterpene backbone. This resulted in 201 compounds being annotated with molecular formulas by manual analysis of the data; see Supplementary Table 1.

For the remaining datasets, we performed spectral library searches against multiple libraries, but did not add manual annotations. We searched compounds in a spectral library combining GNPS[14], MassBank[12], the NIST17 database (National Institute of Standards and Technology, v17) and the 'MassHunter Forensics/Toxicology PCDL library' (Agilent Technologies, Inc.). Prior to library search, we removed peaks above the precursor ion mass from reference and query spectra. Given a spectrum with precursor ion mass M, we removed all peaks with *m/z* above $M - 0.5$ from this spectrum. We computed a similarity score assuming peaks as Gaussians, with the *m/z* values of the centroided peaks as the mean, the standard deviation being the maximum of a relative mass error of 20 ppm and an absolute mass error of 0.005 *m/z*. Precursor ion masses are permitted to differ by 10 ppm or 0.0025 *m/z* at maximum. Only library hits with a similarity score of 0.7 or higher and with at least six shared peaks are considered valid. We computed the score as the mean of the cosine score of the sample spectrum and the cosine score of the mirrored spectrum; to mirror a spectrum with precursor mass M, we replaced peak *m/z* value m by $M - m$. We used the neutral molecular formula of the precursor ion as ground truth. Hence, for precursor ion $[M + H - H_2O]^+$, the neutral ground truth molecular formula is $M - H_2O$. This resulted in 94 annotated compounds for the NIST1950 dataset, 271 for the tomato dataset, 93 for the diatoms datset and 44 for the mice stool dataset; see Supplementary Table 2.

We evaluated SIRIUS and ZODIAC against these 'ground truth' molecular formulas, but we stress that besides the molecules that were isolated in *Euphorbia dendroides* samples and correspond to level 1 of the Metabolomics Standards Initiative ranking system[48], not all of these are necessarily correct. In particular, we refrain from ranking these according to the Metabolomics Standards Initiative ranking system, where level 4 corresponds to an 'unequivocal molecular formula'. An evaluation is nevertheless meaningful because we expect few errors on the molecular formula assignment level.

In a few cases, the correct molecular formula was not ranked in the top 50 SIRIUS candidates; we also dropped these from our evaluation, because it would not be possible for ZODIAC to find the correct molecular formula in our evaluation. We discarded four compounds for the dendroides dataset, zero for the NIST1950 dataset, one for the tomato dataset, zero for the diatoms dataset and one compound for the mice stool dataset because of this criterion.

See Extended Data Fig. 1 for details, and see Extended Data Fig. 2 for the mass distribution of the 'ground truth' compounds. Compounds in the dendroides dataset with reference annotations have high mass, and 75% of all reference annotations have an *m/z* of 605 or higher. The NIST1950 dataset resulted in library hits over a broad range of *m/z* values. The diatoms library hits have a median *m/z* of 301 but the sample itself is highly complex, as described above. Only a few compounds remain in the mice stool dataset after filtering out chimeric and low-quality compounds, see above.

Posterior probability of an assignment. We used a probabilistic view for the molecular formula assignment problem[26]. For each hypothetical compound in the LC-MS run, we are given data such as an isotope pattern and a fragmentation pattern. This allows us to determine, for each compound $c \in \mathcal{C}$, a set of candidate molecular formulas that may explain the observed data. Let V be the set of all molecular formula candidates, such that $V(c) \subseteq V$ is the subset of molecular formula candidates for compound $c \in \mathcal{C}$. It is possible that different compounds share an identical molecular formula explanation, but we ignore this in our presentation, solely for the sake of readability; clearly, ZODIAC can assign the same molecular formula to isomeric compounds. An assignment is a mapping $\alpha : \mathcal{C} \rightarrow V$

where $\mathfrak{a}(c) \in V(c)$ is the molecular formula assigned to compound c. The posterior probability of an assignment $\mathfrak{a}$ is

$$\mathbb{P}(\mathfrak{a}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathfrak{a}) \cdot \mathbb{P}(\mathfrak{a})}{\mathbb{P}(\mathcal{D})} \propto \mathbb{P}(\mathcal{D}|\mathfrak{a}) \cdot \mathbb{P}(\mathfrak{a}) \quad (1)$$

where $\mathcal{D}$ is the observed data. We use the terms 'prior probability', 'likelihood' and 'posterior probability' according to this Bayesian point of view. Let $\mathcal{D}(c)$ be the observed data for compound $c \in \mathcal{C}$, that is, the isotope pattern and fragmentation pattern of c. We assume that the likelihoods of molecular formulas for different compounds are independent, and that the likelihood of any compound c only depends on its data $\mathcal{D}(c)$, so

$$\mathbb{P}(\mathcal{D}|\mathfrak{a}) = \prod_{c \in \mathcal{C}} \mathbb{P}(\mathcal{D}(c)|\mathfrak{a}(c)).$$

Next, we define the prior probability of an assignment as the product of priors for pairs of compounds:

$$\mathbb{P}(\mathfrak{a}) \propto \prod_{c,c' \in \mathcal{C}, c \neq c'} \prod_{u \in V(c)} \prod_{v \in V(c')} \mathbb{P}(u, v|\mathfrak{a}(c) = u, \mathfrak{a}(c') = v).$$

Here, $\mathbb{P}(u, v|$ 'true') is the prior probability that two compounds with molecular formulas u, v co-occur in the dataset; analogously, $\mathbb{P}(u, v|$ 'false') if u, v do not co-occur. To simplify our calculations, we introduce a mapping $\mathfrak{c} : V \to \mathcal{C}$ that maps any molecular formula to the compound it belongs to: $\mathfrak{c}(v) = c$ for all $v \in V(c)$, for $c \in \mathcal{C}$. Note that $\mathfrak{c}(\mathfrak{a}(c)) = c$ for all $c \in \mathcal{C}$. Now,

$$\mathbb{P}(\mathfrak{a}|\mathcal{D}) \propto \prod_{c \in \mathcal{C}} \mathbb{P}(\mathcal{D}(c)|\mathfrak{a}(c)) \cdot \prod_{u,v \in V, \mathfrak{c}(u) \neq \mathfrak{c}(v)} \mathbb{P}(u, v|\mathfrak{a}(\mathfrak{c}(u)) = u, \mathfrak{a}(\mathfrak{c}(v)) = v). \quad (2)$$

Unlike in Rogers *et al.*[26], we are able to formulate the posterior probability of an assignment in closed form. A natural question is whether we can find a maximum a posteriori estimate for equation (2); unfortunately, this turns out not to be easy, as the underlying computational problem is NP-complete. Another natural question is to sample from the posterior distribution; this will be addressed below.

Graph-theoretical formulation. We now give a graph-theoretical formulation of the problem; this will allow us to establish its computational complexity, but also to come up with a more efficient algorithm. Let V, the molecular formula candidates, be the nodes of an undirected graph $G = (V, E)$ with edge set $E \subseteq \binom{V}{2}$. We will write uv as shorthand for a tuple $\{u, v\} \in \binom{V}{2}$. We use $\mathfrak{c} : V \to \mathcal{C}$ as a node colouring with colour set $\mathcal{C}$. Now, an assignment is a subset $A \subseteq V$ such that each colour from $\mathcal{C}$ appears exactly once; in this case, A is also called multicoloured. Using the notation of the previous section, we have $A = \mathfrak{a}(\mathcal{C})$; remember that $\mathfrak{c}(\mathfrak{a}(c)) = c$ for all $c \in \mathcal{C}$. Let $w : V \cup E \to \mathbb{R}$ be weights for all nodes and edges of the graph. The weight of the assignment A is

$$w(A) := \sum_{v \in A} w(v) + \sum_{uv \in E, u,v \in A} w(uv) \ . \quad (3)$$

This corresponds to the node plus edge weights of a node-induced subgraph of G, for node set $A \subseteq V$.

We consider the following optimization problem:

Maximum Multicoloured Subgraph problem. We are given a graph $G = (V, E)$, a node colouring $\mathfrak{c} : V \to \mathcal{C}$ and weights $w : V \cup E \to \mathbb{R}$. We search for an assignment $A \subseteq V$ of maximum weight, that is, a node-induced multicoloured subgraph of maximum weight.

How does this problem correspond to our probabilistic problem from the previous section? Setting $E = E^* := \{uv | u, v \in V, \mathfrak{c}(u) \neq \mathfrak{c}(v)\}$ (the set of all node pairs with different colours) and

$$w(v) := \log \mathbb{P}(\mathcal{D}(\mathfrak{c}(v))|v) \ \text{ and } \ w(uv) := \log \mathbb{P}(u, v|\text{'true'}) - \log \mathbb{P}(u, v|\text{'false'}) \quad (4)$$

we can show that these problems are in fact equivalent. We have $\log \mathbb{P}(\mathfrak{a}|\mathcal{D}) = w(\mathfrak{a}(\mathcal{C})) + \alpha$ for some constant $\alpha \in \mathbb{R}$. Here, we assumed that $E = E^*$ contains all possible edges; we call (V, E^*) a complete assignment graph. But we can encode any edge set $E \subsetneq E^*$ using zero edge weight for all $e \notin E$, so both problems are equivalent. Hence, it is natural to ask for an optimal solution of the problem, which would correspond to a maximum a posteriori estimator.

Complexity of the problem. For the decision version, we ask if there is an assignment with weight above some threshold $\tau \in \mathbb{R}$. In its simplest form, all edges have weight one and all nodes have weight zero, $w|_E \equiv 1$ and $w|_V \equiv 0$.

Lemma 1. *The Multicoloured Subgraph problem is NP-complete, even for unit edge weights and zero node weights.* **Proof of Lemma 1** It is clear that the Multicoloured Subgraph problem is in NP. We show that the problem is NP-hard by reduction from CLIQUE[49]: let $G = (V, E)$ be an undirected, simple graph; is there a clique of size k in G? Clearly, $k \leq n := |V|$.

We construct a graph $H := G \square \bar{K}_k$ as the Cartesian graph product of G and the empty graph $\bar{K}_k$with k nodes and no edges. That is, for every node $v \in V$ we generate k copies $(v, 1), \ldots, (v, k)$ in H, and there is an edge $\{(u, i), (v, j)\}$ with $i \neq j$ in H if and only if there is an edge uv in G. Now, $k \leq n$ implies that H contains at most n^2 nodes. We define node colours $1, \ldots, k$ such that $c((v, i)) = i$ for $v \in V$ and $1 \leq i \leq k$. We assign zero node weights and unit edge weights for all nodes and edges in H. Now, any assignment in H corresponds to a k-node induced subgraph in G, and the weight of the assignment equals the number of edges in the node-induced subgraph; to this end, an assignment of weight $\binom{k}{2}$ would correspond to a k-clique in G.□

The Multicoloured Subgraph problem is a generalization of the Multicoloured Clique problem; to this end, Lemma 1 can also be inferred from the complexity of Multicoloured Clique, which is W[1]-hard[50]. Assuming zero node and unit edge weights, the above construction implies that for any $\epsilon > 0$, there is no polynomial time algorithm that approximates the maximum assignment weight to within a factor better than $O(n^{1-\epsilon})$, unless P = NP (ref. [51]). Furthermore, finding an assignment of weight k cannot be done in time $n^{o(k)}$, unless the exponential time hypothesis fails[52,53]. Finally, we noted above that we can encode an arbitrary edge set $E \subsetneq E^*$ using zero edge weight for all $e \notin E$, so:

Corollary 1. *The Multicolored Subgraph problem is NP-complete, even for a complete assignment graph, binary edge weights and zero node weights.* Finally, we consider two problem variants. First, we may allow that some colours from $\mathcal{C}$ are absent from A; in this case, A is called colourful. We can encode this variant in the original problem, by adding a dummy node for each colour that is connected to no other node. Second, we may assume that only edges carry weight. We can encode the Multicoloured Subgraph problem in this variant, by adding a dummy colour for each colour and a dummy node for each node, such that if a node has a certain colour, then the dummy node has the corresponding dummy colour. We connect each node to its dummy node, and transfer the weight of the node to the corresponding edge. Hence, our complexity results also hold for these variants.

On the algorithmic side, it is easy to see that the Multicoloured Subgraph problem can be solved by a simple integer linear program (one variable per edge and one variable per colour). We omit the straightforward technical details. We will not proceed in this direction, as this approach results in a single optimal solution, whereas we want to consider suboptimal solutions and marginal probabilities, which allow us to judge our individual confidence when assigning molecular formulas to compounds.

Likelihoods, prior probabilities and graph topology. The likelihood $\mathbb{P}(\mathcal{D}(\mathfrak{c}(v))|v)$ of a molecular formula candidate v can be computed from the posterior probability of the fragmentation tree and the isotope pattern analysis as estimated by SIRIUS 4.0 (refs. [7,25]). For the Gibbs sampler, we treat these probabilities as likelihoods, although the analysis SIRIUS 4.0 also integrates certain priors[25]. To guarantee rapid computations, we usually limit further computations to the, say, 50 best-scoring molecular formulas for each compound. For each compound, we also introduce a node representing 'molecular formula not identified' that receives likelihood from the remaining molecular formulas, and is not connected to any other nodes.

Furthermore, we assume that some compounds were identified by searching in a library of tandem MS spectra, plus potentially by comparison of retention times. We refer to these compounds and the corresponding molecular formulas as 'anchors'. Such library search results can also be wrong, so we do not exclude other molecular formula explanations, but rather give a bonus to the likelihood of the identified molecular formula. The 'quality' of a spectral library hit can, to a certain extent, be evaluated using its score, usually the dot product (cosine score) between query and reference. Hence, the bonus may be dependent on the corresponding library search score. Given the library search score $s_l \in [0, 1]$ and a minimum score to consider a library hit $\min_l$, we multiply the candidate's likelihood by

$$\psi(s_l, \min_l) = \exp\left(\lambda \cdot \frac{\max\{s_l, \min_l\}}{1 - \max\{s_l, \min_l\}}\right), \quad (5)$$

where $\lambda > 0$ is a fixed weighting parameter. Candidates that disagree with the library hit or without any library hit are scored using $s_l = \min_l$. We note that any 'perfect match' with score of 1.0 will be chosen in any case. We remove any other candidate for this compound. We refrain from normalizing ψ to one.

For estimating priors, we will consider similarity of fragmentation patterns[31,32]: More precisely, we use similarity between fragmentation trees that were computed by SIRIUS in the previous step. For each pair of compounds, we have to compare up to 50 times 50 fragmentation trees: for swift computations, we refrain from using fragmentation tree alignments[54] but instead, simply count the number of common fragments and precursor (root) losses in the two trees[54]. Evaluations indicate that this method, while performing worse than fragmentation tree alignments, is still able to detect structural similarity between compounds[54]. When counting common root losses, the empty root loss is ignored. We introduce two modifications to the score from ref. [54]: let n_1, n_2 be the sizes of the two fragmentation trees, defined by the number of fragments and root losses. Instead of normalizing the number of common fragments plus root losses s by the size of the smaller tree $\min\{n_1, n_2\}$, we use

$$s/n_1 + s/n_2 \tag{6}$$

as the normalized score; by doing this, we slightly penalize large trees, because having common fragments or root losses is more likely against a large tree than a small tree. But this score favours small trees and, hence, inferior molecular formula candidates. To this end, we use the size of the largest fragmentation tree, among all candidate molecular formulas, for the normalization of each compound; this is the maximum number of explainable peaks in the MS/MS data of the compound. Fragments and root losses can be weighted by importance ι. The weight of two common fragments or root losses m_1 and m_2 is $\iota(m_1)\iota(m_2)$. The weighted size of a tree is

$$n_w = \sum\nolimits_{g\in F}(\iota(g)) + \sum\nolimits_{h\in R}(\iota(h)) \tag{7}$$

with fragments F and root losses R. For two molecular formulas $u, v \in V$ we denote the resulting score as $s(u, v)$.

How can we transform this count into a prior probability? Natural choices include significance estimates such as P values and posterior error probabilities. We do not have a reasonable model for the score distribution of 'true' edges; in fact, it is not known how to clearly distinguish between 'true' and 'false' edges in such a model. To this end, we resort to a simple prior based on P value estimation:

$$\mathbb{P}(u, v|'\text{true}') = f(\tau) \text{ and } \mathbb{P}(u, v|'\text{false}') = \begin{cases} f(s(u,v)) & \text{if } s(u,v) \geq \tau, \\ f(\tau) & \text{otherwise}, \end{cases} \tag{8}$$

where $\tau \in \mathbb{R}$ is a thresholding parameter, and $f : \mathbb{R} \to [0, 1]$ is a monotonically decreasing function. We introduce threshold τ because scores below a certain threshold are uninformative in practice and should not be considered in our estimations. For $f(x)$ we estimate the P value of score x, under the null model that scores follow a certain distribution. We note that prior probabilities do in fact depend upon the MS data.

We now assign node and edge weights according to equation (4). Clearly, many of these edges have zero weight and can be removed from the graph. To avoid nodes being isolated, we want to keep some edges incident to any node. This can be formulated by individual thresholds $\tau_c \in \mathbb{R}$ for each colour $c \in \mathcal{C}$ and, for an edge uv, edge weight

$$w(uv) := \max\{0, -\log f(s(u, v)) + \log f(\tau_{uv})\}$$

for threshold $\tau_{uv} := \min\{\tau_{c(u)}, \tau_{c(v)}\}$. This will change the weight of any assignment by an additive constant and, hence, posterior probability by a multiplicative constant.

(Faster) Gibbs sampling. We say that a node v is active in an assignment A if $v \in A$, and that an edge uv is active if both $u \in A$ and $v \in A$; then, the weight of an assignment is the sum of weights of all active nodes and edges.

Gibbs sampling is a Markov chain Monte Carlo algorithm for obtaining a sequence of observations approximated from a multivariate probability distribution[55]. Sampling assignments according to (2) can be seen as an archetype application of a Gibbs sampler: We start with some assignment, such as the highest likelihood node (molecular formula) for each compound (colour). Each epoch of the Gibbs sampler consists of $|\mathcal{C}|$ *steps*, where we iterate over all colours $c \in \mathcal{C}$ in random order: We update the active node with colour c by drawing a node with colour c according to its posterior probability, conditional the current assignment of all nodes with colour different from c. At the end of the epoch we output the current assignment, and repeat until we have reached a sufficient number of samples. This generates a Markov chain of samples converging to the posterior probability distribution of assignments. In practice, we discard samples from the beginning of the chain (burn-in period), and to avoid correlation between nearby samples, we output only every, say, tenth sample.

Assume that $u \in \mathcal{A}$ with colour $c := \mathfrak{c}(u)$ is to be (potentially) replaced by a new node v with the same colour. The probability of $v \in V(c)$, conditional all other nodes $z \in A$ with $\mathfrak{c}(z) \neq c$, can naively be computed as

$$\mathbb{P}(v|A - \{u\}) \propto \exp\Big(w(v) + \sum\nolimits_{z\in A, vz\in E} w(vz)\Big). \tag{9}$$

Computing all conditional probabilities for drawing a node v, requires time proportional to the sum of node degrees for all nodes from $V(c)$. That means running time for one step is of order $O(|V(c)| \cdot |V|)$ and, hence, $\Theta(|V|^2)$ for certain graph families.

To apply Gibbs sampling in practice, the critical point is to quickly reach a large number of samples, so that probability estimates become reliable. To further decrease running time, we assume that we have, at any step, knowledge about all (log) conditional probabilities, for all nodes $v \in V(c)$ and all colours $c \in \mathcal{C}$. We assume that conditional probabilities are not normalized; to sample a new active node, we uniformly draw a random number between zero and the sum of conditional probabilities, over all nodes with this colour. To improve the sampling speed, we want to estimate conditional probabilities without performing a full calculation using (9).

Lemma 2. One step of the Gibbs sampler, exchanging some node u by another node v with the same colour, can be carried out in $O(|V(c)| + \deg(u) + \deg(v))$ time.

Proof of Lemma 2. Let $A \subseteq V$ be the current assignment with $u \in A$. We want to choose a new node $v \in U$ from the set of candidate nodes $U := V(c)$ for colour $c := \mathfrak{c}(u)$. We know the conditional probabilities $\mathbb{P}(v|\mathcal{A} - \{u\})$ for all $v \in U$; we sum up the conditional probabilities, then uniformly choose a random number between zero and this sum and, finally, use this random number to select one $v \in U$. This can be carried out in time $O(|U|)$. If $u = v$ then we can stop at this point.

Second, we have to estimate conditional probabilities for all nodes $z \in V$. From equation (9), we infer that the conditional probability only changes for those nodes z where there is a change in the neighbourhood $N(z)$ of z, and remains constant for all others. To this end, we iterate over all $z \in N(u)$, and decrease the log conditional probability of z by $w(uz)$; then, we iterate over all $z \in N(v)$, and increase the log conditional probability of z by $w(vz)$. Finally, for any node $z \in N(u) \cup N(v)$, we recompute its conditional probability using the exponential function. This can be carried out in time $O(\deg(u) + \deg(v))$; afterwards, all conditional probabilities are correct for the new assignment $\mathcal{A} - \{u\} \cup \{v\}$.□

Comparing a naive graph-based implementation of a Gibbs sampler with one that uses Lemma 2, we can estimate that the speedup is of order $\Theta(|V(c)|)$.

For the first iteration, we use an arbitrary assignment, then compute all conditional probabilities using equation (9). The method requires $O(|V| + |E|)$ memory for storing the graph, and $O(|V|)$ memory for storing (log) conditional probabilities. The probability of a particular molecular formula v being correct can now be estimated as its marginal probability: that is, the ratio of assignments in the output that contain v.

ZODIAC parameters. We use identical parameters for all five datasets; see equation (5) above. We weight fragments and root losses when comparing fragmentation trees of molecular formula candidates. Here, we use the SIRIUS 4 noise intensity scoring as importance ι in equation (7). The probability that a peak p that corresponds to a fragment and root loss is not noise is $\iota = 1 - \text{par}(\text{int}(p))$, where par is the Pareto cumulative distribution function with $x_{\min} = 0.002$, $x_{\text{median}} = 0.015$ and $\text{int}(p) \in [0, 1]$ the relative peak intensity, see ref. [25]. To establish a threshold for the minimal similarity of fragmentation trees, we decrease score s and tree sizes n_1 and n_2 each by 1.0; see equation (6).

The empirical score distributions resemble a log-normal distribution (see Supplementary Fig. 7) so we use its cumulative distribution function to estimate P values for equation (8). For the robust estimation of parameters μ and σ^2, we sampled 100,000 non-zero scores for each dataset, and used the median score as parameter μ and the median absolute deviation as parameter σ^2. We naturally expect most edges to be false edges and chose the score threshold τ so that 95 % of the non-zero scores are smaller than this threshold. Finally, we used individual thresholds for each compound (colour) so that at least ten molecular formulas of this colour are incident to ten or more edges.

Each molecular formula candidate of some compound receives a score $s_1, \ldots, s_n$, where $s_{\max}$ is the largest score. We transformed SIRIUS scores to probabilities using the softmax function, where $p_j = \exp(s_j - s_{\max})$ are normalized to sum to one. To adjust for the fact that the correct molecular formula may not be in the top 50, we added a dummy node receiving the combined probability of all unconsidered candidates. Dummy nodes are not connected to any other node. SIRIUS does not report the score of all candidates, as one compound may have tens of thousands of candidates. Hence, we estimated the probability of all unconsidered candidates by multiplying the number of unconsidered candidates to the lowest probability of the top 50 candidates.

Finding and scoring ZODIAC anchors. ZODIAC can use (potentially incorrect) spectral library hits as anchors to improve annotations. To find a reasonable number of anchors, we performed spectral library search in analogue mode. Resulting molecular formula annotations are not considered ground truth identifications but are sufficient as anchors. Only those hits were considered that have mass differences between query and reference corresponding to a frequent biotransformation; this assumption is only valid for biological datasets. We used the following molecular formula differences as valid biotransformations[26,56,57]: C_2H_2, C_2H_2O, C_2H_3NO, $C_2H_3O_2$, C_2H_4, C_2H_4, C_2O_2, $C_3H_2O_3$, C_3H_5NO, $C_3H_5NO_2$, C_3H_5O, $C_4H_2N_2O$, $C_4H_3N_3$, $C_4H_4O_2$, C_5H_7, C_5H_7NO, C_5H_9NO, CH_2, CH_2ON, CH_3N_2O, CHO_2, CO, CO_2, H_2, H_2O, N, NH, NH_2, NH_3, O, H_4O_2 and H_6O_3. This list is probably not comprehensive; we use a small list to ensure that the approach is conservative, avoiding spurious hits. As noted, false anchor identifications will be tolerated by ZODIAC. Also remember that ZODIAC is not using these biotransformations itself, and that ZODIAC can also be executed without anchors.

We used identical parameters for all five datasets. When scoring anchors according to equation (5), we used the maximum of the cosine score between the spectrum and the cosine score of the mirrored spectrum as the similarity measure, and $\min_l = 0.5$ as the score threshold parameter and $\lambda = 1{,}000$ as the weighting parameter. For anchors found by spectral library match in analogue mode (that is, non-identical m/z), spectral similarity was reduced by 0.1 to account for increased uncertainty.

Searching for anchors as described above resulted in 96 anchors for the dendroides dataset, 254 anchors for the NIST1950 dataset, 749 for the tomato dataset, 372 for the diatoms dataset and 176 for the mice stool dataset. All spectral hits described in the previous section are anchors, too; remember that for dendroides, the ground truth was established manually and those annotations do not serve as anchors.

Burn-in and number of Gibbs sampling epochs. We determined a reasonable number of Gibbs sampling iterations using the dendroides dataset. One iteration, or 'epoch', is defined as one round in which each compound is updated once by choosing a new 'active' molecular formula candidate. We run ten independent Markov chains (Fig. 6): the total score summed over all active candidate at a specific epoch increases swiftly over the first 500 epochs. Similarly, the number of correct annotations at a specific epoch increases quickly for most Markov chains until the chain seems to remain at a local optimum. We note that this number of correct molecular formula is determined at each epoch, whereas ZODIAC scores are computed from the average over many epochs. From this data, we estimated a burn-in of 1,000 epochs and sampling of 2,000 iterations. Larger values increase running times but should never worsen results.

In application, we used ten Markov chains in parallel, a burn-in of 1,000 epochs, and sampled 2,000 epochs; we retained only every tenth sample. This resulted in a total of $10 \times 200 = 2,000$ samples.

Evaluation against competing methods. *Seven Golden Rules*. We analysed whether molecular formula annotations adhere to the Seven Golden Rules by Kind and Fiehn[9]. First, we evaluated whether ZODIAC molecular formula annotations adhere to the Seven Golden Rules; this concerns only those rules involved with molecular formulas. Second, we used the Seven Golden Rules to rank candidate molecular formulas, to evaluate against SIRIUS and ZODIAC.

To evaluate ZODIAC annotations (Extended Data Fig. 5), we used the Seven Golden Rules as a filter, ignoring the measured isotope pattern. The original version of the Seven Golden rules does not include the elements iodine, boron and selenium, but these are straightforward to include in the RDBE computation of the valency filter. Additionally to the valency filter, we applied the common range of element ratios and element probability check[9]. We applied the Seven Golden Rules to the neutral molecular formula of the best ZODIAC hit.

For evaluation of annotation rates (Extended Data Fig. 4), we used the Excel sheet implementation version 46 provided by the authors of ref. [9], available from https://fiehnlab.ucdavis.edu/projects/seven-golden-rules/software. We assumed the same adduct candidates and isotope patterns as described for SIRIUS and ZODIAC; if multiple adduct candidates were considered, we searched for the candidate with the overall highest score among these alternatives. We either used elements CHNOPS, or CHNOPS plus bromine and chlorine. The NIST1950 data contains five chlorinated ground truth compounds, which are the only ground truth compounds in all datasets that are not solely comprised of CHNOPS. Excluding the other elements makes it easier for the Seven Golden Rules to annotate the correct molecular formula, since fewer candidates have to be considered. We use the same mass accuracy as assumed for SIRIUS computations.

Exact mass search and GenForm. We also evaluated ZODIAC against annotation by exact mass and GenForm (Extended Data Fig. 4). GenForm (https://sourceforge.net/projects/genform/) is an open-source implementation of MOLGEN-MS/MS[58]. For all methods we assumed the same adduct candidates as described for SIRIUS and ZODIAC (undocumented adduct M + K for kalium adducts). We required molecular formulas from exact mass annotation to have an RDBE value of −1 or greater. We considered only elements CHNO or CHNOPS for exact mass, to reduce the number of candidates; for larger sets of elements, exact mass performs worse. For GenForm, we merged all MS/MS spectra of one compound into a single MS/MS spectrum assuming 15 ppm mass accuracy and use the same isotope pattern as for SIRIUS and ZODIAC. We enabled the RDBE filter ('exist' option) and set the 'rej' and the 'ppm' parameters to the same mass accuracy as assumed for SIRIUS computations. We used elements CHNOPS and tested different values for parameters MS1 match value 'msmv' and MS/MS intensity weighting 'wi'. We found that the parameter settings 'msmv=ndp' and 'wi=lin', which are reported here, performed slightly better than other settings, but this is inconsistent between datasets, and differences are not substantial. We also evaluated the Heuerding and Clerk[59] filter option ('hcf') on element ratios but found that it does not change results substantially; notably, this filter discards between one in four and three in four correct molecular formulas.

Evaluation of molecular structure annotation. We evaluated how improved molecular formula annotations affect the number of correct molecular structure identifications with CSI:FingerID. For four of the datasets (NIST1950, tomato, diatoms and mice stool), ground truth structure annotations were established by spectral library search. We limited our evaluation to those hits with cosine score greater of equal to 0.8, since the ground truth may contain more false-positive structure hits than false molecular formula hits. We used only structures with adducts $[M + H]^+$, $[M + Na]^+$ and $[M + K]^+$, since our molecular formula annotation did not take further adducts into account. This resulted in 60 ground truth structures for the NIST1950 dataset, 88 for the tomato dataset, 47 for the diatoms dataset and 19 for the mice stool dataset. We searched query compounds with CSI:FingerID version 1.1.3 in PubChem using the top-ranked SIRIUS or ZODIAC molecular formula annotation. Two molecular structures are assumed to be identical if they have the same constitution; this is tested by comparing the first 14 characters of the InChIKey. For 94.39% of the query compounds, an MS/MS spectrum with the same molecular structure was contained in the training data of CSI:FingerID. On the NIST1950 dataset, we found the correct structure for both SIRIUS and ZODIAC molecular formula in 76.67% of the cases; on the tomato dataset, we correctly annotated 69.32% with SIRIUS and 71.59% with ZODIAC; on the diatoms dataset, we correctly annotated 65.96% with SIRIUS and 72.34% with ZODIAC; and on the mice stool dataset, we correctly annotated73.68% with SIRIUS and 78.95% with ZODIAC. Merging all search results, we annotated the correct structure for 71.03% using SIRIUS molecular formulas and for 73.83% using ZODIAC molecular formulas. In practice, SIRIUS uses a 'soft threshold' to counter the effect of falsely annotated molecular formulas[36]. We note that reported structure annotation rates are 'inflated', in the sense that we did not remove data for the correct structure from the CSI:FingerID training data[36].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Input mzML/mzXML files for the five datasets are available at MassIVE (https://massive.ucsd.edu/), with the following accession numbers for dendroides (MSV000080502), for NIST1950 (MSV000081364), for tomato (MSV000081463), for diatoms (MSV000081731) and for the mice stool (MSV000079949) datasets. SIRIUS and ZODIAC results and a virtual machine on which to reproduce the data are available from https://bio.informatik.uni-jena.de/data/ and https://doi.org/10.6084/m9.figshare.12911171. Source data are provided with this paper.

Code availability

ZODIAC has been integrated into the SIRIUS software and is written in Java. It is open source under the GNU General Public License (version 3), and works on Windows, macOS X and Linux. A command-line version allows batch processing and results can be visualized in a graphical user interface. We provide executable binaries, example files and additional information on the ZODIAC website (https://bio.informatik.uni-jena.de/software/zodiac/). A source copy is hosted on GitHub (https://github.com/boecker-lab/sirius-libs[60]); the branch 'zodiac_in_sirius_4_release' contains the SIRIUS and ZODIAC code used for evaluation in this paper.

Received: 1 April 2020; Accepted: 4 September 2020;
Published online: 13 October 2020

References

1. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucl. Acids Res.* **46**, D608–D617 (2018).
2. Kim, S. et al. PubChem substance and compound databases. *Nucl. Acids Res.* **44**, D1202–D1213 (2016).
3. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **8**, 3 (2016).
4. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
5. Schymanski, E. L. et al. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.* **9**, 22 (2017).
6. Samaraweera, M. A., Hall, L. M., Hill, D. W. & Grant, D. F. Evaluation of an artificial neural network retention index model for chemical structure identification in nontargeted metabolomics. *Anal. Chem.* **90**, 12752–12760 (2018).
7. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
8. Dührkop, K. et al. Classes for the masses: systematic classification of unknowns using fragmentation spectra. Preprint at https://www.biorxiv.org/content/10.1101/2020.04.17.046672v1 (2020).
9. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.* **8**, 105 (2007).
10. Stein, S. E. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.* **84**, 7274–7282 (2012).
11. Vinaixa, M. et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *Trends Anal. Chem.* **78**, 23–35 (2016).
12. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
13. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
14. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
15. Guijas, C. et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
16. Alon, T. & Amirav, A. Isotope abundance analysis methods and software for improved sample identification with supersonic gas chromatography/mass spectrometry. *Rapid Commun. Mass Spectrom.* **20**, 2579–2588 (2006).

17. Böcker, S., Letzel, M., Lipták, Z. S. & Pervukhin, A. Decomposing metabolomic isotope patterns. In *Proc. Works. Algorithms in Bioinformatics (WABI 2006)* Vol. 4175,12–23 (Springer, Berlin, 2006).
18. Ojanperä, S. et al. Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **20**, 1161–1167 (2006).
19. Böcker, S., Letzel, M., Lipták, Zs & Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **25**, 218–224 (2009).
20. Pluskal, T., Uehara, T. & Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal. Chem.* **84**, 4396–4403 (2012).
21. Valkenborg, D., Mertens, I., Lemière, F., Witters, E. & Burzykowski, T. The isotopic distribution conundrum. *Mass Spectrom. Rev.* **31**, 96–109 (2012).
22. Loos, M., Gerber, C., Corona, F., Hollender, J. & Singer, H. Accelerated isotope fine structure calculation using pruned transition trees. *Anal. Chem.* **87**, 5738–5744 (2015).
23. Böcker, S. & Rasche, F. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* **24**, i49–Ii55 (2008).
24. Stravs, M. A., Schymanski, E. L., Singer, H. P. & Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J. Mass Spectrom.* **48**, 89–99 (2013).
25. Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *J. Cheminform.* **8**, 5 (2016).
26. Rogers, S., Scheltema, R. A., Girolami, M. & Breitling, R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* **25**, 512–518 (2009).
27. Daly, R. et al. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* **30**, 2764–2771 (2014).
28. da Silva, R. R. et al. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics* **30**, 1336–1337 (2014).
29. Del Carratore, F. et al. Integrated probabilistic annotation: a Bayesian-based annotation method for metabolomic profiles integrating biochemical connections, isotope patterns, and adduct relationships. *Anal. Chem.* **91**, 12799–12807 (2019).
30. Tziotis, D., Hertkorn, N. & Schmitt-Kopplin, P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. *Eur. J. Mass Spectrom.* **17**, 415–421 (2011).
31. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci. USA* **109**, E1743–E1752 (2012).
32. Morreel, K. et al. Systematic structural characterization of metabolites in *Arabidopsis* via candidate substrate-product pair networks. *Plant Cell* **26**, 929–945 (2014).
33. Quinn, R. A. et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
34. Esposito, M. et al. *Euphorbia dendroides* latex as a source of jatrophane esters: isolation, structural analysis, conformational study, and anti-CHIKV activity. *J. Natural Prod.* **79**, 2873–2882 (2016).
35. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
36. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
37. Nothias, L.-F. et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J. Natural Prod.* **81**, 758–767 (2018).
38. Pence, H. E. & Williams, A. ChemSpider: an online chemical information resource. *J. Chem. Ed.* **87**, 1123–1124 (2010).
39. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **11**, 395 (2010).
40. Nothias, L. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
41. Simón-Manso, Y. et al. Metabolite profiling of a NIST standard reference material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal. Chem.* **85**, 11725–11731 (2013).
42. Vos, R. C. H. D. et al. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protocols* **2**, 778–791 (2007).
43. Agarwal, V. et al. Complexity of naturally produced polybrominated diphenyl ethers revealed via mass spectrometry. *Environ. Sci. Technol.* **49**, 1339–46 (2015).
44. Andersen, R. & America, P. S. *Algal Culturing Techniques* (Elsevier Science, 2005).
45. Dittmar, T., Koch, B., Hertkorn, N. & Kattner, G. A simple and efficient method for the solid-phase extraction of dissolved organic matter (SPE-DOM) from seawater. *Limnol. Oceanogr. Meth.* **6**, 230–235 (2008).
46. Petras, D. et al. High-resolution liquid chromatography tandem mass spectrometry enables large scale molecular characterization of dissolved organic matter. *Front. Mar. Sci.* **4**, 405 (2017).
47. Meusel, M. et al. Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal. Chem.* **88**, 7556–7566 (2016).
48. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* **3**, 211–221 (2007).
49. Karp, R. M. in *Complexity of Computer Computations* (eds Miller, R. E. & Thatcher, J. W.) 85–103 (Plenum Press, 1972).
50. Downey, R. G. & Fellows, M. R. *Parameterized Complexity* (Springer, Berlin, 1999).
51. Zuckerman, D. Linear degree extractors and the inapproximability of max clique and chromatic number. In *Proc. ACM Symp. on Theory of Computing (STOC 2006)* 681–690 (2006).
52. Chen, J., Huang, X., Kanj, I. A. & Xia, G. Strong computational lower bounds via parameterized complexity. *J. Comp. Syst. Sci.* **72**, 1346–1367 (2006).
53. Impagliazzo, R. & Paturi, R. On the complexity of k-SAT. *J. Comp. Syst. Sci.* **62**, 367–375 (2001).
54. Rasche, F. et al. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.* **84**, 3417–3426 (2012).
55. Geman, S. & Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
56. Ludwig, M., Dührkop, K. & Böcker, S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics* **34**, i333–i340 (2018).
57. Li, L. et al. MyCompoundID: using an evidence-based metabolome library for metabolite identification. *Anal. Chem.* **85**, 3401–3408 (2013).
58. Meringer, M., Reinker, S., Zhang, J. & Muller, A. MS/MS data improves automated determination of molecular formulas by mass spectrometry. *MATCH Commun. Math. Comput. Chem.* **65**, 259–290 (2011).
59. Heuerding, S. & Clerc, J. T. Simple tools for the computer-aided interpretation of mass spectra. *Chemometr. Intell. Lab. Syst.* **20**, 57–69 (1993).
60. Dührkop, K. et al. boecker-lab/sirius-libs: SIRIUS 4.0.1 including ZODIAC (Version v4.0.1_with_ZODIAC). https://doi.org/10.5281/zenodo.3985859 (2020).

Acknowledgements

We thank M. Witting for discussions and F. Kretschmer for the fragmentation tree visualization. We acknowledge financial support by the Deutsche Forschungsgemeinschaft to S.B., K.D., M.F., M.A.H. and M.L. (grant BO 1910/20) and D.P. (grant PE 2600/1). I.K. acknowledges funding from the Blasker Environmental Grant, San Diego Foundation. F.V. was funded by the Department of Navy, Office of Naval Research Multidisciplinary University Research Initiative (MURI) Award (award number N00014-15-1-2809). L.-F.N. was supported by European Union's Horizon 2020 grants (MSCA-GF, 704786). M.M. acknowledges funding from the National Science Foundation (award number 1354050). We acknowledge financial support by the US National Institutes of Health to P.C.D. for the Center for Computational Mass Spectrometry (grant P41 GM103484), the re-use of metabolomics data (grant R03 CA211211) and the tools for rapid and accurate structure elucidation of natural products (grant R01 GM107550). P.C.D. also acknowledges support from the Sloan Foundation and from the Gordon and Betty Moore Foundation.

Author contributions

S.B. designed the research. S.B. and M.L. developed the computational method with help from K.D. M.L. implemented the computational method with contributions from K.D. and M.F. M.L. and L.-F.N. performed the method evaluation, coordinated by S.B. L.-F.N., I.K. and L.A. contributed to the interpretation of results. M.F. and M.L. integrated ZODIAC into SIRIUS. M.A.H. contributed to the visualization of the novel compound's data. Mass spectrometry experiments were performed for the dendroides dataset by L.F.N., for the NIST1950 dataset by F.V., for the tomato dataset by L.-F.N and M.M. and for the diatoms dataset by I.K. and D.P. L.A. and P.C.D. coordinated the experimental part of the study. S.B. and M.L. wrote the manuscript, to which L.F.N. and I.K. contributed, in cooperation with all other authors.

Competing interests

S.B, K.D., M.F., M.A.H. and M.L. are founders of Bright Giant GmbH. P.C.D. is the scientific advisor for Sirenas LLC.

Additional information

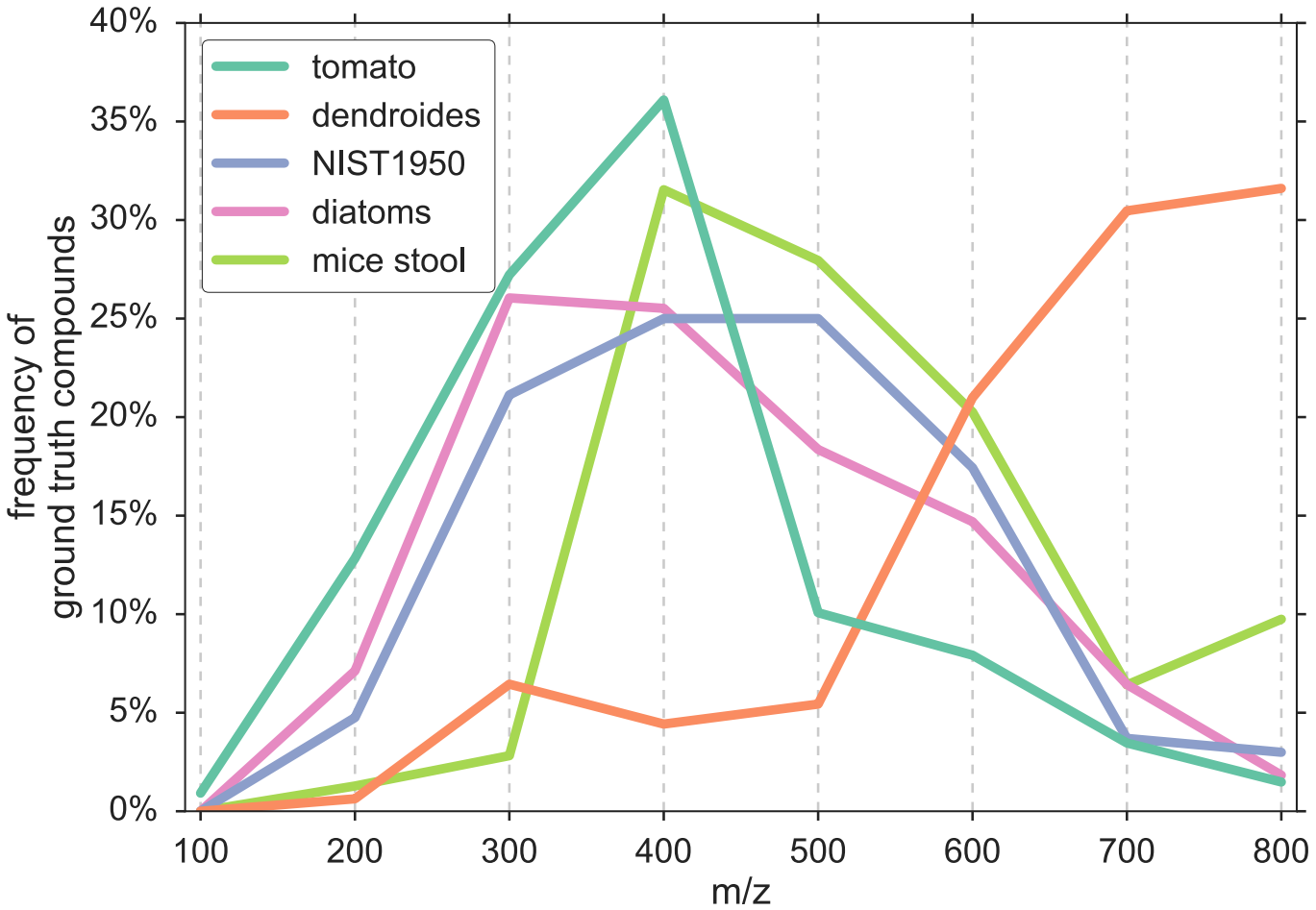
Extended data is available for this paper at https://doi.org/10.1038/s42256-020-00234-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00234-6.

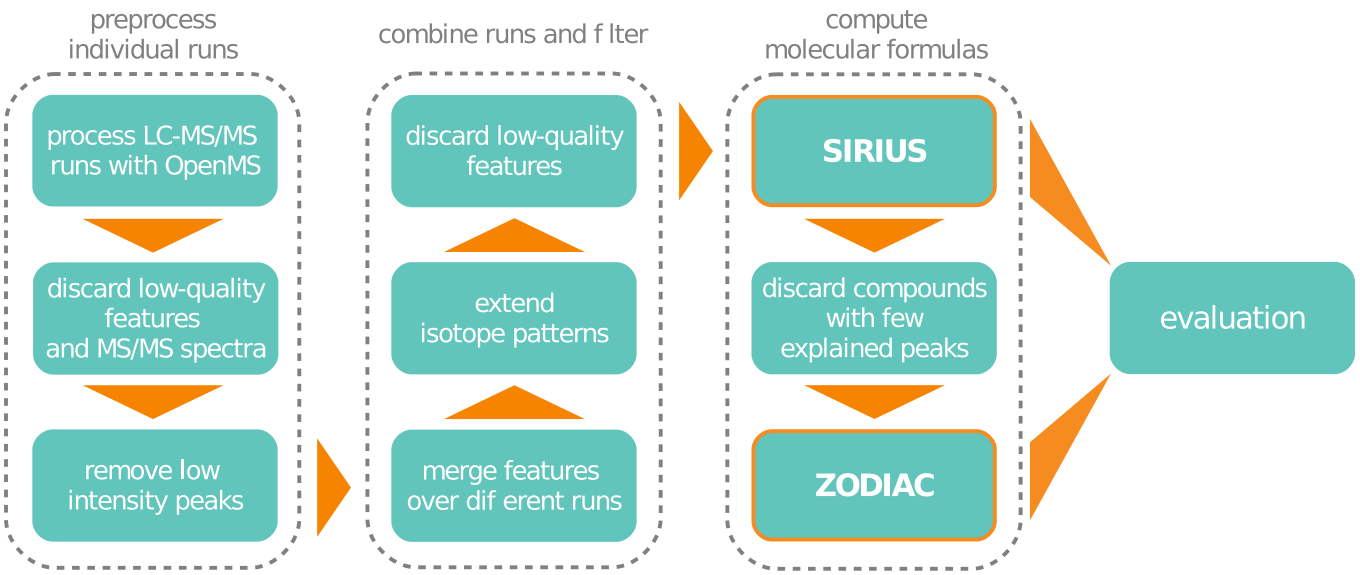
Correspondence and requests for materials should be addressed to S.B.

dataset	total compounds	ground truth	# in top 50	1st quartile m/z	median m/z	3rd quartile m/z	mass error (ppm) max	mass error (ppm) STD
dendroides	792	201	197	605.310	705.274	759.353	7.40	2.85
NIST1950	571	94	94	286.390	373.800	477.335	3.79	1.83
tomato	2902	271	270	207.814	271.713	334.526	6.75	1.32
diatoms	2472	93	93	253.195	301.216	349.237	2.07	1.09
mice stool	398	44	43	373.274	454.292	516.298	5.95	1.84

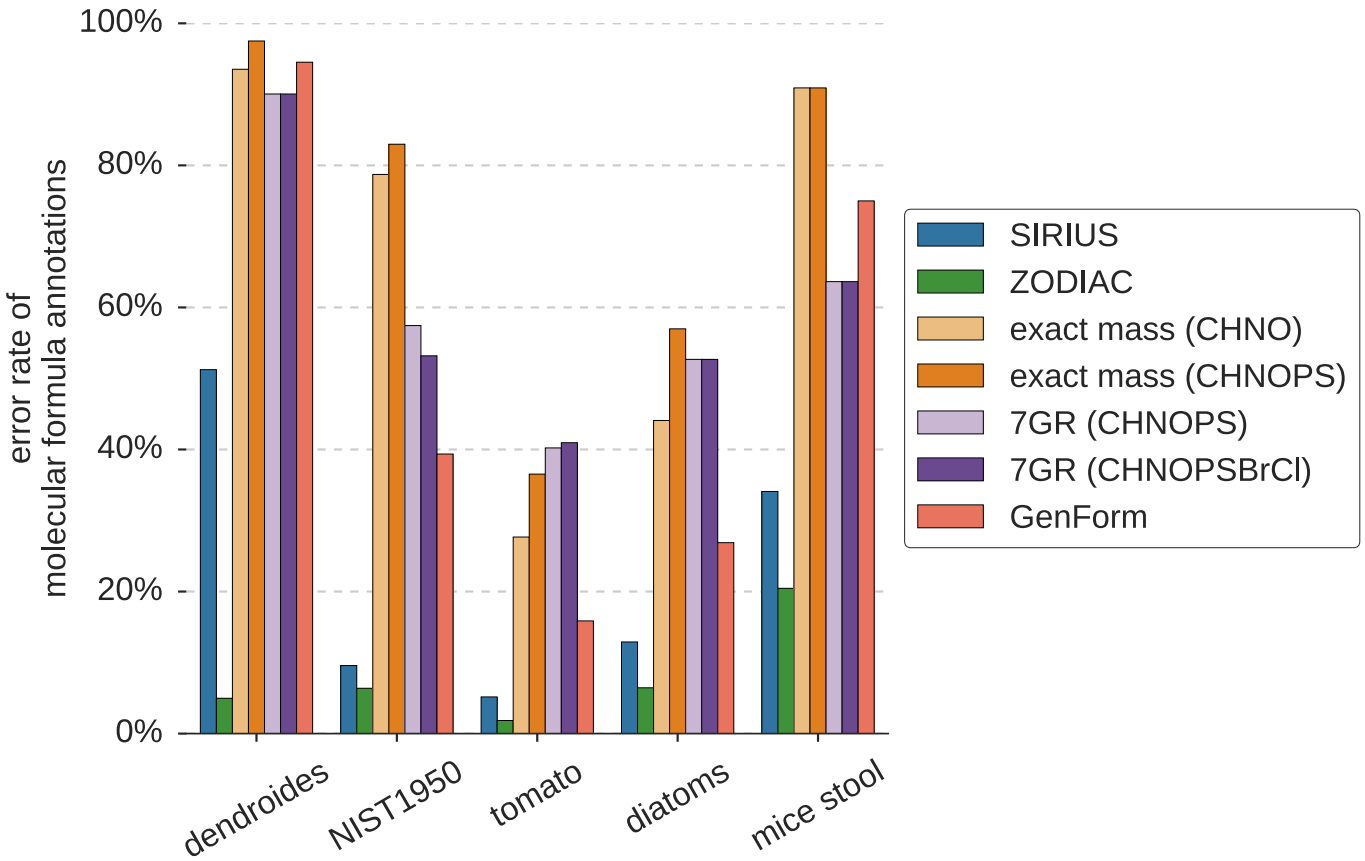
Extended Data Fig. 1 | Statistics on compounds with annotated ground truth molecular formulas. Given is the number of total compounds, the number of compounds with a ground truth molecular formula and the number which are in the top 50 of SIRIUS-ranked candidates. The median *m/z* and 25 and 75 percentile considers only candidates in the top 50. We report the maximum absolute value of all relative mass errors in a dataset. Finally, sample standard deviations (STD) of relative mass errors are computed assuming a mean mass error of zero.



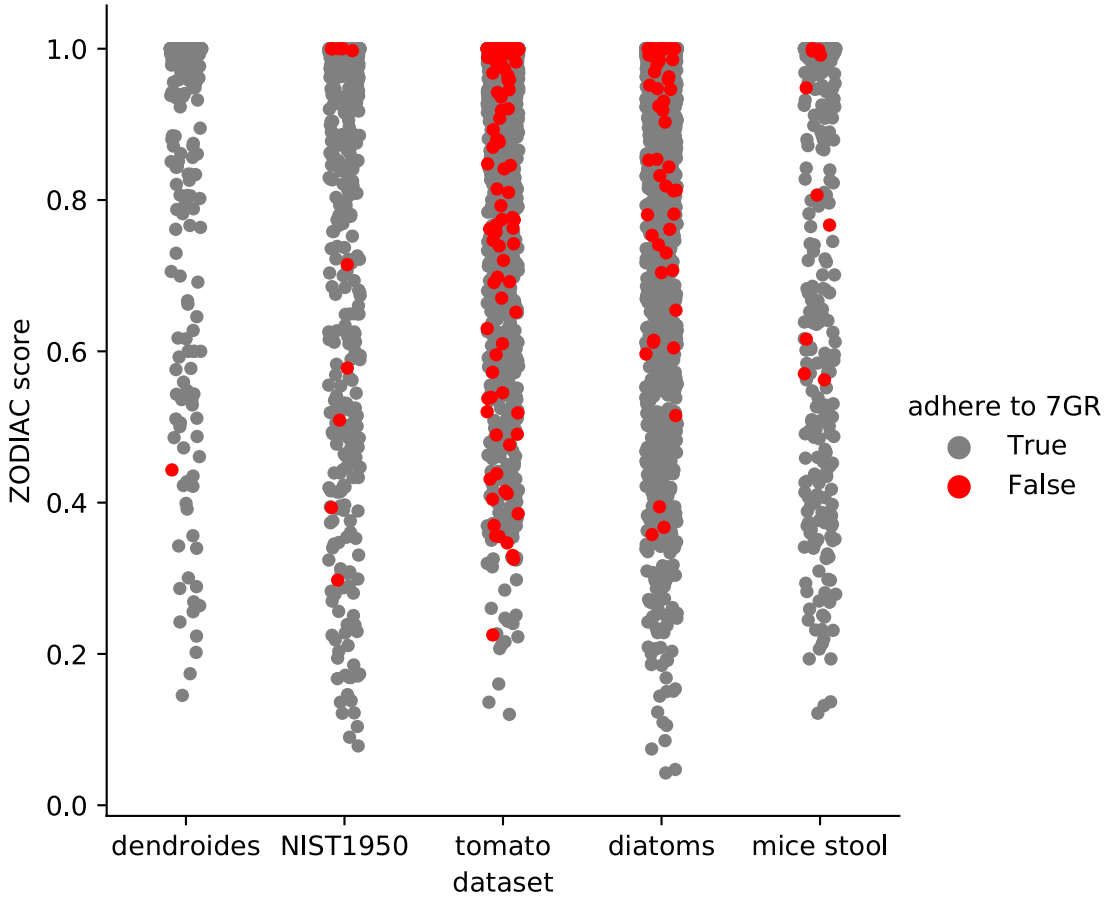
Extended Data Fig. 2 | Distribution of compound masses. Distribution of precursor ion *m/z* of the compounds used as ground truth for the evaluation of the molecular formula annotation on the five datasets. Bins of width 100 are centred at 100, 200, …, 800 *m/z*.



Extended Data Fig. 3 | ZODIAC processing and evaluation workflow. (1) Each LC-MS/MS run is processed individually; input mzML/mzXML files are processed using OpenMS, performing feature and adduct detection and producing files in SIRIUS input format. Resulting features combine MS1, MS/MS and adduct information. (2), (3) Filtering is performed on feature, MS/MS and peak level. (4) Similar features are merged between different runs using hierarchical clustering; MS/MS are combined and a best isotope pattern is selected per feature. (5) Missing isotope peaks are searched in MS1 spectra to extend isotope patterns. (6) A final feature filtering step is performed; the remaining features are considered as compounds. (7) SIRIUS is executed. (8) Compounds with few explained peaks are discarded, since a badly explained MS/MS spectrum indicates low quality. (9) ZODIAC is run on the remaining compounds. (10) SIRIUS and ZODIAC are evaluated on the same set of compounds.



Extended Data Fig. 4 | Molecular formula annotation error rates. Error rates on five datasets. Methods are SIRIUS; ZODIAC (without anchors); exact mass over elements carbon, hydrogen, nitrogen and oxygen ('exact mass (CHNO)'); exact mass over CHNO plus phosphorus and sulfur ('exact mass (CHNOPS)'); Seven Golden Rules with elements CHNOPS ('7GR (CHNOPS)'); Seven Golden Rules with elements CHNOPS plus bromine and chlorine ('7GR (CHNOPSBrCl)'); and GenForm. Between 44 an 271 compounds were processed per dataset, see Extended Data Fig. 1 for details. GenForm is the only publicly available tool for molecular formula inference besides SIRIUS, and considers both the isotope pattern and the fragmentation spectrum. GenForm was restricted to elements CHNOPS, and 7GR (CHNOPSBrCl) cannot annotate iodine-containing compounds; to this end, only SIRIUS and ZODIAC are in theory capable of annotating the two novel molecular formulas $C_{24}H_{47}BrNO_8P$ and $C_{15}H_{30}ClIO_5$ reported here. Error rates are based on all compounds with established ground truth, resulting in slightly higher error rates for SIRIUS and ZODIAC on dendroides, tomato and mice stool compared to Fig. 1. Error rates on the five datasets agree well with the mass of compounds in the respective dataset, see Extended Data Fig. 1: larger compounds result in substantially more candidates to be considered, in particular for a larger set of elements, and result in worse annotation rates. For evaluation details see the Methods section.



Extended Data Fig. 5 | Seven Golden Rules applied to annotated molecular formulas. For each ZODIAC molecular formula annotation, we test whether it meets the molecular formula subset of the Seven Golden Rules (7GR). Each dot represents one annotated compound; molecular formulas are sorted by ZODIAC score.

dataset	molecular formula	# comp.	max score
NIST1950	$C_{15}H_{33}N_9O_9P_2$	1	0.982
diatoms	$C_{24}H_{47}BrNO_8P$	6	1.0
diatoms	$C_{24}H_{49}BrNO_8P$	3	1.0
diatoms	$C_{24}H_{49}INO_8P$	3	1.0
diatoms	$C_{25}H_{41}ClO_{11}$	3	1.0
diatoms	$C_{12}H_{24}ClIO_4$	1	1.0
diatoms	$C_{15}H_{30}ClIO_5$	1	0.999
diatoms	$C_{16}H_{34}N_3O_5$	1	1.0
diatoms	$C_{19}H_{43}ClN_{10}O_{10}$	1	0.992
diatoms	$C_{19}H_{43}NO_3P_2$	1	1.0
diatoms	$C_{21}H_{41}INO_8P$	1	0.9995
diatoms	$C_{21}H_{43}INO_8P$	1	0.9965
diatoms	$C_{22}H_{48}N_5O_7P$	1	0.991
diatoms	$C_{25}H_{45}O_7PS$	1	0.996
diatoms	$C_{25}H_{49}INO_8P$	1	0.996
diatoms	$C_9H_{19}BN_4O_4$	1	1.0
mice stool	$C_{16}H_{45}N_{10}O_6$	1	0.9915
tomato	$C_{11}H_{10}N_4O_{13}$	3	1.0
tomato	$C_8H_{23}N_2O_{15}P_5$	3	1.0
tomato	$C_{11}H_{14}N_4O_{15}$	2	1.0
tomato	$C_6H_{14}N_2O_{16}P_4$	2	1.0
tomato	$C_6H_{16}N_2O_{13}P_4$	2	0.9985
tomato	$C_{10}H_{20}N_6O_{16}P_2$	1	0.9955
tomato	$C_{20}H_{34}NO_{20}P$	1	0.9925
tomato	$C_{20}H_{51}N_7O_3S$	1	1.0
tomato	$C_4H_{13}N_4O_6P$	1	1.0
tomato	$C_6H_{16}N_2O_{11}P_4$	1	1.0
tomato	$C_8H_{12}N_3O_9P_3$	1	0.995
tomato	$C_8H_{15}N_2O_{14}P_3$	1	0.9935
tomato	$C_8H_{20}NO_{17}P_5$	1	0.9885
tomato	$C_8H_{21}N_7O_8$	1	0.9995
tomato	$C_9H_{16}N_2O_{12}$	1	0.98

Extended Data Fig. 6 | Novel molecular formulas. All molecular formulas are absent from the largest molecular structure databases PubChem and ChemSpider. Only molecular formula annotations with a minimum ZODIAC score of 0.98 are reported such that at least 95% of the MS/MS spectrum intensity is being explained by the SIRIUS fragmentation tree, and at least one molecular formula of the compound is connected to 5 or more compounds. There may be more than one hypothetical compound in an LC-MS run being annotated with one molecular formula, potentially corresponding to different isomers. For such cases, '#comp.' is the number of hypothetical compounds being annotated with the given molecular formula, and 'max score' is the maximum ZODIAC score among these annotations. The corresponding compounds are given in Supplementary Table 5. For 90.00% of the compounds, SIRIUS top-ranks the same molecular formula.

nature research

Corresponding author(s): Sebastian Böcker

Last updated by author(s): Apr 16, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	ZODIAC has been integrated into the SIRIUS software and is written in Java. It is open source under the GNU General Public License (version 3), and works on Windows, macOS X, and Linux. A command-line version allows batch processing and results can be visualized in a graphical user interface. We provide source code, executable binaries, example files, and additional information on the SIRIUS website (https://bio.informatik.uni-jena.de/sirius/). A source copy is hosted on GitHub (https://github.com/boecker-lab/sirius-libs); the branch "zodiac_in_sirius_4_release" contains the SIRIUS and ZODIAC code used for evaluation in this paper.
Data analysis	We use OpenMS (https://www.openms.de/), patch provided.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Input mzML/mzXML files for the five datasets are available at MassIVE (https://massive.ucsd.edu/), with the following accession numbers: dendroides (MSV000080502), NIST1950 (MSV000081364), tomato (MSV000081463), diatoms (MSV000081731) and mice stool dataset (MSV000079949). SIRIUS and ZODIAC results and a virtual machine to reproduce the data are available from https://bio.informatik.uni-jena.de/data/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging