











A spectroscopic test suggests that fragment ion structure annotations in MS/MS libraries are frequently incorrect

Lara van Tetering ¹, Sylvia Spies ¹, Quirine D. K. Wildeman ¹, Kas J. Houthuijs ¹,
Rianne E. van Outersterp ¹, Jonathan Martens ¹, Ron A. Wevers ², David S. Wishart ³, Giel Berden ¹ &
Jos Oomens ^{1,4}✉

Modern untargeted mass spectrometry (MS) analyses quickly detect and resolve thousands of molecular compounds. Although features are readily annotated with a molecular formula in high-resolution small-molecule MS applications, the large majority of them remains unidentified in terms of their full molecular structure. Collision-induced dissociation tandem mass spectrometry (CID-MS²) provides a diagnostic molecular fingerprint to resolve the molecular structure through a library search. However, for de novo identifications, one must often rely on in silico generated MS² spectra as reference. The ability of different in silico algorithms to correctly predict MS² spectra and thus to retrieve correct molecular structures is a topic of lively debate, for instance in the CASMI contest. Underlying the predicted MS² spectra are the in silico generated product ion structures, which are normally not used in de novo identification, but which can serve to critically assess the fragmentation algorithms. Here we evaluate in silico generated MSⁿ product ion structures by comparison with structures established experimentally by infrared ion spectroscopy (IRIS). For a set of three dozen product ion structures from five precursor molecules, we find that virtually all fragment ion structure annotations in three major in silico MS² libraries (HMDB, METLIN, mzCloud) are incorrect and caution the reader against their use for structure annotation of MS/MS ions.

¹Radboud University, Institute for Molecules and Materials, FELIX Laboratory, Toernooiveld 7, 6525ED Nijmegen, The Netherlands. ²Department of Laboratory Medicine, Translational Metabolic Laboratory, Radboud University Medical Center, Geert Grooteplein Zuid 10, 6525GA Nijmegen, The Netherlands. ³Departments of Computing Science and Biological Sciences, University of Alberta, Edmonton, AB, Canada. ⁴van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands. ✉email: jos.oomens@ru.nl

Owing to its high sensitivity and resolution, mass spectrometry (MS) has become indispensable in the detection and identification of molecular species in complex mixtures, e.g., in metabolomics and many other small-molecule applications¹. Modern MS instruments detect and resolve thousands of molecular compounds in a sample in a matter of just seconds. However, most features detected in untargeted MS analyses are not identified in terms of their full molecular structure. While the resolving power is usually sufficient to assign a unique chemical formula to the detected ions, the mass value alone gives little information on the arrangement of the atoms within the molecule. Consequently, definitive identification of the molecular structure remains challenging as it requires structural and stereoisomers corresponding to the same mass-to-charge ratio (m/z) to be distinguished.

Tandem mass spectrometry (MS/MS = MS²) is commonly used to advance structure annotation beyond the chemical formula. Collision-induced dissociation (CID) of a precursor ion selected in the first MS stage produces a structurally diagnostic fragmentation pattern in the second MS stage. To delineate a molecular structure for the precursor ion, this fragmentation pattern is compared against the entries in MS/MS spectral libraries². Many application-specific libraries exist^{2–4}, e.g. for metabolites, agrochemicals, toxicological substances, drug compounds, etc. However, even taken together, MS/MS reference libraries cover only a minute fraction of chemical space, estimated as perhaps 1% or so^{3,5–7}. For *de novo* identifications beyond these ‘known unknowns’, *in silico* strategies to identify structures have been developed. High-level quantum-chemical computation of MS/MS spectra⁸ is developing, but far too costly to screen large numbers of candidate structures. Much faster methods originally relied mostly on rule-based^{9,10} and combinatorial^{9,11} fragmentation approaches⁷, while more recently, these heuristic models

are being updated with strategies involving elements of machine learning^{12–16}.

The combinatorial approach (Fig. 1) to predict MS/MS spectra is still widely in use and underlies some of the more recent machine-learning strategies^{13,14}. A large compound database is screened for entries with the accurate mass (MS¹) of the unknown query molecule. For each hit, a quick rule-based algorithm first determines possible small neutral losses (H₂O, NH₃, etc.), which are common in MS/MS and often occur early on in the CID breakdown cascade. A list of possible fragment m/z values is then generated *in silico* by breaking each bond in the molecule (excluding X-H bonds). The resulting combinatorial fragment m/z 's are fed back into the *in silico* fragmenter for a second (and third) round of fragmentation, forming so-called fragmentation trees (not to be confused with experimental MSⁿ spectral trees). Each type of chemical bond represents a preset bond dissociation energy, so that the probability of generating a specific m/z fragment—and hence its relative intensity in the MS/MS spectrum—can be quantified from the summed bond dissociation energies along the fragmentation tree ending up at that m/z . The resulting *in silico* MS/MS spectrum is then matched against the query MS/MS spectrum.

A caveat involves the neglect of possible structural rearrangement occurring upon bond dissociation^{7,10}. Rearrangement constitutes the replacement of broken bonds with new ones. In the *in silico* fragmenter, rearrangement would thus lead to a different set of combinatorial fragments being generated upon a next round of fragmentation. Neglecting rearrangement, as indicated by the question marks in Fig. 1, may thus lead to fragment ion m/z values in the *in silico* MS/MS spectrum that are incorrect. Especially rearrangements occurring early on in the fragmentation tree may strongly affect the entire downstream tree.

Commonly used *in silico* fragmenters include MetFrag, which generates MS/MS spectra for the METLIN (<https://metlin.scripps.edu>) database^{17,18} and Competitive Fragmentation Modeling (CFM-ID)^{9,15} which provides spectra for the Human Metabolome Database (HMDB) (<https://hmdb.ca>)^{19–23}. mzCloud implements heuristic approaches by combining general fragmentation rules with fragmentation mechanisms published and (partly) relies on manual evaluation²⁴. mzCloud (<https://www.mzcloud.org>) also uses *ab initio* and density functional theory (DFT) quantum-chemical calculations, assigning the lowest-energy isomer found as the fragment ion structure and thus ignoring the kinetic aspects of the CID reaction.

The question that arises is whether the annotated structures of the MS/MS fragment ions (that are often included in *in silico* MS/MS libraries) are indeed correct. Even though these structural annotations are not used directly in MS/MS spectral analysis, the *in silico* MS/MS spectra that are compared with the experiment are derived from the combinatorial algorithm that involves these structures. An estimate of the reliability of these structures is, therefore, an indirect measure for the reliability of predicted MS/MS spectra. Incorrect structures are expected to degrade the quality of the predicted MS/MS spectrum, especially in the lower mass range, where ions occur that are formed upon cleavage of more than one bond. Deviations are, therefore also expected to be more severe at higher-energy CID settings.

Beyond their role in predicting MS/MS spectra, annotated CID-fragment ion structures may serve other purposes as well. In MSⁿ libraries, the MSⁿ spectrum may be used to identify the MS^{n–1} ion structure. As such, the annotated structures increase the number of entries (molecular structures) in the database, including compounds that may not be available as reference standards²⁵. An example of making use of this approach is implemented in METLIN and mzCloud, where a

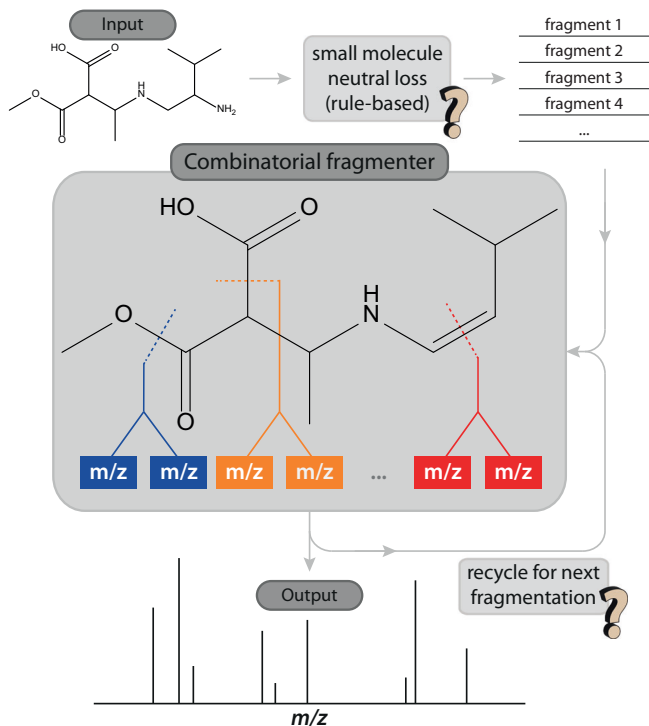


Fig. 1 Simplified schematic showing the combinatorial fragmentation algorithm. Combinatorial fragmentation is used to generate *in silico* MS/MS spectra (figure loosely based on Wolf et al.¹¹). The question marks indicate steps where rearrangements can occur that are not accounted for in the algorithm.

precursor-independent similarity search can be performed that can identify molecular substructures based on annotated fragment structures. In addition, annotations consolidate the internal consistency of the database, as families of structurally similar compounds become connected through shared substructures at deeper stages of the fragmentation tree²⁶.

To address the question at hand, we verify a selection of annotated structures in various MS/MS libraries by infrared ion spectroscopy (IRIS). IRIS records an infrared (IR) spectrum for a mass-isolated CID MSⁿ ion in an ion trap mass spectrometer. Structure determination is then achieved by comparison of the experimental IR spectrum to theoretical IR spectra for candidate structures computed at the density functional level of theory (DFT). In this work, we evaluate the fragment ion structures for a total of 36 MSⁿ ions derived from five precursor molecules, all listed as small-molecule human metabolites. Although this is admittedly a very small sample dataset, it serves as a random selection and ought to give us some insight into the reliability of fragment ion structure annotations in common mass-spectral libraries. We find that the structure annotation of almost all fragment ions investigated is incorrect, which can often be attributed to the neglect of cyclization reactions (METLIN and HMDB) or to an incorrectly selected cyclized product ion (mzCloud). As an indication of the importance of rearrangements in CID MS/MS reactions, a large body of work has been reported on the dissociation reactions of small protonated peptides, employing theoretical and experimental methods in fundamental ion chemistry, including ion spectroscopy²⁷. Furthermore, there are indeed indications that predicted MS/MS spectra are less accurate at higher collision energy²⁸, possibly exposing the effects of rearrangements occurring in between two bond cleavages, which are not accounted for in the *in silico* modeling.

Results and discussion

Workflow. IRIS spectra were recorded for 36 CID-fragment ions derived from five protonated precursors (α -amino-adipic acid, urocanic acid, citrulline, homocitrulline, arginine). Investigated species include MS², MS³ and a few MS⁴ ions. Experimental IRIS spectra were manually compared with predicted IR spectra. In cases where multiple conformers with similar hydrogen bonding interactions are computed to be within 6 kJ mol⁻¹, a Boltzmann-weighted average of the predicted spectra was taken. The degree of spectral matching was determined qualitatively by visual inspection, and the structure providing the best match was assigned as the actual fragment ion structure. Based on this structure, we suggest plausible reaction pathways; transition-state calculations needed to confirm these mechanisms are beyond the scope of the present study. An overview of the workflow is depicted in Fig. 2.

All experimental spectra, comparisons with computed spectra, and the structural identifications proposed are shown in Supplementary Figs. S1–S33. Below, we highlight selected examples and discuss different aspects of structural identification. Figures 3–5, 7, and 9 in this section summarize our spectroscopy-based structure determinations compared against annotations from *in silico* MS/MS libraries. Suggested reaction pathways from precursor to product ions are provided in Supplementary Figs. S36–S61.

α -amino-adipic acid—cyclization rearrangements are common.

As an example of the typical workflow, the results for α -amino-adipic acid (α -AAA) are shown in Fig. 2. Upon CID, the MS² spectrum of protonated α -AAA reveals three prominent fragment ions at m/z 144, 116, and 98. These MS² ions were consecutively mass-isolated, and their IRIS spectra were recorded, see

Supplementary Figs. S2–S4. The IRIS spectrum of the m/z 98 fragment is shown as an example in Fig. 2, along with computed spectra for the annotated structure for this fragment ion taken from the three databases. The elemental composition of the m/z 98 fragment was established as C₅H₈NO⁺ using FTICR-MS (see Supplementary Table S1) and is presumably formed by the neutral loss of CO + 2H₂O. The HMDB^{19–23} and mzCloud²⁶ libraries give the correct chemical formula for this fragment. The annotation in METLIN^{17,18,25} is ambiguous (see Fig. 2): at the aldehyde terminus, either a CC double bond is missing or some, but not all, hydrogens are implicit. To arrive at the correct m/z value, we assume that there is a double bond, forming a ketene moiety; the formal charge is then at the carbon atom adjacent to the amine, so that the METLIN annotation is identical to the one in mzCloud.

Comparisons of experimental and theoretical spectra in Fig. 2 shows that the spectra calculated for the annotations in the libraries deviate substantially from the measured spectrum. Most notably, the strong band observed at 1850 cm⁻¹ is missing for the HMDB structure and shifted for the METLIN/mzCloud structure. We, therefore, devised new structures by chemical intuition and calculated their IR spectra (see Fig. 2 and Supplementary Fig. S4). The N-protonated 6-oxo-1-piperidine ion provides a good match with the experimental spectrum, attributing the 1850 cm⁻¹ band to the carbonyl stretch mode. With this structure for the m/z 98 fragment ion, a reaction mechanism may be proposed involving proton migration upon activation and nucleophilic attack, driving a cyclization and concomitant expulsion of H₂O. Subsequently, the carboxylic acid moiety detaches from the ring and (probably) leaves as H₂O + CO. These processes are not uncommon in gas-phase ion chemistry, although a detailed confirmation of the reaction path requires transition-state calculations.

Figure 3 summarizes all IRIS-based CID-fragment ion annotations for α -AAA, based on the experiment-versus-theory spectral comparisons in Supplementary Figs. S1–S4. One notices that all structures proposed by the databases are linear, as implied by the *in silico* algorithms that were used to derive these structures. However, two out of three spectroscopically established fragment structures are cyclized products.

Urocanic acid—another example of cyclization upon CID.

As another example of facile cyclization upon CID, Fig. 4 shows results for the water loss fragment ion at m/z 121 of protonated urocanic acid, which is the main CID fragment. Our IR spectroscopic analysis shown in Supplementary Fig. S6 unambiguously demonstrates that cyclization occurs to form a bicyclic structure (see Fig. 4). None of the annotations in the MS/MS libraries predict this cyclized structure and instead contain only the original imidazole ring of the precursor. It is interesting to note that MoNA and Wishart Lab sources in the HMDB give different annotations (Splash Key identifiers for the MS/MS spectra used are given in Supplementary Table S6). The MoNA data are based on an experimental MS/MS spectrum to which annotations have been added. The Wishart Lab MS/MS spectrum and its structural annotations are predicted by CFM-ID.

Citrulline—lowest-energy isomer is not necessarily correct structure.

For citrulline, several MS² as well as MS³ fragments were explored, using the MSⁿ capabilities of the ion trap instrument. Figure 5 presents the MSⁿ spectral tree including the annotations from the libraries and from our IR spectroscopic identifications. The underlying experimental IRIS spectra and their matching against DFT computed spectra for hypothesized fragment ion structures is given in Supplementary Figs. S7–S15.

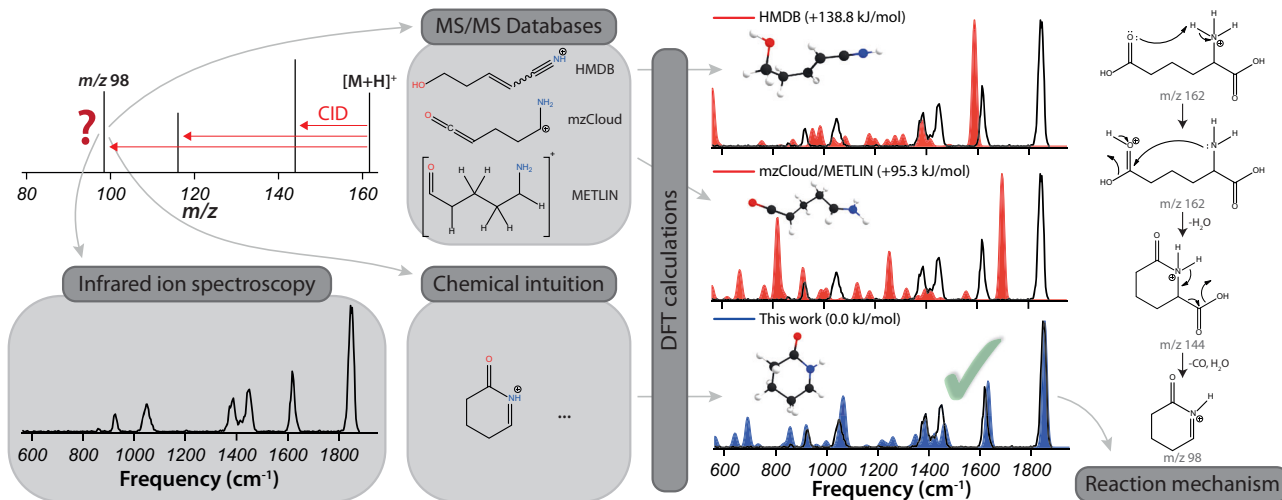


Fig. 2 Workflow to establish the structure of CID MSⁿ fragment ions. The MSⁿ fragment ion of interest is generated by CID and isolated in the ion trap mass spectrometer, where its IRIS spectrum is recorded. DFT geometry optimizations and vibrational frequency calculations are performed for the structures listed in MS/MS libraries as well as for alternative structures suggested by chemical intuition. The structure providing the predicted spectrum with the best match to experiment is annotated as the actual MSⁿ fragment ion structure. Based on this identification, we propose plausible reaction paths from precursor to product ion.

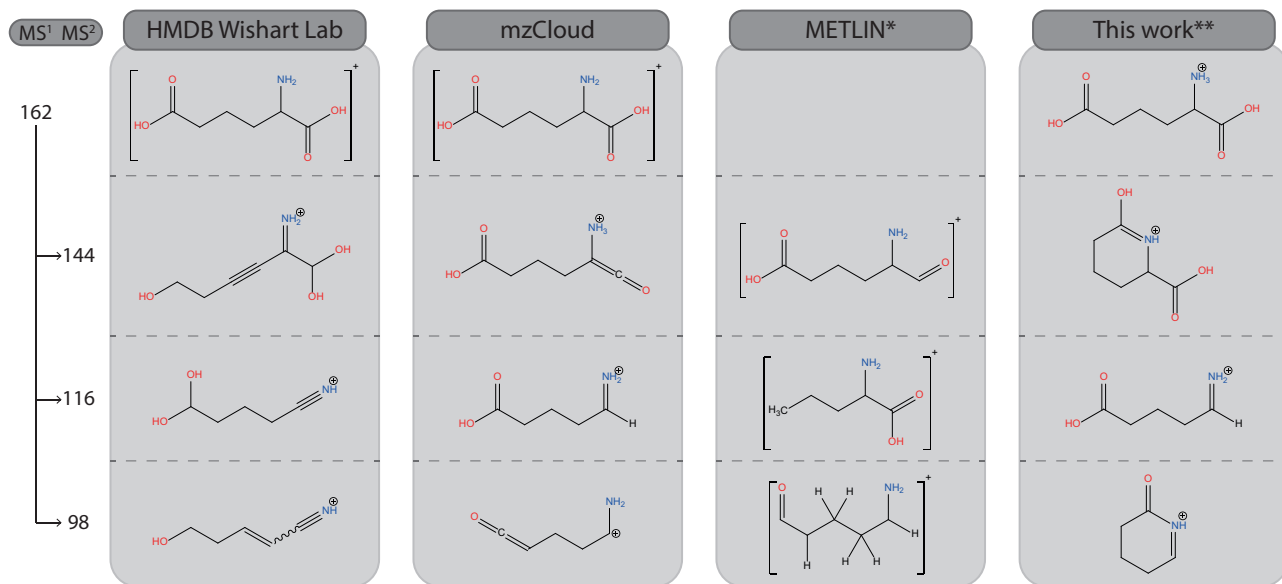


Fig. 3 MS² spectral tree for protonated α -amino-adipic acid with structural annotations. Structures derived in this work from the IR spectra of the fragment ions are compared with those listed in three databases. * For the METLIN structures, a proton must be added to obtain an ion of the correct m/z value. ** Annotations in this column are based on the spectroscopic analysis shown in Supplementary Figs. S1-S4.

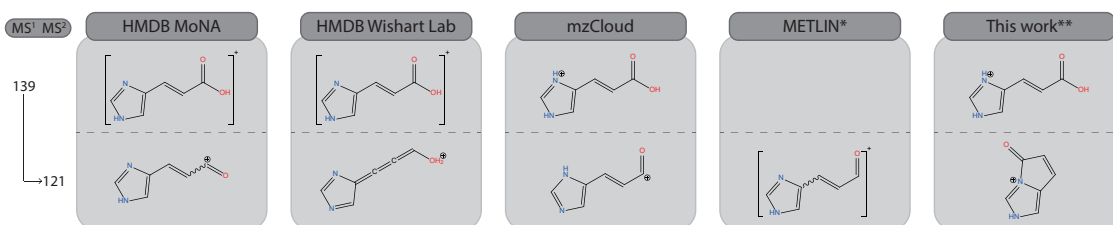


Fig. 4 Spectral tree for protonated urocanic acid with structural annotations. The bicyclic structure identified by ion spectroscopy is not reproduced in any of the databases. * For the METLIN structure, a proton must be added to obtain an ion of the correct m/z value. ** Annotations in this column are based on the spectroscopic analysis shown in Supplementary Figs. S5-S6.

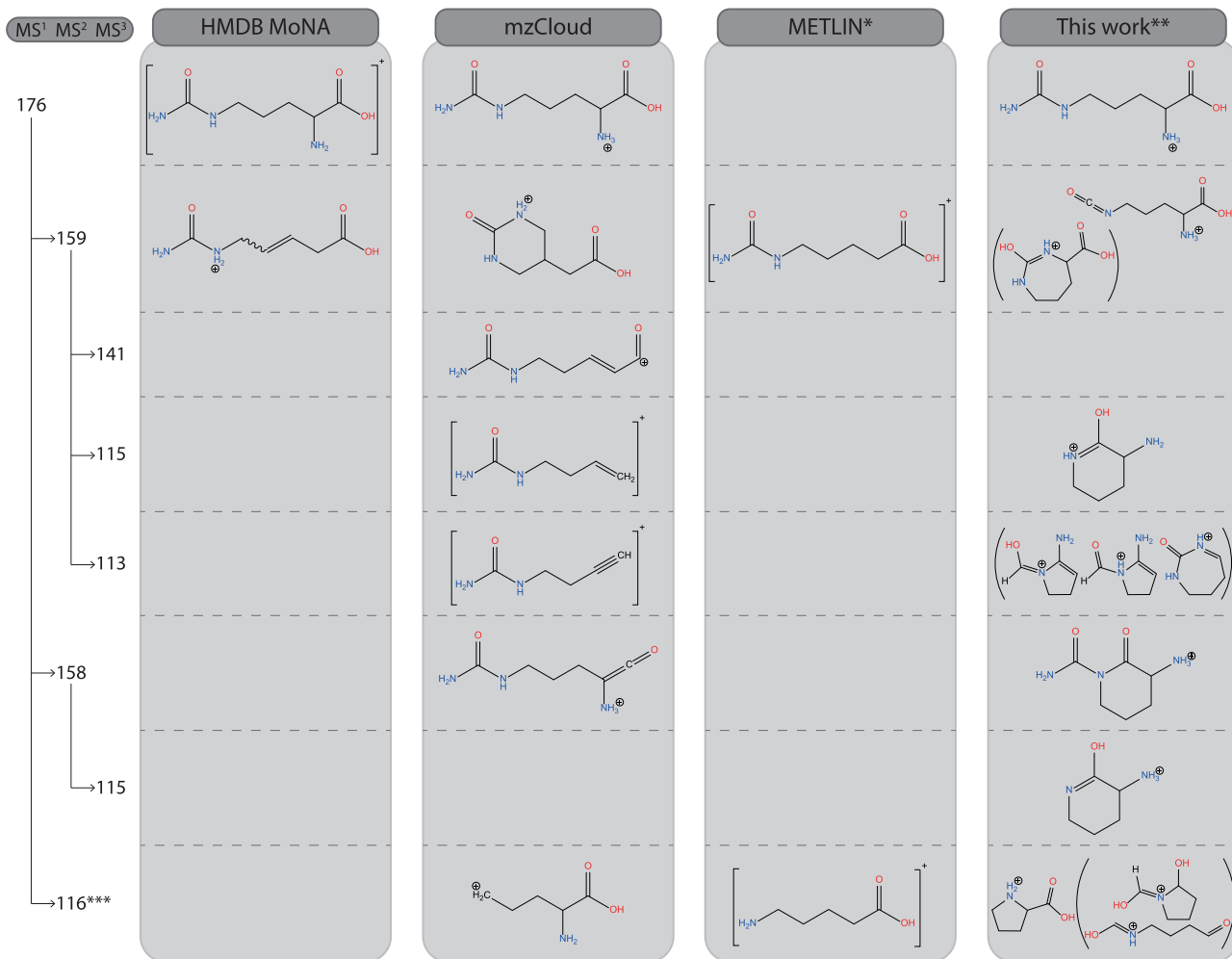


Fig. 5 MSⁿ spectral tree for protonated citrulline with structural annotations. Structures from the databases are contrasted against spectroscopically identified structures. Structures in brackets are tentative. * For METLIN structures, a proton must be placed at one of the carbon atoms to obtain an ion of the correct *m/z* value. ** Annotations in this column are based on the spectroscopic analyses shown in Supplementary Figs. S7–S15. *** A further analysis of the *m/z* 116 product ion from citrulline and from arginine is presented in Supplementary Fig. S34 and Supplementary Note 1.

The HMDB and METLIN libraries only contain MS² spectra, whereas mzCloud also includes fragments of MS³ and higher.

Among the various libraries, mzCloud uniquely employs quantum-chemical calculations to annotate the structure of MSⁿ fragment ions. Calculations are performed for multiple candidate structures and the lowest-energy isomer is proposed as the actual product ion structure. As such, this strategy considers only thermodynamic aspects of the CID reaction and ignores possible kinetic effects. This may lead to incorrect annotations, as is illustrated here for the *m/z* 159 fragment. mzCloud provides PM6 calculations for various candidate structures, giving the 6-membered ring structure shown in Fig. 5 as the global minimum. Figure 6a compares the experimental IRIS spectrum of the *m/z* 159 fragment ion with the B3LYP-calculated spectrum of this structure, revealing a clear mismatch. Most notable is the poor match between the computed carbonyl stretch band at 1840 cm⁻¹ and the strong experimental band at 1770 cm⁻¹. In contrast, the computed spectrum for a linear isomer formed by direct NH₃-loss from the urea moiety, shown in Fig. 6b, provides a reasonable match. The carbonyl stretch as well as the strong feature around 1000–1200 cm⁻¹ and the series of weaker bands between 600 and 1000 cm⁻¹ overlap convincingly. The mismatch near 1500 cm⁻¹ is due to an amine NH bending vibration, which often show small shifts in harmonic

frequency calculations^{29–31}. The OH stretch mode near 3550 cm⁻¹ is closely reproduced, whereas the NH stretch band computed near 3360 cm⁻¹ appears broadened and redshifted, which is likely a consequence of the shared-proton character of the N-H...N moiety in combination with the room-temperature environment of the experiment^{32–34}.

The experimentally observed absorption at 3600 cm⁻¹ is not accounted for in the computed spectrum, nor is the plateau between 1550 and 1700 cm⁻¹. It appears that a fraction of the *m/z* 159 ion population adopts the 7-membered ring structure shown in Fig. 6c, which corresponds to the global minimum at our level of theory, 5 kJ mol⁻¹ lower than the 6-membered ring global minimum structure of mzCloud. From a kinetic point of view, the 7-membered ring appears more plausible since it can form through NH₃-loss from the urea moiety with concomitant nucleophilic attack of the amino nitrogen onto the urea carbon atom (see Fig. 6d). In contrast, formation of the 6-membered ring would require additional displacement of the amino group along the alkyl backbone.

The IRIS spectrum suggests that the global minimum 7-membered ring structure is a minor contributor to the total ion population, with the linear isomer at a relative energy of 22 kJ mol⁻¹ being the dominant contributor. We speculate that pathways leading to rings of larger sizes face increasing entropic barriers³⁵. The PM6 calculations in mzCloud place this linear

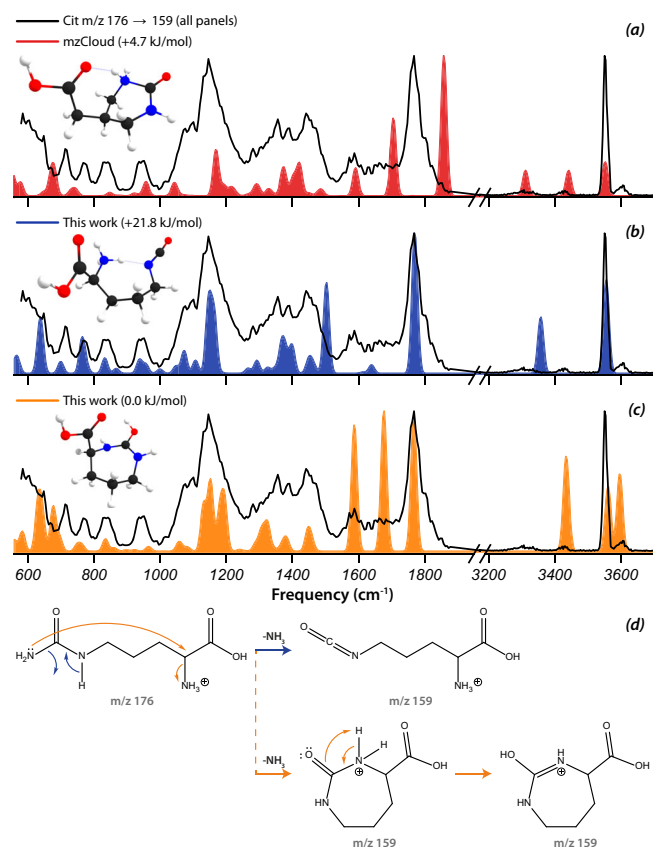


Fig. 6 IRIS spectrum of the m/z 159 MS^2 ion from $[Cit+H]^+$. The measured spectrum is shown as the black trace in **a–c**. The spectrum predicted for the annotation in *mzCloud* deviates substantially from the measured spectrum (**a**). The non-cyclized isomer provides a reasonable match for most bands observed experimentally (**b**). To account for the experimental intensity near 1550–1700 cm^{-1} and at 3600 cm^{-1} , a minor contribution of the 7-membered ring structure is invoked (**c**). Proposed reaction mechanisms leading to these fragment structures (**d**).

isomer 190 $kJ\ mol^{-1}$ above the 6-ring structure. This example shows that the lowest-energy structure is not necessarily the actual fragment structure. This is not uncommon in CID reactions, as for instance, illustrated by the oxazolone structure adopted by b_2 sequence ions of many protonated peptides^{36,37}, which does not correspond to the global minimum, as confirmed in various ion spectroscopy studies^{27,38–42}.

Homocitrulline— MS^n ions of same m/z may have different structures. For homocitrulline, MS^2 as well as MS^3 and MS^4 fragments were investigated, as listed in the MS^n spectral tree in Fig. 7, with annotations from the HMDB library and from our IR spectroscopic identifications. The HMDB libraries only contain MS^2 spectra; homocitrulline is not included in *mzCloud*.

In MS^n spectra with $n \geq 3$, fragments of the same m/z value may be generated via different MS^n paths, i.e., involving different MS^2 intermediates. It appears obvious in such cases that an identical m/z value does not necessarily imply that these fragments correspond to the same structure. However, strategies that assign by definition the lowest-energy isomer give identical annotations to such fragments.

Protonated homocitrulline, $[hCit+H]^+$ at m/z 190, shows an MS^2 fragment at m/z 130 that can also be generated in an MS^3 experiment via the m/z 147 intermediate MS^2 ion. IRIS spectra of the m/z 130 ion produced via either the MS^2 or the MS^3 pathway

are identical, see Fig. 8a. Hence, the two ions possess the same structure, which we identify as protonated pipercolic acid based on a comparison with its predicted IR spectrum. Moreover, a physical reference standard of pipercolic acid is available and its IRIS spectrum coincides closely with those of the MS^2 and MS^3 product ions. The HMDB predicts a linear dienol iminium ion structure for the MS^2 ion (Fig. 7). Pipercolic acid was identified previously as the m/z 130 MS^2 ion of protonated L-lysine⁴³.

The m/z 127 fragment ion of $[hCit+H]^+$ can also be generated via two distinct MS^n routes, either directly in MS^2 or indirectly in MS^3 with the m/z 173 ion as intermediate. In this case, IRIS spectra of the MS^n ions generated through different pathways are clearly distinct, indicating that they do not share the same molecular structure. For the MS^2 fragment ion, the computed spectrum for the 1-piperidine-N-carboxamide ion provides a convincing match, see Fig. 8b. The IRIS spectrum of the m/z 127 MS^3 fragment ion formed via m/z 173 is more difficult to assign. We suspect that two isomeric ions co-exist, as none of the computed spectra reproduces the number of bands observed (see Supplementary Fig. S22). We tentatively propose a mixture of two protomers of 2-amino-2-piperidine-N-aldehyde, protonated either on the amine or on the piperidine nitrogen, shown in Fig. 8c, d.

Arginine—a comparison with the ion chemistry literature. The vast ion chemistry literature documents numerous studies proposing MS^n structure annotations based on manual interpretation of MS^n spectra. For protonated arginine, Zhang et al.⁴⁴ employed high-resolution ESI-MS and HCD- MS^2 on a high-resolution Orbitrap platform to derive product ion structures and reaction mechanisms from MS^3 spectra as well as from MS^2 analyses of deuterated analogs⁴⁵. An overview of fragment ion annotations is given in Fig. 9 along with annotations from libraries and from our spectroscopic investigation. Spectral overlays that form the basis for our annotations are presented in Supplementary Figs. S25–S33.

A quick glance at Fig. 9 suggests that the library annotations are again incorrect. In general, the large majority of fragment ion structures established by IRIS are cyclized species, versus nearly exclusively linear structures in the databases. An exception is the m/z 112 ion in *mzCloud*, which is correctly annotated, although different MS^n pathways lead to this m/z value, and the spectroscopic results show that they do not all correspond to the same molecular structure. The manual annotations of Zhang et al.⁴⁴ include many cyclic product ion structures that generally match closely with our results. Nonetheless, some deviations in assigned structures are also noted. Especially, cyclization upon nucleophilic attack leads mostly to 5- or 6-membered rings according to our spectroscopic analyses, whereas several 8-membered rings are proposed in Zhang et al.⁴⁴.

Conclusions

We conclude that the library annotations rarely agree with the fragment ion structures established spectroscopically in this work. Based on the 36 spectroscopically established CID product ion structures in this work, the fractions of correct and incorrect annotations are displayed in the diagram in Fig. 10, where it should be noted that not all 36 ions are in each database. We consider a database entry to be correct when the molecular structure matches, ignoring discrepancies in protonation site. *mzCloud* appears to slightly outperform HMDB and METLIN, although the fraction of correct annotations is too low to be reliably used.

Admittedly, our test set is extremely small and clearly too limited to derive statistically significant percentages of correct and incorrect annotations. Nonetheless, this set of 36 spectroscopically established CID product ion structures is probably the largest published thus far. Furthermore, 45 additional

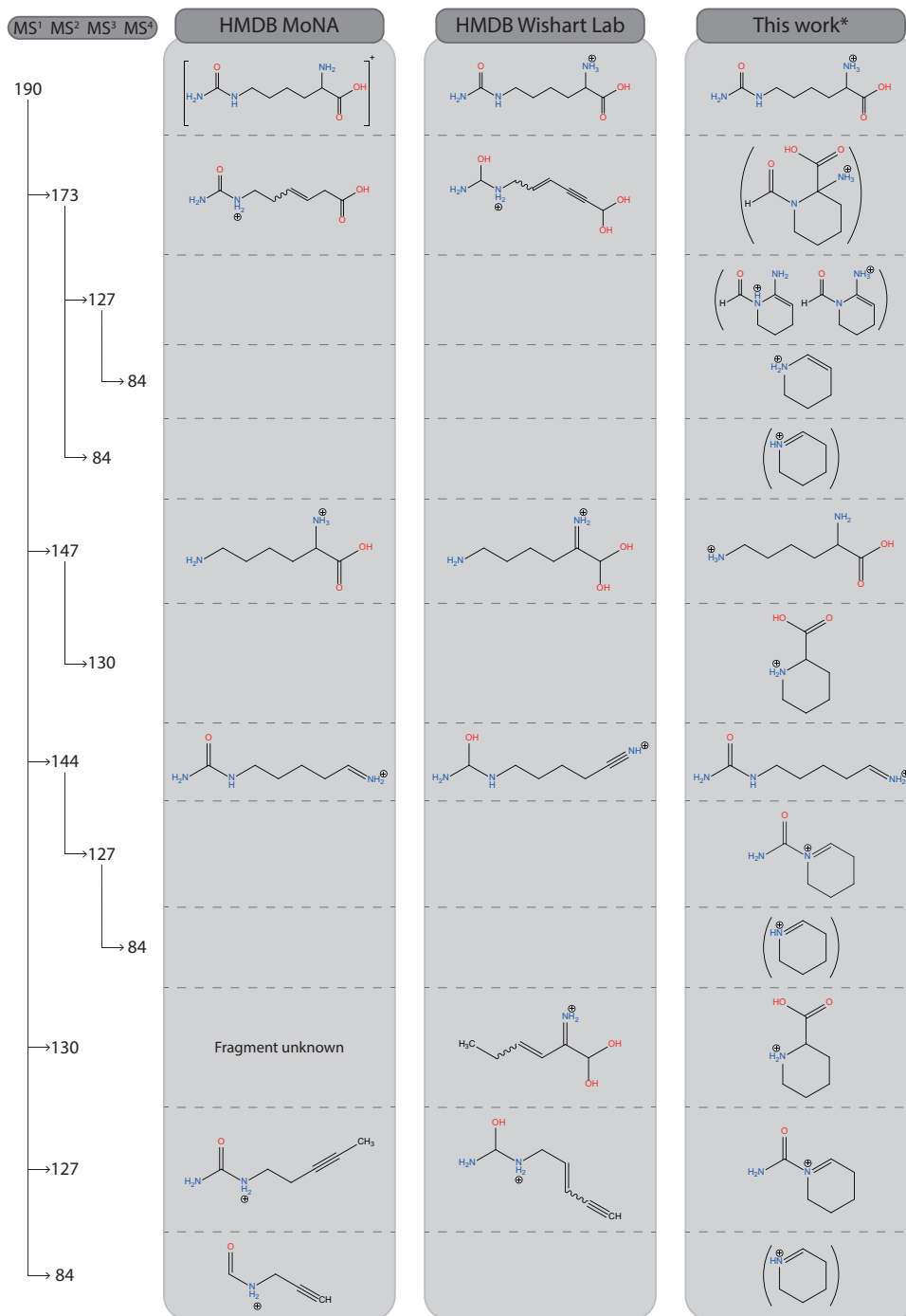


Fig. 7 MSⁿ spectral tree for [homocitrulline + H]⁺ with annotations. HMDB library annotations are compared with structures identified through ion spectroscopy in this work. Annotations in brackets are tentative. * Annotations in this column are based on the spectroscopic analysis shown in Supplementary Figs. S16–S24.

spectroscopically established CID product ion structures were collected from literature; see Section 7 in the Supplementary Information. This dataset contains mostly peptide sequence ions and was also contrasted against structural annotations in the MS/MS libraries, which confirms the trend: only 1 out of 29 HMDB entries is correct and 2 out of 6 mzCloud entries are correct (see Fig. 10). Overall, this test reveals the generic deficiency of library annotations: many of the product ion structures identified here possess cyclic chemical moieties that are formed by rearrangements in the CID reaction. The *in silico* algorithms of MetFrag and CFM-ID do not consider such rearrangements.

Small-molecule neutral losses (e.g., H₂O, NH₃) are common in ESI-MS/MS and are also represented in our small test set. They are of particular interest since they often occur already at low CID activation energies and, hence early on in the MSⁿ spectral tree. In the *in silico* algorithms, these neutral losses are often treated separately and are implemented early on in the fragmentation tree¹¹. We observe in our IR spectroscopic analysis that many H₂O and NH₃ loss events induce cyclization of the ionic fragment. A significant improvement may be achieved by devising new algorithms that correctly annotate the products of these small-molecule neutral losses, which would better model the

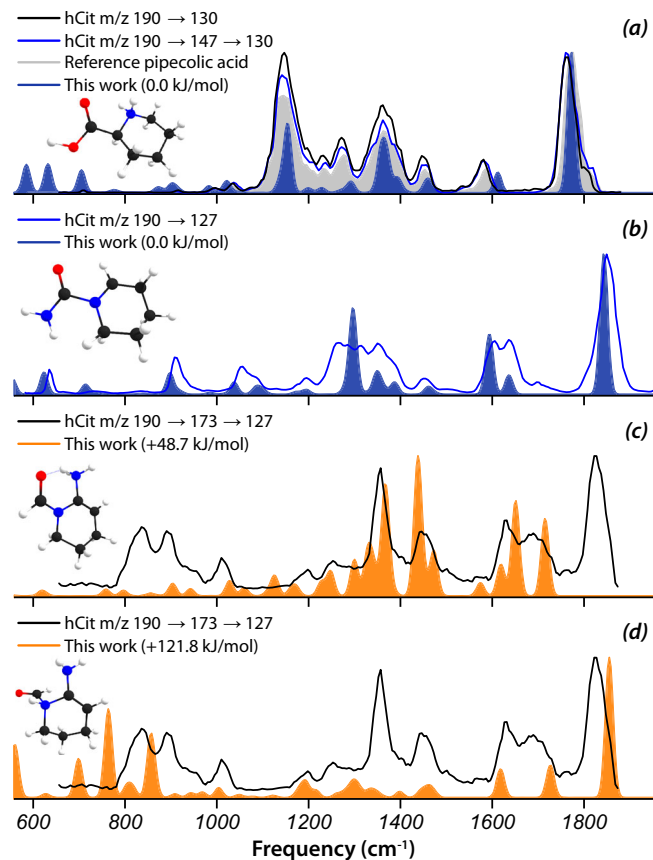


Fig. 8 The m/z 130 fragment ion of homocitrulline generated via different MS^n pathways. **a** IRIS spectra of the m/z 130 CID product ion from $[hCit + H]^+$, generated either directly (MS^2) or via the m/z 147 intermediate (MS^3). Both m/z 130 ions share the same spectrum and hence the same structure, which is identified as protonated pipercolic acid (computed spectrum shaded blue and reference IRIS spectrum shaded gray). **b** IRIS spectrum of the m/z 127 fragments produced directly (MS^2) from the m/z 190 precursor or (**c, d**) via the m/z 173 intermediate (MS^3). Clearly, different species are generated: MS^2 produces the 1-piperidine-N-carboxamide ion (**b**), while MS^3 is suggested to produce two protonation isomers of 2-amino-2-piperidine-N-aldehyde (**c, d**).

entire downstream fragmentation tree and, hence, the predicted MS/MS fragment masses.

The current machine-learning revolution is having a strong impact on the development of *in silico* MS/MS algorithms. The performance of fragmentation tools currently available is benchmarked in numerous reports in the MS literature, see, for instance, the CASMI contest^{46,47}. The performance is usually evaluated based on the ability to correctly identify molecular (sub)structures from an experimental MS/MS spectrum, without other a priori knowledge of the query compound. In the most recent competition, although the elemental composition was often correctly identified from the mass-spectral data, the correct 2D chemical structure was retrieved from the MS/MS spectrum in only about 10% of cases⁴⁸. A benchmark on the ability to correctly annotate MS/MS fragment ions may serve as an alternative test that could provide deeper insight into the underlying fundamentals of why one algorithm performs better than another one.

Methods

Chemicals and sample preparation. Homocitrulline, citrulline, arginine, and α -amino-adipic acid were obtained from MetaSci (Toronto, Canada). Urocanic acid and reference compounds were

acquired from Sigma-Aldrich (St. Louis, USA). Stock solutions were prepared by dissolving the samples in 50:50 MeOH-H₂O (LC-MS grade, obtained from Sigma-Aldrich (St. Louis, USA)). Approximately 0.5% of formic acid (LC-MS grade, obtained from Sigma-Aldrich (St. Louis, USA)) was added to promote protonation.

Experimental. A modified 3D quadrupole ion trap mass spectrometer (Bruker, Amazon Speed ETD) is used to perform IRIS experiments. The instrument was modified to enable optical access to the trapped ions⁴⁹. Solutions were diluted to approximately 1 μ M and introduced into the MS by direct infusion at 120–180 μ l/h flow rates through an electrospray ionization (ESI) source. Typical operating parameters were capillary and end plate voltages of 4500 V and 500 V, respectively, and a nitrogen dry gas temperature of 180 °C flowing at 4 L min⁻¹. MS/MS and MS^n fragments were generated by collision-induced dissociation (CID) of the mass-isolated protonated precursor ion. The CID amplitude was optimized to maximize the signal on the fragment mass peak of interest. Additionally, a Fourier transform ion cyclotron resonance (FTICR-MS, Bruker, Solarix XR 7 T) mass spectrometer equipped with an identical ESI source was used to determine the accurate mass and hence the molecular formula of all MS^n fragments (see Supplementary Figs. S62–S66 and Supplementary Tables S1–S5).

In the ion trap, the MS^n fragment of interest is mass-isolated and subjected to IR analysis using the FELIX free-electron laser. The FELIX IR laser was set to produce IR radiation between 560 and 1950 cm^{-1} in the form of 5–10 μ s long macropulses of 20–160 mJ at a repetition rate of 5 or 10 Hz and with a bandwidth of ~0.4% of the center frequency. An optical parametric oscillator (LaserSpec, Belgium)⁵⁰ was also used to obtain IR spectra in the 3250–3800 cm^{-1} spectral range for some MS/MS fragments of arginine and citrulline. The OPO produces 5.6 nJ pulses of 35 ps duration at a 80-MHz repetition rate and with a bandwidth of 0.5 cm^{-1} .

When the laser frequency is in resonance with a vibrational transition, multiple-photon absorption by the ions results in an increase of their internal energy and, eventually, in their fragmentation. The extent of fragmentation in the mass spectrum can be monitored as a function of the IR laser frequency. Hence, an IR vibrational spectrum can be reconstructed from a series of mass spectra by monitoring the IR-induced dissociation yield, defined here as $\ln\left[\frac{\sum I_{all\ ions}}{I_{precursor\ ion}}\right]$, as function of IR frequency⁵¹. The yield is obtained from 6 averaged mass spectra at each laser frequency, advancing in steps of 5 cm^{-1} . The frequency of the laser was calibrated using a grating spectrometer and the yield was linearly corrected for variations in the laser pulse energy. For displaying purposes, each experimental spectrum is normalized to the most intense peak.

Computational. Theoretical spectra were generated for possible fragment ion structures using density functional theory (DFT) at the B3LYP/6-31++G(d,p) level. As input structures, we used annotations suggested in the HMDB, mzCloud, and METLIN databases, as well as many alternative structures based on chemical intuition. In our computational workflow, the SMILES 2D-structure format was used as input for the cheminformatics toolbox RDKit⁵² that generates all possible protonation isomers as well as stereoisomers. Structures containing only one stereocenter were randomly assigned to be R or S, as enantiomers have the same IR spectrum. For each protonation isomer, a conformational search was performed using a distance geometry algorithm to find 500 random 3D-conformations, after which the structures were optimized with a classical force field⁵³ and

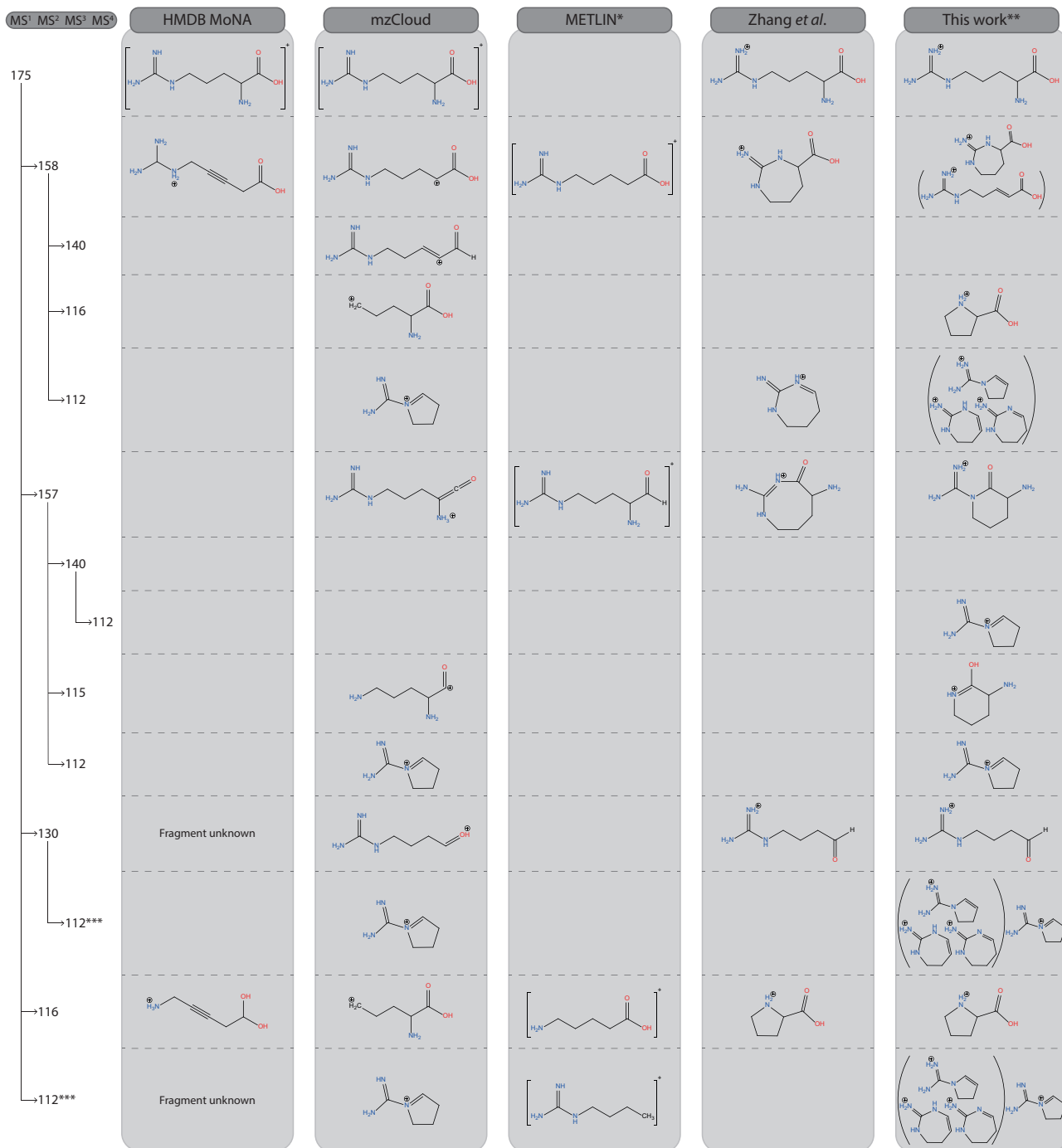


Fig. 9 MS^n spectral tree of protonated arginine with annotations. Annotations proposed by the MS libraries, Zhang et al.⁴⁴, and ion spectroscopy (this work) are compared. Structures in brackets are tentative assignments. * For METLIN structures, a proton must be placed at one of the carbon atoms to arrive at an ion of the correct m/z value. ** Annotations based on the spectroscopic analysis shown in Supplementary Figs. S25–S33. *** A more detailed analysis of the m/z 112 fragment ion is presented in Supplementary Fig. S35 and Supplementary Note 2.

clustered based on similarity. A maximum of 20 unique structures were then selected based on root-mean-square deviations of atomic positions, and these were submitted to Gaussian16⁵⁴ for geometry optimizations and frequency calculations at the semi-empirical PM6 level. Optimizations that converge to one of the other isomers, unconverged calculations and structures with broken bonds are automatically removed. In addition, conformations were filtered by a relative energy cut-off of 40 kJ mol⁻¹. The remaining structures were then re-optimized at

the B3LYP/6-31++G(d,p) level of theory^{55,56} and harmonic vibrational frequencies were calculated and scaled by a factor of 0.975 to correct for anharmonicity⁵⁷. A factor of 0.955 was applied for the H-stretching frequencies in the 3- μ m wavelength range⁵⁷. A convolution of the stick spectra with a Gaussian line shape of 20 cm⁻¹ full width at half maximum produced the theoretical spectra that were compared with the experimental spectra. The B3LYP-optimized geometries were also employed for a single-point electronic energy calculation at the MP2/6-31+

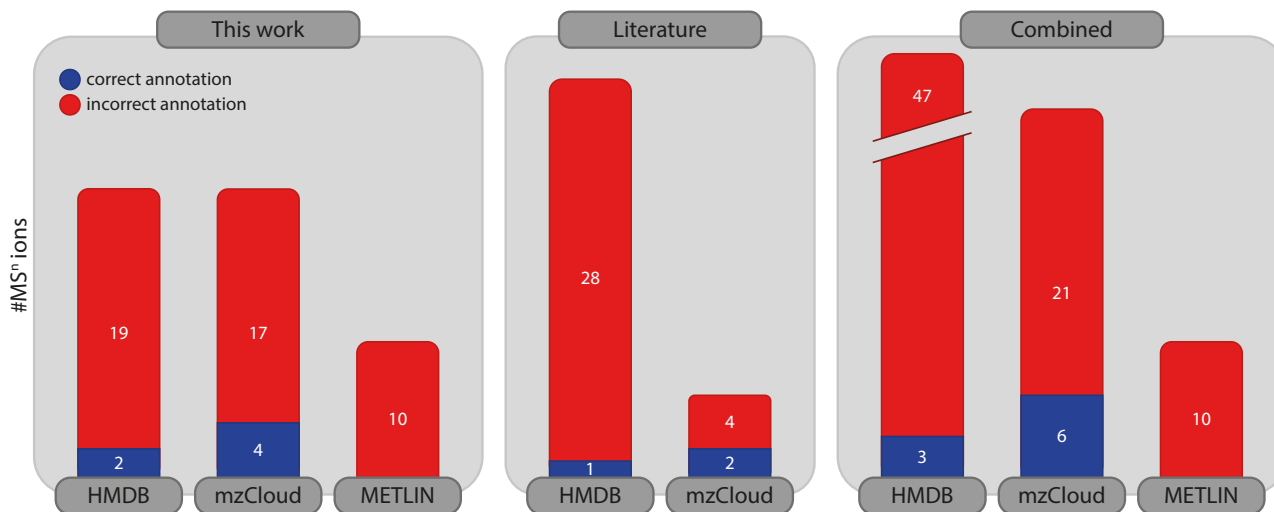


Fig. 10 Number of correct and incorrect MS/MS structure annotations. Summary of correct/incorrect MSⁿ structure annotations as verified by IR ion spectroscopy in this work; the large majority of annotations in major MS/MS libraries are likely incorrect. Additional spectroscopically established CID product ion structures collected from literature confirm this trend (see Supplementary Figs. S67–S89).

+G(d,p) level of theory to improve relative energies of isomeric structures.

Data availability

In the Supplementary Information file, we provide comparisons of experimental versus theoretical spectra underlying our proposed structural annotations in Figs. 3–5, 7, and 9. Furthermore, we provide all DFT-optimized geometries of fragment ion structures considered. Plausible reaction pathways connecting precursor ion and CID-fragment ion structure established through IRIS are suggested. Recorded high-resolution MS/MS spectra and accurate mass values of CID fragments studied are listed. We provide a list of Splash Keys identifying the accessed MS/MS spectra in the HMDB. Finally, additional spectroscopically identified CID-fragment ion structures from various precursors taken from the literature (mainly small peptides) are presented. The experimental IRIS and MS/MS data underlying the figures in the main text and in the Supplementary Information is available through the Radboud Data Repository (<https://doi.org/10.34973/e77p-ac39>).

Received: 18 November 2023; Accepted: 22 January 2024;

Published online: 14 February 2024

References

- Dettmer, K., Aronov, P. A. & Hammock, B. D. Mass spectrometry-based metabolomics. *Mass Spectrom. Rev.* **26**, 51–78 (2007).
- Vinaixa, M. et al. Mass spectral databases for Lc/Ms- and Gc/Ms-based metabolomics: state of the field and future prospects. *TrAC Trends Anal. Chem.* **78**, 23–35 (2016).
- Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 0054. <https://doi.org/10.1038/s41570-017-0054> (2017).
- Bittremieux, W., Wang, M. & Dorrestein, P. C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. *Metabolomics* **18**, 94 (2022).
- Dobson, C. M. Chemical space and biology. *Nature* **432**, 824–828 (2004).
- Collins, S. L., Koo, I., Peters, J. M., Smith, P. B. & Patterson, A. D. Current challenges and recent developments in mass spectrometry-based metabolomics. *Annu. Rev. Anal. Chem.* **14**, 467–487 (2021).
- Hufsky, F., Scheubert, K. & Böcker, S. Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends Anal. Chem.* **53**, 41–48 (2014).
- Martin Somer, A. et al. Role of chemical dynamics simulations in mass spectrometry studies of collision-induced dissociation and collisions of biological ions with organic surfaces. *J. Am. Soc. Mass Spectrom.* **31**, 2–24 (2020).
- Djombou-Feunang, Y. et al. CFM-ID 3.0: significantly improved ESI-MS/MS prediction and compound identification. *Metabolites* **9**, 72 (2019).
- Tsugawa, H. et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-finder software. *Anal. Chem.* **88**, 7946–7958 (2016).
- Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform.* **11**, 148 (2010).
- Duhrkop, K., Shen, H. B., Meusel, M., Rousu, J. & Bocker, S. Searching molecular structure databases with tandem mass spectra using CSI:fingerid. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
- Cao, L. et al. Moldiscovery: learning mass spectrometry fragmentation of small molecules. *Nat. Commun.* **12**, 3718 (2021).
- Hoffmann, M. A. et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat. Biotechnol.* **40**, 411–421 (2022).
- Wang, F. et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* **93**, 11692–11700 (2021).
- Singh, A. Annotating unknown metabolites. *Nat. Methods* **20**, 33 (2023).
- Xue, J., Guijas, C., Benton, H. P., Warth, B. & Siuzdak, G. Metlin MS² molecular standards database: a broad chemical and biological resource. *Nat. Methods* **17**, 953–954 (2020).
- Smith, C. A. et al. Metlin: a metabolite mass spectral database. *Ther. Drug Monit.* **27**, 747–751 (2005).
- Wishart, D. S. et al. HMDB: the human metabolome database. *Nucleic Acids Res.* **35**, D521–D526 (2007).
- Wishart, D. S. et al. Hmdb: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **37**, D603–D610 (2009).
- Wishart, D. S. et al. HMDB 3.0—the human metabolome database in 2013. *Nucleic Acids Res.* **41**, D801–D807 (2013).
- Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
- Wishart, D. S. et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
- Yi, Z. & Zhu, Z. J. Overview of tandem mass spectral and metabolite databases for metabolite identification in metabolomics. *Methods Mol. Biol.* **2104**, 139–148 (2020).
- Guijas, C. et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
- Sheldon, M. T., Mistrík, R. & Croley, T. R. Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *J. Am. Soc. Mass Spectrom.* **20**, 370–376 (2009).
- Patrick, A. L. & Polfer, N. C. Peptide fragmentation products in mass spectrometry probed by infrared spectroscopy. In: *Gas-Phase IR Spectroscopy and Structure of Biological Molecules*, Rijs, A. M.; Oomens, J. Eds. Springer International Publishing: Cham; pp 153–181. https://doi.org/10.1007/128_2014_576 (2015)
- Demarque, D. P., Crotti, A. E., Vescechi, R., Lopes, J. L. & Lopes, N. P. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Nat. Prod. Rep.* **33**, 432–455 (2016).
- van Outersterp, R. E. et al. Metabolite identification using infrared ion spectroscopy—novel biomarkers for pyridoxine-dependent epilepsy. *Anal. Chem.* **93**, 15340–15348 (2021).

30. Hofstetter, T. E., Howder, C., Berden, G., Oomens, J. & Armentrout, P. B. Structural elucidation of biological and toxicological complexes: investigation of monomeric and dimeric complexes of histidine with multiply charged transition metal (Zn and Cd) cations using IR action spectroscopy. *J. Phys. Chem. B* **115**, 12648–12661 (2011).
31. Oomens, J., Moore, D. T., Meijer, G. & von Helden, G. Infrared multiple photon dynamics and spectroscopy of cationic PABA and its dehydroxylated fragment ion. *Phys. Chem. Chem. Phys.* **6**, 710–718 (2004).
32. Li, X., Moore, D. T. & Iyengar, S. S. Insights from first principles molecular dynamics studies toward infrared multiple-photon and single-photon action spectroscopy: case study of the proton-bound dimethyl ether dimer. *J. Chem. Phys.* **128**, 184308 (2008).
33. Roscioli, J. R., McCunn, L. R. & Johnson, M. A. Quantum structure of the intermolecular proton bond. *Science* **316**, 249–254 (2007).
34. Martínez-Haya, B. et al. A dynamic proton bond: $\text{MH}^+\text{-H}_2\text{O} \rightleftharpoons \text{M-H}_3\text{O}^+$ interconversion in loosely coordinated environments. *J. Phys. Chem. Lett.* **14**, 1294–1300 (2023).
35. Tirado, M. & Polfer, N. C. Defying entropy: forming large head-to-tail macrocycles in the gas phase. *Angew. Chem. Int. Ed.* **51**, 6436–6438 (2012).
36. Yalcin, T., Khouw, C., Csizmadia, I. G., Peterson, M. R. & Harrison, A. G. Why are B ions stable species in peptide spectra? *J. Am. Soc. Mass Spectrom.* **6**, 1164–1174 (1995).
37. Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24**, 508–548 (2005).
38. Gucinski, A. C., Chamot-Rooke, J., Nicol, E., Somogyi, A. & Wysocki, V. H. Structural influences on preferential oxazolone versus diketopiperazine b_2^+ ion formation for histidine analogue-containing peptides. *J. Phys. Chem. A* **116**, 4296–4304 (2012).
39. Oomens, J. et al. Water loss from protonated XxxSer and XxxThr dipeptides gives oxazolone–not oxazolone–product ions. *J. Am. Soc. Mass Spectrom.* **31**, 2111–2123 (2020).
40. Oomens, J., Young, S., Molesworth, S. & van Stipdonk, M. Spectroscopic evidence for an oxazolone structure of the b_2 fragment ion from protonated Tri-alanine. *J. Am. Soc. Mass Spectrom.* **20**, 334–339 (2009).
41. Perkins, B. R. et al. Evidence of diketopiperazine and oxazolone structures for HA b_2^+ Ion. *J. Am. Chem. Soc.* **131**, 17528–17529 (2009).
42. Zou, S., Oomens, J. & Polfer, N. C. Competition between diketopiperazine and oxazolone formation in water loss products from protonated ArgGly and GlyArg. *Int. J. Mass Spectrom.* **316–318**, 12–17 (2012).
43. van Outersterp, R. E. et al. Reference-standard free metabolite identification using infrared ion spectroscopy. *Int. J. Mass Spectrom.* **443**, 77–85 (2019).
44. Zhang, P. et al. Revisiting fragmentation reactions of protonated alpha-amino acids by high-resolution electrospray ionization tandem mass spectrometry with collision-induced dissociation. *Sci. Rep.* **9**, 6453 (2019).
45. Chen, J., Shiyarov, P., Schlager, J. J. & Green, K. B. A pseudo MS^3 approach for identification of disulfide-bonded proteins: uncommon product ions and database search. *J. Am. Soc. Mass Spectrom.* **23**, 225–243 (2012).
46. Schymanski, E. L. et al. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.* **9**, 22 (2017).
47. Blazenovic, I. et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. *J. Cheminform.* **9**, 32 (2017).
48. Fiehn-Lab CASMI results. <https://fiehnlab.ucdavis.edu/casmi/casmi-2022-results> (2022).
49. Martens, J., Berden, G., Gebhardt, C. R. & Oomens, J. Infrared ion spectroscopy in a modified quadrupole ion trap mass spectrometer at the FELIX free electron laser laboratory. *Rev. Sci. Instrum.* **87**, 103108 (2016).
50. van Outersterp, R. E. et al. Evaluation of table-top lasers for routine infrared ion spectroscopy in the analytical laboratory. *Analyst* **146**, 7218–7229 (2021).
51. Martens, J. et al. Infrared ion spectroscopy: new opportunities for small-molecule identification in mass spectrometry—a tutorial perspective. *Anal. Chim. Acta* **1093**, 1–15 (2020).
52. Landrum, G. RDKit: open-source cheminformatics software. (2016).
53. Tosco, P., Stiefl, N. & Landrum, G. Bringing the MMFF force field to the RDKit: implementation and validation. *J. Cheminform.* **6**, 37 (2014).
54. Frisch, M. J. et al. Gaussian 16 Rev. C.01, Wallingford, CT, (2016).
55. He, C. C. et al. Evaluation of hybrid theoretical approaches for structural determination of a glycine-linked cisplatin derivative via infrared multiple photon dissociation (IRMPD) action spectroscopy. *J. Phys. Chem. A* **119**, 10980–10987 (2015).
56. Rodrigues-Oliveira, A. F., M. Ribeiro, F. W., Cervi, G. & C. Correira, T. Evaluation of common theoretical methods for predicting infrared multiphoton dissociation vibrational spectra of intramolecular hydrogen-bonded ions. *ACS Omega* **3**, 9075–9085 (2018).
57. Martens, J. K., Grzetic, J., Berden, G. & Oomens, J. Gas-phase conformations of small polyprolines and their fragment ions by IRMPD spectroscopy. *Int. J. Mass Spectrom.* **377**, 179–187 (2015).

Acknowledgements

We gratefully acknowledge the support of the FELIX technical staff. We thank Ms. Anneke Grotenhuis and Ms. Anna Peters for the preliminary experiments leading to this study. This project received funding from the Dutch Research Council (NWO) under grant numbers TTW-OTP 15769, TKI-LIFT 731.014.419, and GWI Roadmap 184.034.022. Computations were performed at the national supercomputer *Snellius* at SurfSara in Amsterdam with the compute budget kindly provided through NWO Rekenijd grant 2021.055.

Author contributions

LvT, SS, and QDKW recorded and analyzed all MS/MS and IRIS spectra, and performed quantum-chemical calculations. KJH developed, streamlined, and supervised the computational workflow. RvE, JM, and GB supervised all experimental work. RAW and DSW advised on the use of MS^n libraries. JO, GB, and JM conceived the study. JO, GB, and RAW acquired funding. LvT and JO wrote the manuscript, which was edited by all authors.

Competing interests

DS Wishart is the initiator of the Human Metabolome Database (HMDB). All other authors declare no competing interests.

Additional information


Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42004-024-01112-7>.

Correspondence and requests for materials should be addressed to Jos Oomens.

Peer review information *Communications Chemistry* thanks Pieter Dorrestein, Kevin Pagel and the other, anonymous reviewer for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024