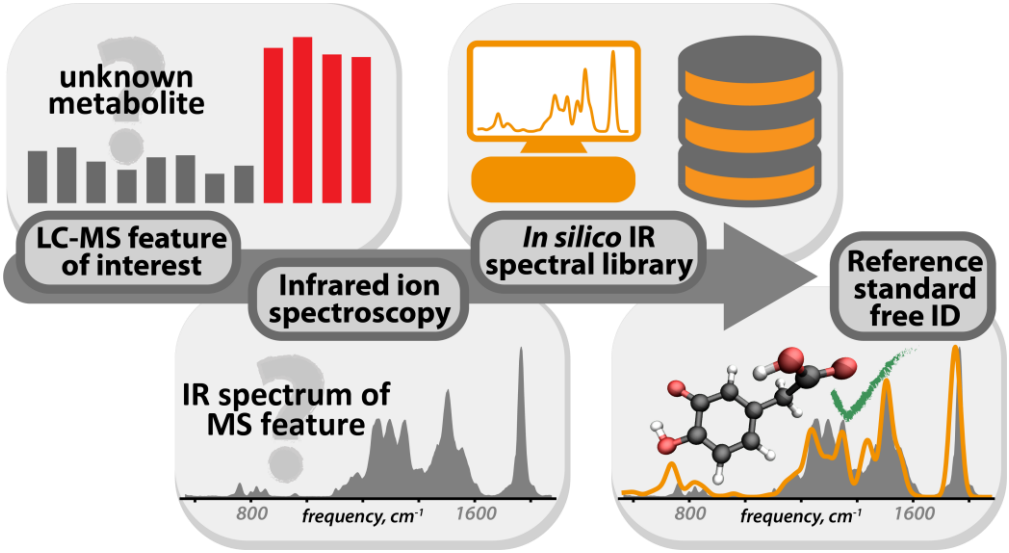
An *in silico* infrared spectral library of molecular ions for metabolite identification

Kas J. Houthuijs,[a] Giel Berden,[a] Udo F.H. Engelke,[b] Vasuk Gautam,[c] David S. Wishart,[c,d,e,f] Ron A. Wevers,[b] Jonathan Martens,[a] Jos Oomens[a,g,*]

[a] Institute for Molecules and Materials, FELIX Laboratory, Radboud University, Nijmegen, 6525 ED, The Netherlands
[b] Department of Genetics, Translational Metabolic Laboratory, Radboud University Medical Center, Nijmegen, 6525 GA, The Netherlands
[c] Department of Biological Sciences, University of Alberta, Edmonton, AB T6G 2E9, Canada
[d] Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8, Canada
[e] Department of Laboratory Medicine and Pathology, University of Alberta, Edmonton, AB T6G 2B7, Canada
[f] Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, AB T6G 2H7, Canada
[g] van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Amsterdam, 1098 XH, The Netherlands

ABSTRACT: Infrared ion spectroscopy (IRIS) continues to see increasing use as an analytical tool for small-molecule identification in conjunction with mass spectrometry (MS). The IR spectrum of an *m/z* selected population of ions constitutes a unique fingerprint that is specific to the molecular structure. However, direct translation of an IR spectrum to a molecular structure remains challenging, as reference libraries of IR spectra of molecular ions largely do not exist. Quantum-chemically computed spectra can reliably be used as reference, but the challenge of selecting the candidate structures remains. Here we introduce an *in silico* library of vibrational spectra of common MS adducts of over 4500 compounds found in the human metabolome database (HMDB). In total, the library currently contains more than 75 000 spectra computed at the DFT level that can be queried with an experimental IR spectrum. Moreover, we introduce a database of 189 experimental IRIS spectra, which is employed to validate the automated spectral matching routines. This demonstrates that 75% of metabolites in the experimental dataset is correctly identified, based solely on their exact *m/z* and IRIS spectrum. Additionally, we demonstrate an approach for specifically identifying substructures by performing a search without *m/z* constraints to find structural analogues. Such an unsupervised search paves the way towards the *de novo* identification of unknowns that are absent in spectral libraries. We apply the *in silico* spectral library to identify an unknown in a plasma sample as 3-hydroyxhexanoic acid, highlighting the potential of the method.



Introduction

Mass spectrometry (MS) is the primary analytical approach in untargeted metabolomics due to its high sensitivity, selectivity and throughput. Routine analyses of complex biological samples yield large numbers of detected features – numbers that are unrivalled by other analytical techniques [1]. Subsequent (biological) interpretation requires the precise molecular identification of the *m/z* features of interest [2]. Herein lies the major challenge of MS-based metabolomics, as little structural information is contained in an *m/z* value alone. To distinguish between structural isomers, MS is often hyphenated with gas or liquid chromatography (GC or LC), tandem MS (MS/MS) [3] or ion mobility spectrometry (IMS) [4]. Annotation is then achieved by comparison against libraries with reference MS/MS spectra [5, 6], retention times or CCS values [6, 7]. Yet, despite the many reference libraries, detailed molecular identification remains the main bottleneck in MS-based metabolomics.

Infrared ion spectroscopy (IRIS) provides an alternative approach to molecular structure determination in MS. IRIS measures the IR spectrum of an *m/z*-selected population of ions in the MS instrument and thus provides detailed structural information [8]. Being rooted in MS, IRIS can be appended to LC/MS-workflows [9, 10], which has enabled its analytical implementation in, for instance, the identification of biomarkers for metabolic diseases [11-13] and other small-molecule isomerism questions [14-16]. Translation of an IR spectrum to a chemical structure is usually performed by formulating (a small number of) candidate structures based on prior knowledge of the underlying (bio)chemistry. Structural confirmation of these candidates is then achieved through comparison of the experimental IRIS spectrum to a reference spectrum from a physical standard and/or from a quantum-chemical prediction [17]. The latter option enables identification without a physical reference standard and is therefore efficient if standards are not available. Density functional theory (DFT) calculations typically give accurate predictions of a compound's vibrational spectrum and have been employed in numerous fundamental ion chemistry studies utilizing IRIS [18-20]. DFT-predictions are therefore useful for (preliminary) metabolite identification and can at least reduce the number of candidate structures substantially.

This approach to spectrum-to-structure conversion is efficient if detailed prior knowledge on the detected feature is available, especially if the candidate structures are well defined and limited in number. However, such information is often not available or not sufficiently detailed. Ideally, one would use a large IRIS spectral library, similar to those available for MS/MS spectra, to facilitate quick and high-throughput screening. Large IR spectral libraries are available for neutral (gaseous) molecules, but their ions formed in MS (*e.g.* $[M+H]^+$, $[M-H]^-$, $[M+Na]^+$) possess drastically different IR spectra [21]. An experimental IRIS library sufficiently large for identification purposes is currently unavailable (although efforts are underway to compile small libraries for specific compound classes, such as glycans [22, 23]). Moreover, in the light of the vastness of small-molecule chemical space [24, 25], experimental reference libraries are intrinsically limited for metabolomic studies, as is also the case for MS/MS and IMS data. Therefore, the focus in the metabolomics community has shifted towards expanding MS/MS and IMS libraries with computational data [5, 26-28]. In addition to their cost-effectiveness, *in silico* libraries have the advantage of a greatly increased coverage, since only a small fraction of molecules are available as physical reference standards.

Using quantum-chemical methods, vibrational spectra can be routinely and accurately calculated (somewhat in contrast to MS/MS spectra). For example, Karunaratne et al. pursued identification of neutral gaseous molecules detected with GC-FTIR spectroscopy by employing DFT-calculated reference spectra of compounds extracted from the PubChem database [29]. This approach yields identification rates comparable to those achieved with MS/MS. However, GC-FTIR is not compatible with typical sample matrices in the metabolomics field, where LC-MS is the method of choice.

Here we present an *in silico* IR spectral library of molecular ions, compatible with IRIS and hence LC-MS workflows. This library was compiled by coupling a standardized workflow for calculating vibrational spectra at the DFT-level to the human metabolome database (HMDB) [6]. This workflow requires only a simple chemical identifier as input to automatically generate relevant IR spectra. This workflow is applied to entries in the HMDB with molecular weights <210 Da. We present benchmark tests of this *in silico* library using a new, extensive set of 189 experimental IRIS spectra, derived from 87 unique metabolites. We demonstrate that the specificity of the IR fingerprint not only allows for the identification of known unknowns, but also enables *de novo* structural elucidation of unknown unknowns.

Methods

Chemicals

Methanol and water (HPLC grade) were obtained from Sigma-Aldrich (St. Louis, USA). The metabolite reference compounds originate from vendors indicated in Table S2 in the SI. The 3-hydroxyhexanoic acid reference metabolite was obtained from Enamine Ltd. (Kiev, Ukraine).

Infrared ion spectroscopy

IR fingerprint spectra (600-1900 cm^{-1}) of the reference compounds were measured using an ion trap mass spectrometer (Bruker AmaZon Speed ETD, Bremen, Germany) coupled to the beamline of the Free-Electron Laser for Infrared eXperiments, FELIX [30, 31]. Solutions of the reference compounds (approx. 10^{-5} M in 1:1 methanol:water) were directly infused into the electrospray ionization source, after which adducts of interest ($[M+H]^+$, $[M-H]^-$, and/or $[M+Na]^+$) were mass-isolated and subjected to IR measurements. LC-IRIS measurements were performed in an online fashion using a single 80 µL sample loop installed between a 6-port switching valve and the ion source [10, 32]. The isolated ions were irradiated with 1 to 10 FELIX macropulses of 10 µs; the pulse energy ranges from 20 to 180 mJ depending on wavelength and the laser bandwidth amounts to 0.5% of the center frequency. The internal energy of the ions increases by frequency-dependent absorption of multiple IR photons until the ions undergo fragmentation. Each IR spectrum was reconstructed from a series of mass spectra by plotting the fragmentation yield, $-\ln[\, I_{precursor} / \sum(I_{all})]$, as a function of IR frequency (scanned in 3-5 cm^{-1} steps). When no IR-induced fragment ions were observed, the depletion of the precursor ion was plotted instead, $-\ln[\, I_{precursor} / I_{precursor}^{no\ IR}]$ [33]. The yield is linearly corrected for frequency-dependent variations in laser pulse energy. A grating spectrometer was used to calibrate the laser frequency.

Computational workflow

The SMILES chemical identifier associated with each HMDB entry was used to generate the starting 2D-structure for the workflow. Using the cheminformatics toolbox RDKit [34], protonated, deprotonated and sodiated adduct ions were constructed by considering all nitrogen, oxygen and sulfur atoms as sites for H^+ or Na^+ addition or removal. This was done for all possible tautomers, after which resonance structures were filtered out. Some HMDB entries contained unspecified stereochemistry and were omitted, unless their stereochemistry would not affect the IR spectrum (enantiomers), in which case the stereochemistry was randomly assigned. For each ionized isomer a conformational search using RDKit's distance geometry algorithm was performed to produce 500 random 3D-conformations. After minimization using the MMFF94 classical force field, ten conformations were selected after clustering, or fewer if conformations were too similar (rms-deviation of atom positions <1.4 Å). The selected 3D-geometries were then submitted to Gaussian16 for geometry optimization and frequency calculation at the semi-empirical PM6 level [35]. Unfavorable ionization sites and unfavorable conformations were filtered by their relative energies (electronic + thermal) using a threshold of +40 kJ/mol from the global

minimum. Additionally, geometries that converged to the same local minimum in the optimization were filtered based on their (nearly) identical vibrational spectrum. The remaining geometries were reoptimized using the B3LYP density functional and 6-311+G(d,p) basis set, followed by a frequency calculation. More accurate electronic energies at the MP2/6-311+G(d,p) level were calculated using the B3LYP geometry and combined with the thermal energy from the B3LYP frequency calculation.

The B3LYP/6-311+G(d,p) frequency calculation was used to generate the reference IR spectra to populate the spectral library. The computed frequencies were scaled by a factor of 0.975 to correct for the harmonic approximation used. The stick spectrum was convolved with a Gaussian profile of 45 cm^{-1} full width at half maximum (FWHM).

Scoring of spectral similarity

A search of the library with experimental IR spectra retrieves computed IR spectra sorted by their spectral similarity S_{spec}, derived from the cosine similarity score *i.e.,* the normalized Euclidean dot product of two spectra with a common x-axis with *n* points, represented as vectors ***a*** and ***b***

$$S_{spec} = 1000 \cdot \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}||\boldsymbol{b}|} = 1000 \cdot \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2}\sqrt{\sum_{i=1}^{n} (b_i)^2}}$$

such that $0 \leq S_{spec} \leq 1000$, with a score closer to 1000 indicating greater similarity. To expedite spectral comparisons, the convolved computed spectra were binned once at 3 cm^{-1} intervals (minimum experimental step size) and saved into the library. Spectral comparisons are then performed by evaluating the experimental intensities at these wavenumber points through linear interpolation, which ensures a common x-axis.

Earlier studies used a log transformation of the normalized spectral intensities I_i to make S_{spec} less sensitive to intensity deviations and hence more sensitive to frequency overlap [36]. A similar effect is achieved with a power function [37],

$$I_i^{transformed} = (I_i)^{0.5}$$

Both the values of the exponent (here 0.5) and the Gaussian line broadening (here 45 cm^{-1}) were optimized to provide the best retrieval of the correct spectra (Table S1 in the SI).

Scoring of structural similarity

To define the structural similarity S_{struc} between two molecules, the ionized metabolites were represented by a Morgan2 bit-vector (based on ECFP4) using RDKit [38]. Conceptually, each bit in this 2048-bit vector represents the presence (1) or absence (0) of a specific chemical substructure in the molecule. Molecules with many substructures in common have similar bit vectors and therefore yield a Dice similarity score closer to 1. The combination of Morgan2 and Dice similarity scores was chosen as this gave a more uniform spread across structural similarities compared to other fingerprints (*e.g.* Morgan3) and similarity measures (*e.g.* Tanimoto) [39].

Results and Discussion

Library and validation set

IR spectra were calculated for the protonated ($[M+H]^+$), deprotonated ($[M-H]^-$) and sodiated ($[M+Na]^+$) forms of all entries in the HMDB 4.0 [40] with a molecular weight lower than 210 Da. This amounts to a total of 11823 ions generated from 4640 metabolites. Since several tautomers and conformers are included for each ion, the library contains a total of 75941 computed IR spectra. The experimental validation set consisted of 12 IR spectra taken from earlier publications [8, 14, 17, 41] and 177 newly

measured IRIS spectra derived from 87 metabolites (Table S2 and Scheme S1). Chemical class information of both the library and validation set was determined by ClassyFire [42]. This assessment showed that all major classes in the library are represented by at least 6 reference metabolites (Figure S1) and that the computational library and experimental validation set cover similar distributions in mass and ion type (Figure S2). The experimental and computational IR spectra are available through the HMDB website (https://hmdb.ca), where they are interactively viewable (via JSpectraView) in the "spectrum" field of the *MetaboCard* of the corresponding metabolite [43].

Identifying 3,4-dihydroxyphenylacetic acid

Figure 1 presents a schematic overview of the proposed workflow, where an LC-MS feature with known accurate mass but unknown molecular structure is characterized by recording its IRIS spectrum. This spectrum is then compared to computed spectra of candidate isomers in the library of *in silico* IR ion spectra. The unknown was annotated with the structure that provides the best IR spectral match in terms of the cosine similarity. As a proof of concept, we demonstrate this workflow using the IRIS spectrum of deprotonated 3,4-dihydroxyphenylacetic acid (DOPAC), a metabolite of dopamine. As previously shown, DOPAC is spectroscopically distinguishable from its structural isomer homogentisic acid (HGA), a biomarker for the genetic disorder alkaptonuria [8]. However, the HMDB contains a total of 18 metabolites that share the elemental composition of DOPAC ($C_8H_8O_4$), 16 of which can be deprotonated; their computed IR spectra are present in the *in silico* IR spectral library.

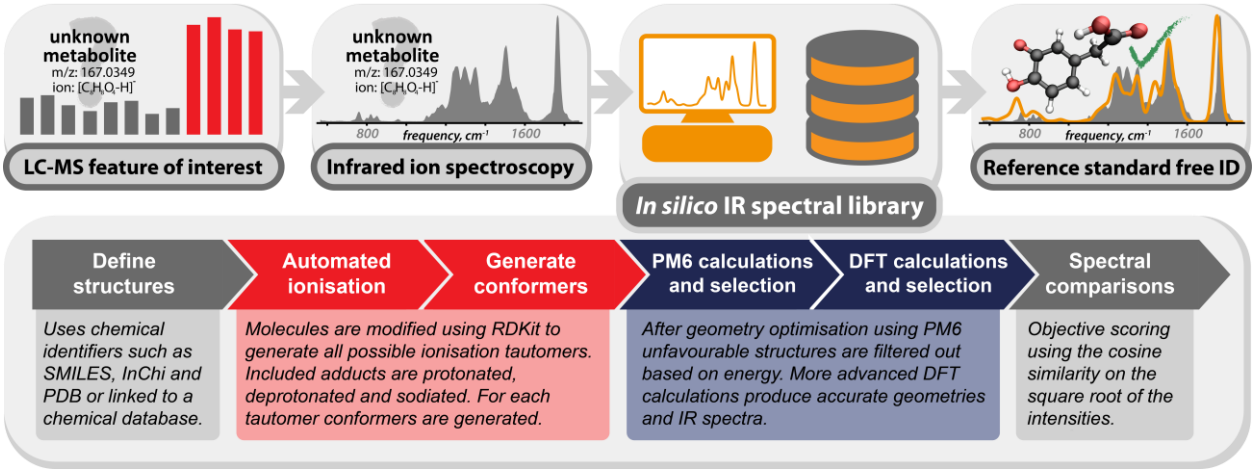


Figure 1. *Schematic workflow for reference standard free metabolite identification. A metabolite with unknown molecular structure, encountered in an untargeted LC-MS screening, is characterized by its IR spectrum measured using infrared ion spectroscopy. The IRIS spectrum is compared against the DFT-computed IR spectra in the library and the ion is annotated with the structure of the best matching library spectrum.*

Figure 2 shows the IR spectrum of $[DOPAC-H]^-$ along with the three best matching spectra in the *in silico* library and their corresponding structures (see results for all 16 isomers in Figure S3). The computed spectrum of $[DOPAC-H]^-$ indeed yields the highest spectral similarity. Interestingly, metabolites that are structurally similar to DOPAC give the second and third highest scores, 3,4-dihydroxymandelaldehyde and 3,5-dihydroxyphenylacetic acid, respectively. These metabolites are the only 1,2-positional isomers of DOPAC in the set of 16 isomers (Figure S3) and have the highest structural similarity with DOPAC (0.56 and 0.78).

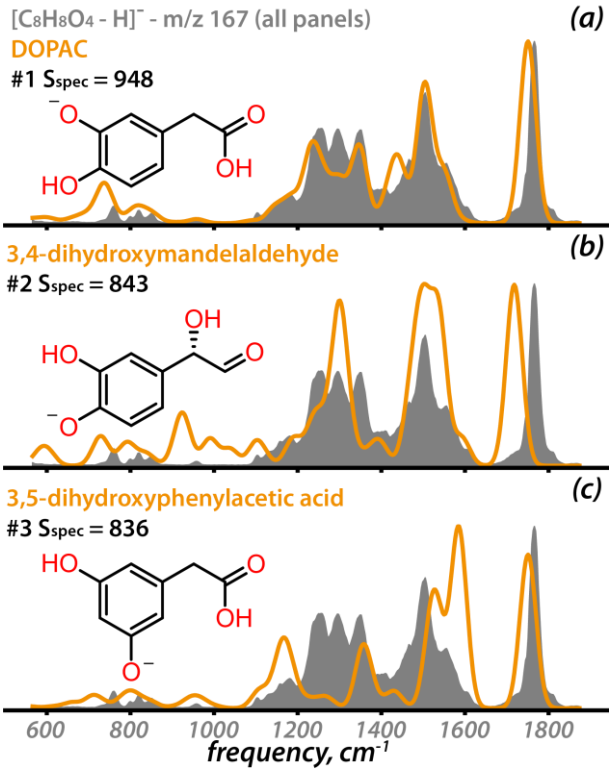


Figure 2. *The three computed vibrational spectra (orange) for $[C_8H_8O_4\text{-}H]^-$ ions in the library giving the best spectral match to the experimental spectrum of $[DOPAC\text{-}H]^-$ (gray). Spectral similarity scores are indicated.*

The use of *in silico* IR spectra as a spectral reference enables us to attribute normal mode vibrations of specific functional groups to each band in the spectrum. This allows spectral mismatches to be correlated with structural mismatches. For example, in Figure 2b, the computed bands in the 900-1050 cm^{-1} range originate from bending vibrations of the aldehyde C-H and the α-hydroxide O-H. Likewise, the band at 1600 cm^{-1} in Figure 2c corresponds to the $C\text{-}O^-$ stretch of the phenoxide. In cases where the actual metabolite is not included in the library, this may suggest new structures based on correctly and incorrectly matching substructure(s). Computing IR spectra for these newly conceived structures is then a route towards *de novo* structure identification of unknown unknowns (*i.e.,* compounds that are not contained in the IR spectral library).

Unsupervised searches for unknown unknowns

The example in Figure 2 demonstrates the potential of IRIS-based identification of known unknowns, based on the inherent sensitivity of an IR spectrum to chemical structure. Not only is the correct metabolite ranked best (known unknown identification), but the search also assigns high scores to entries that are structurally similar. This inherent link between structure and spectroscopy provides venues towards the identification of unknown unknowns. Searching the entire IR library with an adduct constraint, but releasing the constraint on chemical formula (*i.e.,* on exact mass), should assign high spectral similarity scores to compounds with high structural similarity to the unknown, even if the elemental composition is not the same.

For this proof-of-concept, the experimental IRIS spectrum of $[DOPAC\text{-}H]^-$ is again taken as the hypothetical unknown. Figure 3 presents the nine highest-ranked entries upon matching this IRIS spectrum against the entire library, without constraint on chemical formula. The fact that DOPAC itself is ranked #1 out of 2707 deprotonated entries demonstrates both the uniqueness of the experimental IR spectrum, as well as its accurate prediction by DFT calculations. Inspecting the compounds ranked

#2 – #9 reveals the potential for *de novo* structural elucidation via this unsupervised search: the top-9 metabolites are structurally very similar to DOPAC, mostly differing in the length of the alkyl chain, the addition of a methyl or hydroxyl group, or a combination thereof.

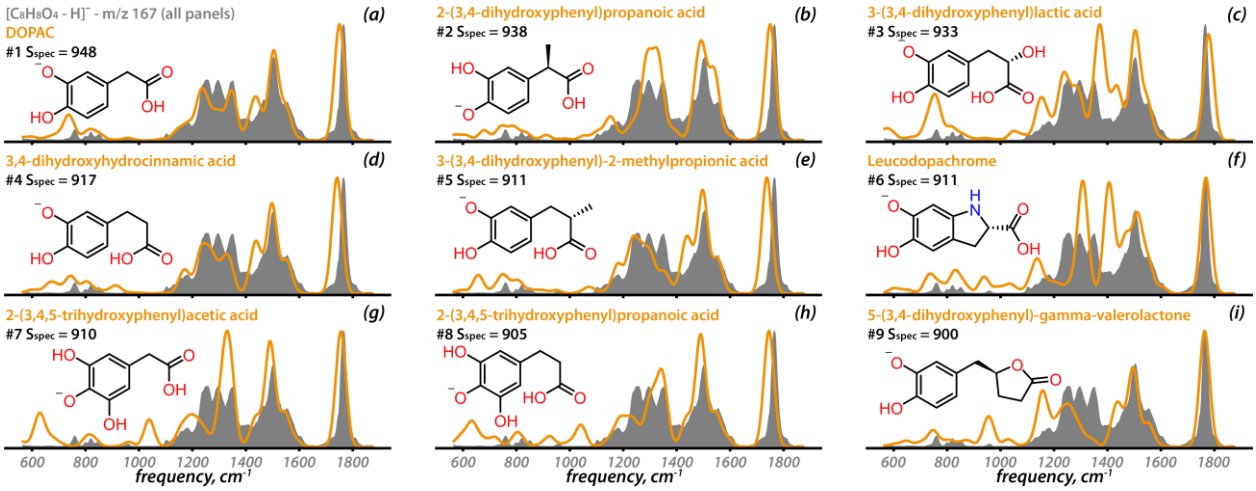


Figure 3. *The nine best matching computed vibrational spectra (orange) in the DFT library with the experimental IR spectrum of [DOPAC-H]⁻ (gray). No m/z or molecular formula constraint was applied (unsupervised). Spectral similarity scores are indicated.*

A strong correlation between molecules with similar structure and molecules with similar spectra is essential for this identification strategy. This is also the basis of recently developed scoring approaches for metabolite identification by tandem MS [39]. To further quantify the ability of the unsupervised IRIS spectral search to find structurally similar metabolites, the structural similarity between each deprotonated metabolite and [DOPAC-H]⁻ was calculated using the Dice similarity score of their bit vectors. For each metabolite in our library, the grey dots in Figure 4 mark the structural similarity to [DOPAC-H]⁻ plotted versus the spectral similarity to its IRIS spectrum. For a large majority of structures with intermediate spectral similarity to [DOPAC-H]⁻, the structural similarity fluctuates greatly. This arises from the fact that coincidental overlap between absorption bands due to different functional groups is common in vibrational spectra. However, the moving average (bin size of 15) shows a steep increase as the spectral score approaches its maximum value of 1000. This indicates that good spectral similarity indeed correlates well with similarity in the type, number and relative positioning of functional groups within a molecule. This is clearly reflected in the top-9 retrieved structures all being very similar to DOPAC (Figure 3).

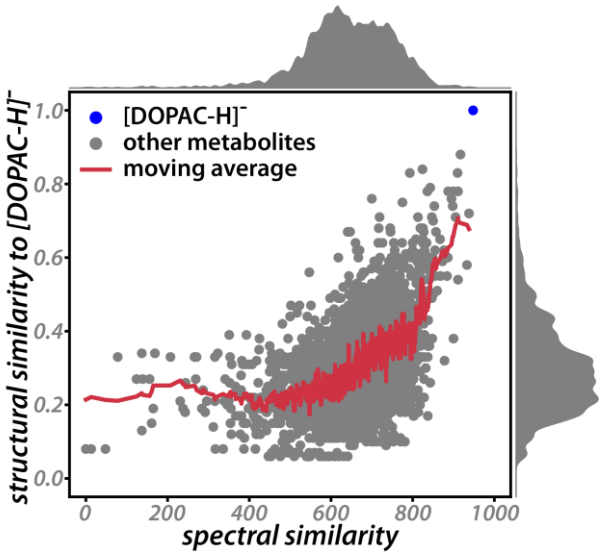


Figure 4. *For each entry in the in silico library, the structural similarity and spectral similarity to [DOPAC-H]⁻ are plotted as the grey dots. [DOPAC-H]⁻ itself is the blue dot. The trend is visualized with a 15-point moving average (red line), excluding the blue data point. Distributions of the structural and spectral similarity are presented on the right and top axis, respectively.*

General performance

To verify how the results for [DOPAC-H]⁻ generalize, a validation set of 189 experimental IRIS spectra was subjected to spectral matching against the *in silico* IR library spectra, analogous to the analysis for DOPAC. The ranking of each IR spectrum match, as well as spectral plots similar to those for DOPAC are available in the Supporting Information. Figure 5a shows how well the metabolites are identified based on their IRIS spectrum and exact *m/z*, plotting the percentage of correctly identified metabolites in the top *k* for increasing *k*. A total of 142 (75%) correct metabolite identifications is achieved when a single experimental IR spectrum is used (red solid trace); 97% of the metabolites are among the top 5 ranked structures. Annotating the experimental IR spectrum to a random isomeric entry from the spectral library correctly identifies 33% of the metabolites (dashed trace), which is relatively high due to entries in the database with only one or a few isomeric structures.

The overall performance can be captured in a single number by taking the geometric mean of the rank of each of the 189 reference metabolites [44]. This rank product (RP) is equal to 1.3 here. When the IR spectra from different adducts ($[M+H]^+$, $[M-H]^-$, $[M+Na]^+$) of the same metabolite are combined (blue solid trace) by ranking on the product of individual spectral similarity scores, the correct identification rate increases to 83% (72 out of 87 metabolites at rank #1) and the RP improves to 1.2. Overall, this performance is higher than typical retrieval rates when employing tandem MS libraries, which typically are 45-70% [45], although true comparison is hampered by the much smaller validation set used here. What the present results do clearly demonstrate is the general applicability of the *in silico* IR spectral library in identifying compounds across a large data set.

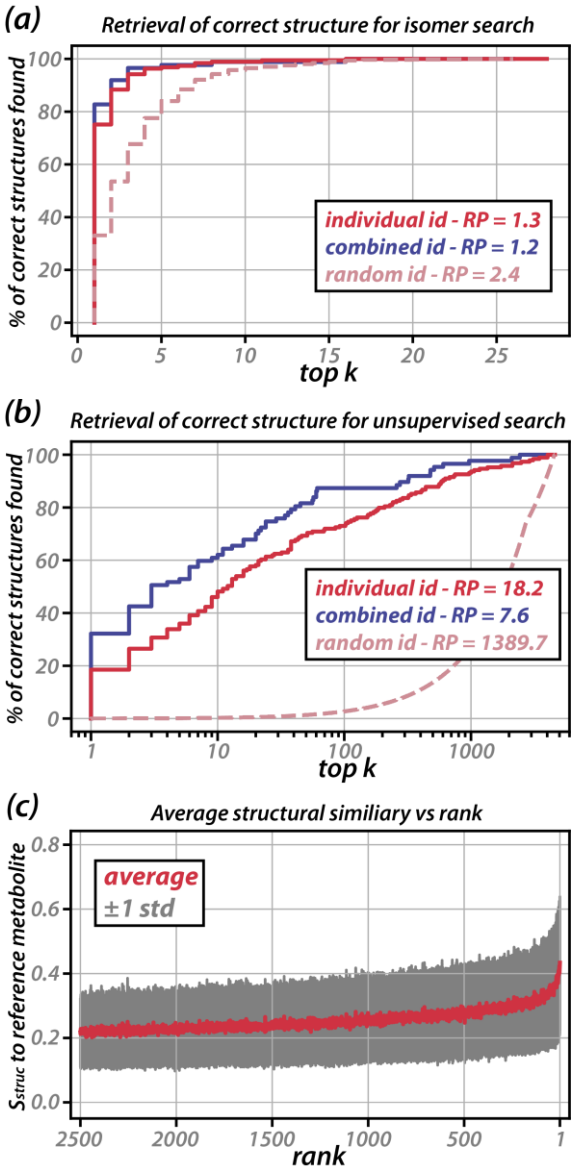


Figure 5. *Percentage of correct structures found in the top k hits when performing an isomer search (a) or an unsupervised search (b). Results for individual adducts are in red (N=189), for adducts combined in purple (N=87); random annotation results are indicated with the dashed line. Panel (c) shows the average structural similarity to the correct metabolite per rank after an unsupervised search.*

A similar plot can be constructed to assess how well metabolites are retrieved in an unsupervised search, *i.e.,* releasing the *m/z* constraint (Figure 5b) and matching the IRIS spectrum against all adduct-constrained computed spectra in the IR library. In this case, 19% of the identifications are correct, while 48% are scored in the top 10. The specificity improves significantly when IR spectra of different adducts ($[M+H]^+$, $[M-H]^-$, $[M+Na]^+$) are combined, boosting the top 1 and top 10 percentages to 32% and 62%, respectively. The RPs follow a similar trend, improving from 18.2 to 7.6. For the unsupervised search, the contrast with randomly annotating the experimental spectra is pronounced, with just 0.03% correct identification, as the number of adduct-constrained candidates increases to 3892 on average.

The effectiveness of combining spectra of different adducts is best illustrated with acetylglycine (HMDB0000532), for which the individual adducts rank 10th ($[M-H]^-$), 20th ($[M+H]^+$) and 6th ($[M+Na]^+$) in an unsupervised search, but combined they rank 1st (Table S3). Decomposition of Figure 5b into the individual adducts shows that the $[M+H]^+$ adducts are retrieved best from the library, followed by

$[M+Na]^+$ and then $[M-H]^-$ (RPs are 12.4, 15.8 and 30.0, respectively). A discussion of this trend is presented in the SI along with Figure S4.

The correlation between spectral similarity and structural similarity for the whole validation set is visualized by plotting the average structural similarity between each reference metabolite and the ranked library structures, where the reference metabolites themselves are removed from the data (Figure 5c). On average, a clear increase in structural similarity is observed when approaching rank 1, with a steep increase for the ~50 best matches. Inspection of the individual structure vs spectral similarity plots (see SI) suggests that the correlation between structural similarity and spectral similarity for some metabolites is limited by the number of structural analogues present in the library (*e.g.* Figure S5 in the SI). An expansion of the library to include more metabolites would therefore not only allow for the direct identification of more metabolites, but also make unsupervised searches more sensitive to substructure, as more structural analogues would be included.

Higher-energy geometries

To assess where the scoring can be further improved, we manually inspected the computed geometries and computed IR spectra for each reference metabolite and compared these with the experimental IR spectra. It appears that about 10% of ions adopt a higher-energy conformational or tautomeric geometry, with a significantly different computed IR spectrum. These higher-energy geometries lie up to 68.8 kJ/mol above the lowest-energy geometry (Figures S6 and S7). The presence of higher-energy geometries is not uncommon and may be due to inaccuracies in the calculated energies or kinetic trapping of solution-phase conformers or tautomers that are transferred to the gas phase [46, 47].

To account for these higher-energy geometries, the spectral library can be searched with an energy tolerance, where the ranking is based on the best matching computed spectrum per entry. For a tolerance of 10 kJ/mol, this results in an overall better performance as derived from the RP, which improves from 18.2 to 17.0 in the unsupervised search. However, the 10 kJ/mol tolerance increases the number of considered spectra about 3-fold (to 10811 on average), possibly leading to a worse ranking for metabolites that already performed well. This effect becomes pronounced when all adduct-constrained spectra in the *in silico* library are searched, *i.e.* without energy constraints (24865 spectra on average, Table S3), yielding a poorer RP of 19.0.

For some ions, a mix of conformers or tautomers may be present, each with distinct IR spectra (Figures S8 and S9), which may be addressed by optimizing linear combinations of computed vibrational spectra [48]. However, this is not pursued here, as it would severely slow down the evaluation of spectral similarities and likely give poorer performance for metabolites that exhibit a "pure" population. Moreover, the benefits are likely small as mixtures are not frequently observed (<15% based on manual inspection) and minimally affect the IR spectrum when the contribution of the minor population is small (Figure S10). Methods that experimentally deconvolute the mixtures, *e.g.* by separating conformers so tautomers with ion mobility spectrometry [23, 47] or by employing 2-color laser experiments [49], could improver heir identification analogous to using spectra of multiple adducts.

Broader spectra

Severe spectral broadening observed in some IRIS spectra can limit the performance of our method. Extensive broadening is often caused by strong ionic hydrogen bonds that induce shared-proton motifs [50-53]. Such problematic spectra may be avoided by selecting different adducts. For instance, for L-aspartic acid (HMDB0000191), the $[M-H]^-$ ion shows severe broadening [54], while both $[M+H]^+$ and $[M+Na]^+$ ions do not (Figure S11) [55]. We also note that spectral broadening in such systems is often reduced on cryogenic tandem mass spectrometry platforms [15, 56-58]. Alternatively, IRIS identification

may be applied to MS/MS fragments, which typically exhibit better resolved spectra [17, 41], in combination with a bottom-up approach for structure elucidation. This also extends the applicability of the IR spectral library towards molecules beyond its upper mass limit [22, 59].

Identification from a human plasma sample

To demonstrate the applicability of the *in silico* IR spectral library to actual biofluids, we elucidated the molecular structure of an LC-MS feature (-ESI, *m/z* 131.0713, RT=6.27 min.) elevated in a patient's plasma sample (see SI for experimental details). The measured *m/z* suggests a chemical formula of $[C_6H_{12}O_3\text{-}H]^-$ (Δm = +7.6 ppm), which corresponds with 10 entries in our spectral library. Figure 6 shows the experimental IR spectrum of the LC-MS feature and the three best matching computed IR spectra. All ten spectral comparisons are shown in Figure S12. The three top-ranked structures all possess a carboxylate moiety an a nearby hydroxyl group, giving rise to three main spectral bands between 1200 and 1700 cm^{-1} that are also observed in the IRIS spectrum. Specifically, these bands originate from symmetric and antisymmetric carboxylate O-C-O stretching (1300 and 1650 cm^{-1}) [60] and O-H bending (1450 cm^{-1}). The 3-hydroxyhexanoate anion is ranked highest, as the experimental band at 850 cm^{-1} is also reproduced, corresponding to the O-H out-of-plane bending vibration (Figure 6a). For the 2-hydroxy isomers, this O-H bending vibration is computed at 700 cm^{-1} (Figure 6b-c). This assignment is further corroborated by inspecting the top-25 matches of an unsupervised search, which yields the 3-hydroxycarboxylate motif in 13 out of 25 structures, while the 2-hydroxycarboxylate substructure occurs only once (see Figure S13).

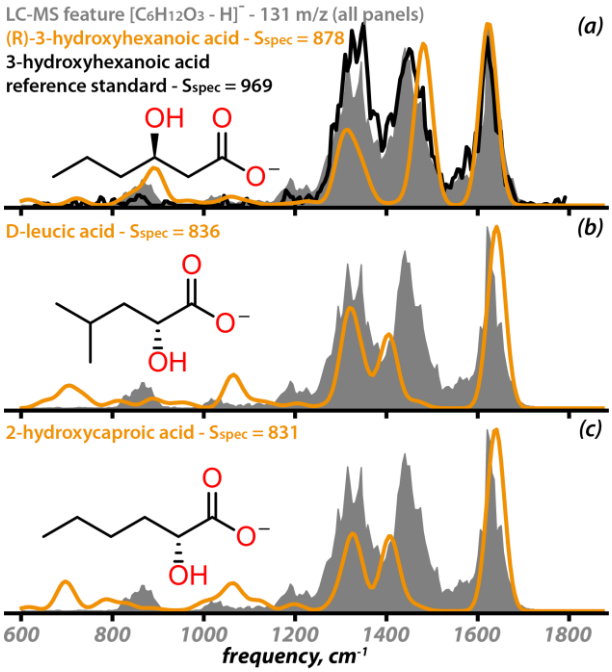


Figure 6. *The three best matching computed vibrational spectra (orange) obtained in a library query with the IRIS spectrum of an unknown LC-MS feature (gray) with chemical formula* $[C_6H_{12}O_3\text{-}H]^-$*. The experimental spectrum of the 3-hydroxyhexanoate ([M-H]⁻) reference standard is added to the top panel (black). Spectral similarity scores are indicated.*

The annotation of the LC-MS feature as 3-hydroxyhexanoic acid was confirmed by measuring an IRIS spectrum for a reference standard of this molecule (Figure 6a), which indeed gives high spectral similarity (S_{spec} = 969) with the spectrum of the LC-MS feature. The stereochemistry of the hydroxide remains ambiguous because it cannot be determined directly from an IR measurement. In general, 3-hydroxycarboxylic acids are biomarkers for fatty acid oxidative disorders of both long- and short-chain

3-hydroxyacyl-CoA dehydrogenases [61]. Specifically, 3-hydroxyhexanoic acid is increased in the serum of diabetic ketoacidotic patients [62] and has been observed in body fluids of patients with 3-hydroxy-3-methylglutaryl-CoA synthase deficiency [63] and metastatic melanoma [64]. The patient in this case was in ketosis as exemplified by the high body fluid concentrations of acetoacetate and 3-hydroxybutyric acid. After the identification of this *m/z* 131.0713 feature, we observed 3-hydroxyhexanoic acid in many plasma samples of ketotic patients. As such 3-hydroxyhexanoic acid may serve as ketosis biomarker.

Conclusions and Outlook

The compilation and utilization of an *in silico* IR spectral library of ionized molecules for the identification of unknown metabolites using MS-based IRIS experiments has been demonstrated. An automated workflow to produce IR spectra of molecular ions generated over 75000 DFT-calculated vibrational spectra for 4640 metabolites taken from the HMDB. A scoring algorithm based on cosine similarity was employed to identify the molecular structures that match favorably with a user supplied experimental IR ion spectrum. By collecting a set of 189 experimental IRIS spectra we evaluated the performance of the *in silico* IR spectral library in the identification of metabolites. With a known accurate mass value and working within the boundaries of our data set, 75% of metabolites was correctly identified, which further improves to 83% by simultaneous identification of multiple ionic adducts of the same metabolite.

We also explored the potential of an *unsupervised* search, where an experimental IRIS spectrum is compared against the entire library, without *m/z* constraints. This strategy to identify molecular substructures in an unknown molecule relies on the strong spectrum-structure correlation of vibrational spectroscopy and is especially valuable for the *de novo* identification of metabolites not included in the library.

Manual inspection of the spectral comparisons revealed that higher-energy conformers and tautomers occur in about 10-15% of cases. Including these higher-energy geometries by using a tolerance on relative energy (that can be set in the spectral scoring procedure) improved the performance by assigning geometries that were not considered before. However, as a trade-off, for metabolites that match well, the inclusion of more candidates reduces the correct identification rate. An intermediate energy cut-off of about 10 kJ/mol was found to yield optimal results. Isomeric mixtures have not been addressed, but experiments involving LC, IMS or 2-color laser spectroscopy are expected to reduce such cases. The utility of the *in silico* IR spectral library was demonstrated in the identification of a human plasma metabolite as 3-hydroxyhexanoic acid.

The datasets presented here are larger than any previously reported set of experimental or computational IR ion spectra and provide further opportunities for improvement of the metabolite identification workflow. For instance, adapting the scoring method to better capture structural similarity [39], applying chemical element-dependent frequency scaling to better correct for anharmonic shifts [65], integrating IRIS scoring with MS/MS scoring algorithms [66], or tackling the spectrum-to-structure conversion directly [66] may be realized with this data set. The experimental and computation spectra are now available through the HMDB website and will be added to the Spectra Search interface in its next release (HMDB 6.0). Moreover, we will continue to expand both the experimental and *in silico* IR spectral library. A promising prospect in this regard is the development of machine-learned density functionals, which should significantly speed-up DFT calculations [67]. This could extend the feasibility of our approach to molecules of increased size and with a much larger coverage of chemical space, thereby further establishing infrared ion spectroscopy as an appealing route for small-molecule identification far beyond metabolomics alone.

Supporting information

- LC-MS procedure, optimization of the spectral similarity scoring, additional sample information and tables/figures in support of the results (word file)
- Excel versions of tables in the supporting information (excel file)
- Spectral comparisons for isomer searching of the library (pdf file)
- Spectral comparisons for unsupervised searching of the library (pdf file)
- Spectral comparisons with all computed conformers and tautomers for all experimental spectra (pdf file)
- Datasets are available at the human metabolome database (https://hmdb.ca) and Zenodo (https://doi.org/10.5281/zenodo.7706021)

Notes

The authors declare no competing financial interest. The work described in this study has been carried out in accordance with The Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving humans. All patients (or their guardians) approved of the possible use of their anonymized left-over samples for method validation purposes, in agreement with institutional and national legislation.

Acknowledgements

We gratefully acknowledge the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) for the support of the FELIX Laboratory through the research program "National Roadmap Grootschalige Wetenschappelijke Infastructuur" 184.034.022. This project further received funding from NWO-TTW (grant 15769) and NWO Rekentijd for the computational resources (grant 2021.055); we thank the SURFsara Supercomputer Centre staff for their continuous support and the Canada Foundation for Innovation (grant 35456) for financial support for the HMDB.

References

1. Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R., Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2013,** *9*, 44-66.
2. Aretz, I.; Meierhofer, D., Advantages and pitfalls of mass spectrometry based metabolome profiling in systems biology. *Int. J. Mol. Sci.* **2016,** *17*, 632.
3. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015,** *12*, 523-526.
4. Luo, M.-D.; Zhou, Z.-W.; Zhu, Z.-J., The application of ion mobility-mass spectrometry in untargeted metabolomics: From separation to identification. *Journal of Analysis and Testing* **2020,** *4*, 163-174.
5. Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; Wolan, D. W.; Spilker, M. E.; Benton, H. P.; Siuzdak, G., METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* **2018,** *90*, 3156-3164.
6. Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, B. L.; Berjanskii, M.; Mah, R.; Yamamoto, M.; Jovel, J.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Lui, V. W.; Varshavi, D.; Allen, D.; Arndt, D.; Khetarpal, N.; Sivakumaran, A.; Harford, K.; Sanford, S.; Yee, K.; Cao, X.; Budinski, Z.; Liigand, J.; Zhang, L.; Zheng, J.; Mandal, R.; Karu, N.; Dambrova, M.; Schiöth, H. B.; Greiner, R.; Gautam, V., HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res.* **2022,** *50*, D622-D631.

7. Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O., FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Anal. Chem.* **2009,** *81*, 10038-10048.
8. Martens, J.; van Outersterp, R. E.; Vreeken, R. J.; Cuyckens, F.; Coene, K. L. M.; Engelke, U. F.; Kluijtmans, L. A. J.; Wevers, R. A.; Buydens, L. M. C.; Redlich, B.; Berden, G.; Oomens, J., Infrared ion spectroscopy: New opportunities for small-molecule identification in mass spectrometry-A tutorial perspective. *Anal. Chim. Acta* **2020,** *1093*, 1-15.
9. Schindler, B.; Laloy-Borgna, G.; Barnes, L.; Allouche, A.-R.; Bouju, E.; Dugas, V.; Demesmay, C.; Compagnon, I., Online separation and identification of isomers using infrared multiple photon dissociation ion spectroscopy coupled to liquid chromatography: application to the analysis of disaccharides regio-isomers and monosaccharide anomers. *Anal. Chem.* **2018,** *90*, 11741-11745.
10. van Outersterp, R. E.; Oosterhout, J.; Gebhardt, C. R.; Berden, G.; Engelke, U. F.; Wevers, R. A.; Cuyckens, F.; Oomens, J.; Martens, J., Targeted small molecule identification using heartcutting liquid chromatography–infrared ion spectroscopy. *Anal. Chem.* **2023,** *95*, 3406-3413.
11. Engelke, U. F. H.; van Outersterp, R. E.; Merx, J.; van Geenen, F. A. M. G.; van Rooij, A.; Berden, G.; Huigen, M. C. D. G.; Kluijtmans, L. A. J.; Peters, T. M. A.; Al-Shekaili, H. H.; Leavitt, B. R.; de Vrieze, E.; Broekman, S.; van Wijk, E.; Tseng, L. A.; Kulkarni, P.; Rutjes, F. P. J. T.; Mecinović, J.; Struys, E. A.; Jansen, L. A.; Gospe Jr., S. M.; Mercimek-Andrews, S.; Hyland, K.; Willemsen, M. A. A. P.; Bok, L. A.; van Karnebeek, C. D. M.; Wevers, R. A.; Boltje, T. J.; Oomens, J.; Martens, J.; Coene, K. L. M., Untargeted metabolomics and infrared ion spectroscopy identify biomarkers for pyridoxine-dependent epilepsy. *The Journal of clinical investigation* **2021,** *131*.
12. van Outersterp, R. E.; Moons, S. J.; Engelke, U. F. H.; Bentlage, H.; Peters, T.; van Rooij, A.; Huigen, M. C. D. G.; de Boer, S.; van der Heeft, E.; Kluijtmans, L. A. J.; Van Karnebeek, C. D. M.; Wevers, R. A.; Berden, G.; Oomens, J.; Boltje, T. J.; Coene, K. L. M.; Martens, J., Amadori rearrangement products as potential biomarkers for inborn errors of amino-acid metabolism. *Communications biology* **2021,** *4*, 1-8.
13. Peters, T. M. A.; Merx, J.; Kooijman, P. C.; Noga, M.; de Boer, S.; van Gemert, L. A.; Salden, G.; Engelke, U. F. H.; Lefeber, D. J.; van Outersterp, R. E.; Berden, G.; Boltje, T. J.; Artuch, R.; Pías-Peleteiro, L.; García-Cazorla, Á.; Barić, I.; Thöny, B.; Oomens, J.; Martens, J.; Wevers, R. A.; Verbeek, M. M.; Coene, K. L. M.; Willemsen, M. A. A. P., Novel cerebrospinal fluid biomarkers of glucose transporter type 1 deficiency syndrome: Implications beyond the brain's energy deficit. *J. Inherit. Metab. Dis.* **2022**, 1-10.
14. Kranenburg, R. F.; van Geenen, F. A. M. G.; Berden, G.; Oomens, J.; Martens, J.; van Asten, A. C., Mass-spectrometry-based identification of synthetic drug isomers using infrared ion spectroscopy. *Anal. Chem.* **2020,** *92*, 7282-7288.
15. Kirschbaum, C.; Greis, K.; Mucha, E.; Kain, L.; Deng, S.; Zappe, A.; Gewinner, S.; Schöllkopf, W.; von Helden, G.; Meijer, G.; Savage, P. B.; Marianski, M.; Teyton, L.; Pagel, K., Unravelling the structural complexity of glycolipids with cryogenic infrared spectroscopy. *Nature communications* **2021,** *12*, 1-7.
16. Vink, M. J. A.; van Geenen, F. A. M. G.; Berden, G.; O'Riordan, T. J. C.; Howe, P. W. A.; Oomens, J.; Perry, S. J.; Martens, J., Structural Elucidation of Agrochemicals and Related Derivatives Using Infrared Ion Spectroscopy. *Environ. Sci. Technol.* **2022,** *56*, 15563-15572.
17. van Outersterp, R. E.; Houthuijs, K. J.; Berden, G.; Engelke, U. F.; Kluijtmans, L. A. J.; Wevers, R. A.; Coene, K. L. M.; Oomens, J.; Martens, J., Reference-standard free metabolite identification using infrared ion spectroscopy. *Int. J. Mass spectrom.* **2019,** *443*, 77-85.
18. MacAleese, L.; Maître, P., Infrared spectroscopy of organometallic ions in the gas phase: from model to real world complexes. *Mass Spectrom. Rev.* **2007,** *26*, 583-605.
19. Roithová, J., Characterization of reaction intermediates by ion spectroscopy. *Chem. Soc. Rev.* **2012,** *41*, 547-559.
20. ter Braak, F.; Elferink, H.; Houthuijs, K. J.; Oomens, J.; Martens, J.; Boltje, T. J., Characterization of Elusive Reaction Intermediates Using Infrared Ion Spectroscopy: Application to the Experimental Characterization of Glycosyl Cations. *Acc. Chem. Res.* **2022**, 6034-6038.

21. Munshi, M. U.; Berden, G.; Martens, J.; Oomens, J., Gas-phase vibrational spectroscopy of triphenylamine: the effect of charge on structure and spectra. *PCCP* **2017,** *19*, 19881-19889.
22. Schindler, B.; Barnes, L.; Renois, G.; Gray, C.; Chambert, S.; Fort, S.; Flitsch, S.; Loison, C.; Allouche, A.-R.; Compagnon, I., Anomeric memory of the glycosidic bond upon fragmentation and its consequences for carbohydrate sequencing. *Nature communications* **2017,** *8*, 973.
23. Ben Faleh, A.; Warnke, S.; Bansal, P.; Pellegrinelli, R. P.; Dyukova, I.; Rizzo, T. R., Identification of mobility-resolved N-glycan isomers. *Anal. Chem.* **2022,** *94*, 10101-10108.
24. Aksenov, A. A.; da Silva, R.; Knight, R.; Lopes, N. P.; Dorrestein, P. C., Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry* **2017,** *1*, 1-20.
25. Collins, S. L.; Koo, I.; Peters, J. M.; Smith, P. B.; Patterson, A. D., Current Challenges and Recent Developments in Mass Spectrometry–Based Metabolomics. *Annual Review of Analytical Chemistry* **2021,** *14*, 467-487.
26. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics: Focus Review. *Chem. Rev.* **2021,** *121*, 5633-5670.
27. Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S., CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* **2021,** *93*, 11692-11700.
28. Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; Renslow, R. S., ISiCLE: a quantum chemistry pipeline for establishing in silico collision cross section libraries. *Anal. Chem.* **2019,** *91*, 4346-4356.
29. Karunaratne, E.; Hill, D. W.; Pracht, P.; Gascón, J. A.; Grimme, S.; Grant, D. F., High-Throughput Non-targeted Chemical Structure Identification Using Gas-Phase Infrared Spectra. *Anal. Chem.* **2021,** *93*, 10688-10696.
30. 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.* **2016,** *87*, 103108.
31. Oepts, D.; Van der Meer, A. F. G.; Van Amersfoort, P. W., The free-electron-laser user facility FELIX. *Infrared physics & technology* **1995,** *36*, 297-308.
32. van Outersterp, R. E.; Engelke, U. F. H.; Merx, J.; Berden, G.; Paul, M.; Thomulka, T.; Berkessel, A.; Huigen, M. C. D. G.; Kluijtmans, L. A.; Mecinović, J.; Rutjes, F. P.; Van Karnebeek, C. D. M.; Wevers, R. A.; Boltje, T. J.; Coene, K. L. M.; Martens, J.; Oomens, J., Metabolite Identification Using Infrared Ion Spectroscopy– Novel Biomarkers for Pyridoxine-Dependent Epilepsy. *Anal. Chem.* **2021,** *93*, 15340-15348.
33. Berden, G.; Derksen, M.; Houthuijs, K. J.; Martens, J.; Oomens, J., An automatic variable laser attenuator for IRMPD spectroscopy and analysis of power-dependence in fragmentation spectra. *Int. J. Mass spectrom.* **2019,** *443*, 1-8.
34. Landrum, G., RDKit: Open-source cheminformatics. **2006**.
35. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.
36. Kempkes, L. J. M.; Martens, J.; Berden, G.; Houthuijs, K. J.; Oomens, J., Investigation of the position of the radical in z 3-ions resulting from electron transfer dissociation using infrared ion spectroscopy. *Faraday Discuss.* **2019,** *217*, 434-452.

37. Stein, S. E.; Scott, D. R., Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.* **1994,** *5*, 859-866.
38. Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010,** *50*, 742-754.
39. Huber, F.; Ridder, L.; Verhoeven, S.; Spaaks, J. H.; Diblen, F.; Rogers, S.; van der Hooft, J. J. J., Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **2021,** *17*, e1008724.
40. Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Loc, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A., HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018,** *46*, D608-D617.
41. Martens, J.; Berden, G.; Bentlage, H.; Coene, K. L. M.; Engelke, U. F. H.; Wishart, D. S.; van Scherpenzeel, M.; Kluijtmans, L. A. J.; Wevers, R. A.; Oomens, J., Unraveling the unknown areas of the human metabolome: the role of infrared ion spectroscopy. *J. Inherit. Metab. Dis.* **2018,** *41*, 367-377.
42. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Wishart, D. S.; Greiner, R., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016,** *8*, 1-20.
43. Houthuijs, K. J.; Berden, G.; Engelke, U. F.; Gautam, V.; Wishart, D. S.; Wevers, R. A.; Martens, J.; Oomens, J., Dataset associated with: "An in silico infrared spectral library of molecular ions for metabolite identification". Zenodo https://doi.org/10.5281/zenodo.7706021: 2023.
44. Breitling, R.; Armengaud, P.; Amtmann, A.; Herzyk, P., Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **2004,** *573*, 83-92.
45. Schymanski, E. L.; Ruttkies, C.; Krauss, M.; Brouard, C.; Kind, T.; Dührkop, K.; Allen, F.; Vaniya, A.; Verdegem, D.; Böcker, S.; Rousu, J.; Shen, H.; Tsugawa, H.; Sajed, T.; Fiehn, O.; Ghesquière, B.; Neumann, S., Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.* **2017,** *9*, 1-21.
46. Almasian, M.; Grzetic, J.; van Maurik, J.; Steill, J. D.; Berden, G.; Ingemann, S.; Buma, W. J.; Oomens, J., Non-equilibrium isomer distribution of the gas-phase photoactive yellow protein chromophore. *The journal of physical chemistry letters* **2012,** *3*, 2259-2263.
47. Warnke, S.; Seo, J.; Boschmans, J.; Sobott, F.; Scrivens, J. H.; Bleiholder, C.; Bowers, M. T.; Gewinner, S.; Schöllkopf, W.; Pagel, K.; von Helden, G., Protomers of benzocaine: solvent and permittivity dependence. *J. Am. Chem. Soc.* **2015,** *137*, 4236-4242.
48. Barnes, L.; Allouche, A.-R.; Chambert, S.; Schindler, B.; Compagnon, I., Ion spectroscopy of heterogeneous mixtures: IRMPD and DFT analysis of anomers and conformers of monosaccharides. *Int. J. Mass spectrom.* **2020,** *447*, 116235.
49. van Geenen, F. A. M. G.; Kranenburg, R. F.; van Asten, A. C.; Martens, J.; Oomens, J.; Berden, G., Isomer-specific two-color double-resonance IR2MS3 ion spectroscopy using a single laser: application in the identification of novel psychoactive substances. *Anal. Chem.* **2021,** *93*, 2687-2693.
50. Asmis, K. R.; Pivonka, N. L.; Santambrogio, G.; Brümmer, M.; Kaposta, C.; Neumark, D. M.; Wöste, L., Gas-phase infrared spectrum of the protonated water dimer. *Science* **2003,** *299*, 1375-1377.
51. Moore, D. T.; Oomens, J.; van der Meer, L.; von Helden, G.; Meijer, G.; Valle, J.; Marshall, A. G.; Eyler, J. R., Probing the vibrations of shared, OH+ O-bound protons in the gas phase. *Chemphyschem* **2004,** *5*, 740-743.
52. Fridgen, T. D.; MacAleese, L.; Maitre, P.; McMahon, T. B.; Boissel, P.; Lemaire, J., Infrared spectra of homogeneous and heterogeneous proton-bound dimers in the gas phase. *PCCP* **2005,** *7*, 2747-2755.
53. Roscioli, J. R.; McCunn, L. R.; Johnson, M. A., Quantum structure of the intermolecular proton bond. *Science* **2007,** *316*, 249-254.
54. Oomens, J.; Steill, J. D.; Redlich, B., Gas-phase IR spectroscopy of deprotonated amino acids. *J. Am. Chem. Soc.* **2009,** *131*, 4310-4319.

55. O'Brien, J. T.; Prell, J. S.; Steill, J. D.; Oomens, J.; Williams, E. R., Interactions of mono-and divalent metal ions with aspartic and glutamic acid investigated with IR photodissociation spectroscopy and theory. *The Journal of Physical Chemistry A* **2008,** *112*, 10823-10830.
56. Gorlova, O.; Colvin, S. M.; Brathwaite, A.; Menges, F. S.; Craig, S. M.; Miller, S. J.; Johnson, M. A., Identification and partial structural characterization of mass isolated valsartan and its metabolite with messenger tagging vibrational spectroscopy. *J. Am. Soc. Mass. Spectrom.* **2017,** *28*, 2414-2422.
57. Bell, M. R.; Tesler, L. F.; Polfer, N. C., Cryogenic infrared ion spectroscopy for the structural elucidation of drug molecules: MDMA and its metabolites. *Int. J. Mass spectrom.* **2019,** *443*, 101-108.
58. Warnke, S.; Ben Faleh, A.; Rizzo, T. R., Toward High-Throughput Cryogenic IR Fingerprinting of Mobility-Separated Glycan Isomers. *ACS measurement science Au* **2021,** *1*, 157-164.
59. Bansal, P.; Faleh, A. B.; Warnke, S.; Rizzo, T. R., Identification of N-glycan positional isomers by combining IMS and vibrational fingerprinting of structurally determinant CID fragments. *Analyst* **2022,** *147*, 704-711.
60. Oomens, J.; Steill, J. D., Free carboxylate stretching modes. *The Journal of Physical Chemistry A* **2008,** *112*, 3281-3283.
61. Jones, P. M.; Moffitt, M.; Joseph, D.; Harthcock, P. A.; Boriack, R. L.; Ibdah, J. A.; Strauss, A. W.; Bennett, M. J., Accumulation of free 3-hydroxy fatty acids in the culture media of fibroblasts from patients deficient in long-chain l-3-hydroxyacyl-CoA dehydrogenase: a useful diagnostic aid. *Clin. Chem.* **2001,** *47*, 1190-1194.
62. Niwa, T.; Yamada, K.; Ohki, T.; Furukawa, H., 3-Hydroxyhexanoic acid: An abnormal metabolite in urine and serum of diabetic ketoacidotic patients. *Journal of Chromatography B: Biomedical Sciences and Applications* **1985,** *337*, 1-7.
63. Thompson, G. N.; Hsu, B. Y. L.; Pitt, J. J.; Treacy, E.; Stanley, C. A., Fasting hypoketotic coma in a child with deficiency of mitochondrial 3-hydroxy-3-methylglutaryl-CoA synthase. *N. Engl. J. Med.* **1997,** *337*, 1203-1207.
64. Frankel, A. E.; Coughlin, L. A.; Kim, J.; Froehlich, T. W.; Xie, Y.; Frenkel, E. P.; Koh, A. Y., Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients. *Neoplasia* **2017,** *19*, 848-855.
65. Pracht, P.; Grant, D. F.; Grimme, S., Comprehensive assessment of GFN tight-binding and composite density functional theory methods for calculating gas-phase infrared spectra. *J. Chem. Theory Comput.* **2020,** *16*, 7044-7060.
66. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences* **2015,** *112*, 12580-12585.
67. Kirkpatrick, J.; McMorrow, B.; Turban, D. H. P.; Gaunt, A. L.; Spencer, J. S.; Matthews, A. G. D. G.; Obika, A.; Thiry, L.; Fortunato, M.; Pfau, D.; Castellanos, L. R.; Petersen, S.; Nelson, A. W. R.; Hohli, P.; Mori-Sánchez, P.; Hassabis, D.; Cohen, A. J., Pushing the frontiers of density functionals by solving the fractional electron problem. *Science* **2021,** *374*, 1385-1389.