Computing fragmentation trees from tandem mass spectrometry data

SUPPORTING INFORMATION

Florian Rasche[1], Aleš Svatoš[2], Ravi Kumar Maddula[2], Christoph Böttcher[3] & Sebastian Böcker[1]*

[1]Chair for Bioinformatics, Friedrich-Schiller-University Jena, Ernst-Abbe-Platz 2, D-07743 Jena, Germany

[2]Research Group Mass Spectrometry, Max Planck Institute for Chemical Ecology, Hans-Knöll-Straße 8, D-07745 Jena, Germany

[3]Department of Stress and Developmental Biology, Leibniz Institute of Plant Biochemistry, Weinberg 3, D-06120 Halle, Germany

Correspondence to: **sebastian.boecker@uni-jena.de**

Contents

Fragmentation tree calculation

In the following, we assume that the molecular formula of the unknown compound is known, and we want to compute its fragmentation tree. The algorithm for deciding upon one molecular formula, based on the fragmentation spectrum, proceeds in the same manner, taking the different decompositions of the mono-isotopic mass as candidates. The computation is performed in three steps: First, we construct a fragmentation graph which contains all possible fragmentation trees that are in accordance with the experimental data; see Fig. 2. In the second step, we score the fragmentation graph: We weight arcs using properties such as peak intensities, mass deviations, common NLs. In the final step, we search for the highest-scoring fragmentation tree inside the fragmentation graph; see Fig. 2 and S-9 for a more complex example.

Generating the fragmentation graph For every peak of the fragmentation spectrum, we compute all molecular formulas that are within the mass accuracy of the instrument and that are sub-formulas of the compound molecular formula. Since the relative mass accuracy drops in the low mass range, an absolute error may also be specified, the larger of which is used. The number of molecular formulas to explain a single peak ranged from 1 to 10, with an average of 2.7 (Thermo Orbitrap), 2.0 (API QSTAR), and 3.1 molecular formulas (Micromass QTOF), respectively. These numbers obviously depend on mass accuracy and the mass of the compound. We use these molecular formulas as the vertices of our fragmentation graph. Vertices are colored so that two molecular formulas corresponding to the same peak also receive the same color. Next, we draw directed edges (arcs) between pairs of vertices: Two vertices are connected by a directed edge if the second molecular formula is a sub-formula of the first.

Weighting the graph Now, we weight the fragmentation graph, based on the probability that a certain vertex or edge is "true": Candidates will be ranked by our algorithm based on the sum of these scores. To this end, it is reasonable to assign scores based on log likelihoods or log odds, because this enables a statistical interpretation of the outcome (i.e., maximum likelihood): Summing log likelihood equals the log product of these likelihoods, and maximum likelihood is identical to maximum log likelihood. For vertices, we use log odds to differentiate between the model (the peak is truly a fragment with the proposed molecular formula) and the background (the peak is noise). Under the model, it is impossible to predict the relative intensity of the fragment peak solely from its molecular formula. But we can use the mass difference between the measured peak and the molecular formula to assess whether it is true: Mass differences are usually assumed to be normally distributed, and we calculate

this likelihood as the two-sided area under the Gaussian curve with SD 1/3 of the relative mass error[1]. For the background model, we cannot use the mass of the peak since, in general, noise peaks may appear at any mass. But we can use the peak intensity for this purpose: Evaluations have shown that noise peak intensities are roughly exponentially distributed; see for instance Fig. 4 in [2]. Let $\lambda e^{\lambda x}$ be the exponential distribution with parameter λ, where *x* is the peak intensity. The likelihood of observing a noise peak with intensity *y* or higher is $\mathrm{P}(\text{intensity} \geq y) = \int_y^\infty \lambda e^{\lambda x} \mathrm{d}x = e^{-\lambda y}$. Taking the natural logarithm, we reach -λ*y* for intensity *y*. Since this likelihood appears in the denominator of the log odds term, we simply add the peak intensity, multiplied by a constant representing the noise in the spectrum, to the score. Finally, we can use prior probabilities, computing the odds ratio that any peak is not noise: We add a constant b, being the logarithm of this odds ratio, to each vertex score.

Weighting edges is more complex: We will consider common neutral losses, unlikely neutral losses containing only one atom type, the mass of the loss, collision energies, and the ratio between carbon and hetero atoms. Our scoring has been adapted from [3]; see there for more details. There are certain NLs that appear often when analyzing organic and biological compounds. We have created a short list of these common neutral losses; see Table 2. We reward the occurrence of a combination of up to three losses from the list by adding log(γ/n); $\gamma > 1$ to the score, where γ is a parameter that has to be chosen individually for each dataset, and n is the number of combined common losses. Unlike [4], we divide by the number of combined losses. Combinations may represent groups detaching together or the loss of an intermediate peak, but these cases are not as strongly rewarded. We penalize losses consisting purely of carbon or purely of nitrogen with $\log(\varepsilon), \varepsilon \ll 1$, as these are unlikely neutral losses. To avoid star-like fragmentation trees where all fragments branch from the root, we penalize large NLs by $\log(1 - \frac{\text{mass neutral loss}}{\text{parent mass}})$. As we have measured tandem spectra at distinct collision energies, we can exploit the fact that fragments often occur at several fragmentation energies. This results in two more parameters α, β; we omit the details[4]. In organic compounds, there is typically a carbon backbone complemented by some hetero atoms (e.g., nitrogen and oxygen). Here, the hetero atom-to-carbon ratio is an indicator for a molecular formula to be true[4]; again, we omit the details[4].

For our analysis, we use the following parameters: For all datasets set $\lambda = 0.1$ and $\varepsilon = 10^{-4}$. Parameters γ, b were chosen to capture instrument-specific properties: For example, the QSTAR instrument produces relatively few fragment peaks, but these often reflect typical losses. For the Orbitrap data, we use $\gamma = 10$ and $b = 5$; for the Micromass QTOF data $\gamma = 10$

and b = 0; and for the API QSTAR data $\gamma = 1000$ and b = 0. For parameters α, β, defaults $\alpha = 0.1$ and $\beta = 0.8$ are used. No parameter optimization was carried out. See Fig. 2 for an example of a fragmentation graph.

Algorithm for fragmentation tree computation Different fragmentation pathways may lead to fragments with identical molecular formulas or even identical structures. This is quite easy to see but, unfortunately, makes it practically impossible to formulate our task as an optimization problem: a small fragment may be generated from almost all other fragments, but we only want to record the most likely explanation. Hence, we slightly oversimplify the problem: We demand that each fragment in the fragmentation spectrum is generated by a single fragmentation pathway. That means that any fragment may have at most one "parent fragment" from which it is generated. For our fragmentation graph, this means that we are searching for a tree inside this graph, denoted a fragmentation tree. This allows us to simplify our problem: For every vertex in the fragmentation tree except for the root corresponding to the unfragmented compound, we select exactly one incoming edge. Hence, we can pull up the weight of each vertex into the incoming edges and assume that the fragmentation graph is edge-weighted.

Similarly, several fragments may result in a single peak in the fragmentation spectrum. We argue that this is an extremely rare event; again, it interferes with our optimization formulation: Adding up scores, we have to make sure that each peak can contribute at most once. To this end, we demand that our fragmentation tree is colorful: Each vertex color and, hence, each peak in the fragmentation spectrum is scored at most once.

Now, we search for a colorful tree inside the fragmentation graph that has the maximal sum of edge weights. This is a NP-hard problem: there cannot exist an algorithm with running time polynomial in the input size unless P = NP. Several heuristics have been proposed for this problem[11], but it turns out that fragmentation trees computed by heuristics are usually of very low quality (see Fig. S-4 and the results section). Thus an efficient exact algorithm for the problem is required. We here suggest an algorithm that follows the paradigm of fixed-parameter tractability[5].

We use dynamic programming over the vertices to find the maximum colorful subtree in the graph. We encode the colors used so far as part of the dynamic programming matrix[6]. Optimal solutions can be computed by combining optimal solutions of sub-problems. Let C represent all the set of all colors, $c(v)$ the color of vertex v and $w(u; v)$ the score of edge uv. We set $W(v; S)$ as the maximal score of a colorful subtree with root v using colors $S \subseteq C$. We set $W(v; S)$ to the maximum of $\max_{u \in V, c(u) \in S \setminus \{c(v)\}} W(u, S \setminus \{c(v)\}) + w(v,u)$ and

$\max_{S_1 \cap S_2 = \{c(v)\}, S_1 \cup S_2 = S} W(v, S_1) + W(v, S_2)$. We initialize $W(v,\{c(v)\})$=0. Non-existent edges are assumed to have weight -∞. The first line extends a tree by introducing a new root; the second line merges two trees that have the same root. Note that the value of $W(v; S)$ is undefined if no corresponding tree exists.

Further MS^n evaluation

For (-)-epicatechin in Fig. S-2, MS^3 of *m/z* 291→273 and 291→165 transitions were recorded. From the 291→273 transition, it was apparent that *m/z* 165, 151, 147, and 123 are on the fragmentation pathway but not *m/z* 139. For the 291→165 transition, *m/z* 147 was observed but again not *m/z* 139. This ion is directly formed from the $[M+H]^+$ precursor, apparently by a retro Diels-Alder (rDA) reaction. The loss of CH_2 from *m/z* 165, assigned as unlikely, was fully excluded. The mechanism of noted ions formation was evaluated and typically consists of rDA, rearrangements, and hydrogen transfer steps. From the calculated tree, the backbone nodes (291-273-165-147) were fully supported by the MS^n spectra; other nodes (151, 139, 123) are formed directly from precursor ions or the one (*m/z* 273) formed after water NLs; see Fig. S-3. For the more complex tree of chelidonine in Fig. 3, MS^3 data also strongly supported the calculated fragmentation tree; see Fig. S-10 and S-11. The main backbone pathway (354-323-295-293-275-247) was fully supported with one exception. The edge connecting nodes 295-293 is incorrect (due to the loss of molecular hydrogen), as *m/z* 293 is formed from *m/z* 323 by the loss of formaldehyde, and nodes 323 and 293 are directly connected. Node 295 remains in the tree but forms a new branch (323-295, loss of CO). The third generation 305 node can be formed both from nodes 323 and 326. This connection is not visible in the calculated tree as this would violate the tree property.

Evaluation against Mass Frontier.

For our second evaluation, we compare the molecular formulas our method assigns to the peaks, with the predictions of the Mass Frontier software. Here, we use the Micromass QTOF dataset, where predictions have previously been carried out in a different experimental context[1]. Version 4 of Mass Frontier was used in protonated ion mode with "rules" fragmentation mechanism and a reaction number of 5. Given the molecular structure of the compound, Mass Frontier predicts tandem mass spectra, which we match to the observed data. Regarding the accuracy of the method, we annotate more than four times as many peaks as Mass Frontier (70.3% vs. 16.8%). Only 19 peaks were annotated by Mass Frontier but not

by our software. For the 1072 peaks that both tools annotate, the same molecular formula is assigned in 97.3% of the cases. This is an excellent agreement, taking into account the completely different paradigms of the two tools: Mass Frontier knows the molecular structure but not the experimental MS data, whereas our tool knows the experimental MS data but not the molecular structure.

The probability that such an agreement can happen by chance (significance) is below 10^{-167}. Because Mass Frontier tends to annotate peaks of small mass, only few molecular formulas are within the mass accuracy. To this end, we discarded all matched peaks with only one possible annotation, keeping 444 peaks with 3.9 explanations on average. For these peaks, we reach a match with Mass Frontier in 93.7% of the cases (significance as above). To assess this agreement, we compared Mass Frontier predictions against two other predictors: A random peak annotator that selects an arbitrary molecular formula within the mass accuracy, reaches only 35.6% agreement with Mass Frontier (significance 0.51). The naive approach, which always uses the molecular formula with the smallest mass difference to each peak, would reach 71.8% agreement (significance 10^{-61}). Clearly, agreement between Mass Frontier and our approach is much higher. See Table S-3.

References

1 Böcker S., Letzel M., Lipták Zs., Pervukhin A. (2009) SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics* 25:218–224.

2 Goldberg, D., Bern, M., Li, B. & Lebrilla, C. B. Automatic determination of O-glycan structure from fragmentation spectra. *J. Proteome Res.* 5, 1429–1434 (2006).

3 Böcker S., Rasche F. (2008) Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* 24:I49–I55. Proc. of European Conference on Computational Biology (ECCB 2008).

4 Kind T., Fiehn O. (2007) Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinf.* 8:105.

5 Niedermeier R. Invitation to Fixed-Parameter Algorithms. Oxford University Press (Oxford), 2006

6 Dreyfus S.E., Wagner R.A. (1972) The Steiner problem in graphs. *Networks* 1:195-207.

Supplementary Tables

Compound name	m/z[a]	molecular formula[b]	measured peaks[c]	rank using isotope pattern[d]	rank using fragmentation pattern[d]	rank combined[d]	annotated NL[e]	evaluation by experts correct[f]	unclear[f]	wrong[f]	note
Adenosine	268.092	C10H13N5O4	4	1	1	1	1	1	0	0	
Anisic acid	153.055	C8H8O3	2	1	1	1	1	1	0	0	
Apomorphine	268.134	C17H17NO2	10	1	2	1	6	5	1	0	
Armentoflavone	539.098	C30H18O10	18	1	19	1	13	11	2	0	
Berberine	335.116[g]	C20H17NO4	11	1	4	1	3	3	0	0	radical loss, pull-up
Bergapten	217.050	C12H8O4	73	1	1	1	12	8	3	1	
Bicuculline	368.113	C20H17NO6	55	1	4	1	34	22	9	3	
Biochanin A	285.076	C16H12O5	74	2	2	1	36	27	3	6	
Chelidonine	354.134	C20H19NO5	69	1	2	1	19	15	4	0	radical loss, pull-up
Cinchonine	295.181	C19H22N2O	29	3	1	1	23	18	4	1	
Emetine	481.307	C29H40N2O4	62	1	5	1	36	31	2	3	
(-)-Epicatechine	291.087	C15H14O6	11	2	1	1	6	5	1	0	pull-up
Erythromycin	734.469	C37H67NO13	2	2	18	1	1	1	0	0	
Genistein	271.061	C15H10O5	36	1	1	1	31	23	2	6	radical loss, pull-up
Harmane	183.092	C12H10N2	4	1	1	1	2	2	0	0	
IAA-Val	275.140	C15H18N2O3	6	1	1	1	3	2	0	1	pull-up
Indol-3-carboxylic acid	162.056	C9H7NO2	3	1	1	1	2	2	0	0	
Kaempherol	287.056	C15H10O6	47	1	1	1	38	24	5	9	
Kinetin	216.089	C10H9N5O	8	1	2	1	7	5	0	2	pull-up
Laudanosin	358.202	C21H27NO4	7	1	1	1	6	6	0	0	pull-up
Methylumbelliferrylglucoronide	353.087	C16H16O9	4	1	1	1	1	1	0	0	
(S,R)-Noscapine	414.155	C22H23NO7	1	2	3	1	7	5	2	0	radical loss
Phenylalanine	166.087	C9H11NO2	5	1	1	1	3	3	0	0	
Phlorizin	437.145	C21H24O10	10	1	3	1	7	7	0	0	
Quercetin	303.050	C15H10O7	45	1	1	1	36	26	3	7	
Reserpine	609.281	C33H40N2O9	31	1	7	1	19	13	6	0	pull-up
Resveratrol	229.086	C14H12O3	20	1	1	1	19	15	0	4	
Rotenone	395.149	C23H22O6	83	2	5	1	45	34	8	3	
Rutine	611.161	C27H30O16	3	1	30	1	2	2	0	0	
Safranin	315.161	C20H18N4	22	1	3	1	7	5	0	2	
Salsolinol	180.102	C10H13NO2	5	1	1	1	4	4	0	0	
Sinapine	310.166[g]	C16H24NO5	5	1	1	1	1	1	0	0	
Tetrahydropapaveroline	288.124	C16H17NO4	8	1	2	1	6	6	0	0	pull-up
3,4,5-Trimethoxycinnamic acid	239.092	C12H14O5	9	1	1	1	5	4	1	0	
Tryptophan	205.098	C11H12N2O2	2	1	2	1	1	1	0	0	
Vitexin-2-O-rhamnoside	579.171	C27H30O14	15	1	58	1	12	11	1	0	
Xanthohumol	355.155	C21H22O5	8	2	1	1	3	2	0	1	pull-up
							458	352	57	49	

Table S-1: Results for the Orbitrap dataset, expert evaluation: Compound, [a]m/z value for [M+H]+ adduct precursor; ion formed by ESI and analyzed in Orbitrap using 30 000 resolution; [b]molecular formula of the compounds; [c]number of peaks in the merged spectra; [d]rank of molecular formula identification using isotope patterns, using fragmentation patterns, and combined identification; [e]number of annotated NLs (edges) in hypothetical fragmentation trees, [f]number of NLs marked "correct", "unclear", or "wrong" by an MS expert. [g]Value for M+, as quaternary nitrogen in the compound. "Radical loss" denotes that MS experts have identified a radical loss in the MS data not annotated by the program, and "pull-up" indicates that NLs may be inserted too deep in the fragmentation tree.

Compound name	m/z[a]	molecular formula[b]	collision energies	measured peaks[c]	rank using isotope pattern[d]	rank using fragmentation pattern[d]	rank combined[d]	annotated NL[e]	evaluation by experts correct[f]	unclear[f]	wrong[f]	note
3-(4-Hexosyloxyphenyl)propanoyl choline	414.214[g]	C20H32NO8	25, 40, 55	5	1	1	1	4	4	0	0	
4-Coumaroyl choline	250.145[g]	C14H20NO3	15, 25, 40	5	1	1	1	4	4	0	0	
4-Hexosylferuloyl choline	442.209[g]	C21H32NO9	15, 25, 40, 55	7	1	3	1	4	4	0	0	
4-Hexosyloxybenzoyl choline	386.182[g]	C18H28NO8	15, 25, 40, 55, 90	7	4	1	1	5	5	0	0	
4-Hexosyloxycinnamoyl choline	412.198[g]	C20H30NO8	25, 40, 55	6	3	2	1	4	4	0	0	
4-Hexosylvanilloyl choline	416.193[g]	C19H30NO9	15, 25, 40, 55, 70	5	7	3	1	3	3	0	0	
4-Hydroxybenzoyl choline	224.130[g]	C12H18NO3	15, 25, 40, 55	5	1	1	1	4	4	0	0	
5-Hydroxyferuloyl choline	296.151[g]	C15H22NO5	15, 25, 40, 55	13	2	5	1	11	8	0	3	radical loss
6-Aminocapronic acid	132.103	C6H13NO2	15, 20, 30, 40	29	1	1	1	17	13	4	0	
Acetyl choline	146.119[g]	C7H16NO2	10, 20,30	4	1	1	1	3	3	0	0	
Alanine	90.056	C3H7NO2	10, 15, 20	2	1	1	1	1	1	0	0	
Arginine	175.120	C6H14N4O2	20, 25, 30	17	1	1	1	14	13	1	0	
Asparagine	133.062	C4H8N2O3	10, 15, 20, 30, 40	26	1	1	1	20	14	3	3	
Aspartic acid	134.046	C4H7NO4	10, 15, 20, 30	13	3	1	1	7	7	0	0	
Benzoyl choline	207.126[g]	C12H18NO2	15, 25, 40, 55	4	1	1	1	3	3	0	0	
Cafeoyl choline	266.140[g]	C14H20NO4	15, 25, 40, 55	10	1	1	1	8	7	0	1	
Cinnamoyl choline	234.150[g]	C14H20NO2	15, 25, 40, 55	4	1	1	1	3	3	0	0	
Citrulline	176.104	C6H13N3O3	10, 15, 20, 25, 30	25	1	1	1	11	11	0	0	
Cysteine	122.028	C3H8NO2S	10, 15, 20, 30	10	1	1	1	6	6	0	0	
Cystine	241.033	C6H12N2O4S2	10, 15, 20, 30, 40	55	1	1	1	16	8	8	0	
Dopamine	154.088	C8H11NO2	10, 20, 30, 40, 50	19	1	1	1	14	10	4	0	
Feruloyl choline	280.156[g]	C15H22NO4	15, 25, 40	9	1	3	1	5	5	0	0	
Glutamic acid	148.062	C5H9NO4	10, 15, 20, 30	8	2	1	1	4	4	0	0	
Glutamine	147.078	C5H10N2O3	10, 15, 20, 30	10	1	1	1	8	8	0	0	
Histidine	156.078	C6H9N3O2	15, 25, 35, 45	17	1	1	1	13	12	0	1	
Isoleucine	132.103	C6H13NO2	10, 15, 25, 40,	18	1	1	1	9	7	2	0	
Leucine	132.103	C6H13NO2	15, 25, 40	19	1	1	1	10	8	2	0	
Methionine	150.060	C5H11NO2S	10, 15, 20, 30	13	1	1	1	10	9	1	0	
Nicotinic acid choline ester	209.130[g]	C11H17N2O2	15, 25, 40, 55	4	1	1	1	3	3	0	0	
Phenylalanine	166.088	C9H11NO2	15, 25, 40	15	1	1	1	12	8	4	0	
Proline	116.072	C5H9NO2	10, 15, 55	9	1	1	1	5	5	0	0	
Serine	106.051	C3H7NO3	10, 15, 20, 30	7	1	1	1	4	4	0	0	
Sinapoyl choline	310.166[g]	C16H24NO5	15, 25, 40	6	2	1	1	5	5	0	0	
Spermidine	146.167	C7H19N3	15, 25, 35, 45	21	1	1	1	13	10	3	0	
Spermine	203.224	C10H26N4	15, 25, 35, 45	13	1	1	1	12	7	4	1	pull-up
Syringoyl choline	284.151[g]	C14H22NO5	15, 25, 40, 55	17	2	1	1	9	6	3	0	
Threonine	120.067	C4H9NO3	10, 15, 20, 30	9	1	1	1	5	5	0	0	
Tryptophane	205.099	C11H12N2O2	15, 25, 40, 55	33	1	1	1	22	15	4	3	radical loss
Tyramine	138.093	C8H12NO	15, 20, 30, 40, 50	21	1	1	1	10	7	3	0	
Tyrosine	182.083	C9H11NO3	10, 15, 25, 30, 40	25	1	1	1	13	11	1	1	
Valine	118.088	C5H11NO2	10, 25, 40, 55	15	1	1	1	10	7	3	0	
Vanilloyl choline	254.140[g]	C13H20NO4	15, 25, 40, 55	10	1	1	1	6	5	1	0	
								350	286	51	13	

Table S-2: Results for the QSTAR dataset, expert evaluation: Compound, [a]m/z value for [M+H]+ adduct precursor; [b]molecular formula of the compounds; [c]number of peaks in the merged spectra; [d]rank of molecular formula identification using isotope patterns, using fragmentation patterns, and combined identification; [e]number of annotated NLs (edges) in hypothetical fragmentation trees, [f]number of NLs marked "correct", "unclear", or "wrong" by an MS expert. [g]Value for M+, as quaternary nitrogen in the compound. "Radical loss" denotes that MS experts have identified a radical loss in the MS data not annotated by the program, and "pull-up" indicates that NLs may be inserted too deep in the fragmentation tree.

Compound	PubChem ID	molecular formula	monoisotopic mass	measured peaks	Mass Frontier prediction sensitivity	Mass Frontier prediction specificity	Mass Frontier prediction F-value	our method sensitivity	common peaks	non-trivial common peaks	non-matching explanations
6a-Methylprednisolone	4159	$C_{22}H_{30}O_5$	374,209	192	0,135	0,268	0,180	0,479	17	0	0
Acepromazine	6077	$C_{19}H_{22}N_2OS$	326,145	44	0,182	0,421	0,254	0,591	8	4	0
Acetophenazine	441185	$C_{23}H_{29}N_3O_2S$	411,198	47	0,404	0,250	0,309	0,660	18	14	1
Adenosine Diphosphate	197	$C_{10}H_{15}N_5O_{10}P_2$	427,029	16	0,313	0,238	0,270	0,750	5	5	1
Adiphenine	2031	$C_{20}H_{25}NO_2$	311,189	15	0,333	0,076	0,123	0,733	5	0	0
Albuterol	2083	$C_{13}H_{21}NO_3$	239,152	45	0,133	0,133	0,133	0,644	6	1	0
Alfentanil	51263	$C_{21}H_{32}N_6O_3$	416,254	60	0,350	0,119	0,178	0,783	19	11	0
Amfenac	2136	$C_{15}H_{11}NO_3$	255,090	59	0,119	0,350	0,177	0,695	7	0	0
Aminophylline	2153	$C_7H_8N_4O_2$	180,065	36	0,139	0,357	0,200	0,556	5	3	0
Ampicillin	2174	$C_{16}H_{19}N_3O_4S$	349,110	58	0,328	0,066	0,110	0,845	18	14	3
Anileridine	8944	$C_{22}H_{28}N_2O_2$	352,215	16	0,188	0,046	0,074	0,813	3	0	0
Antipyrine	2206	$C_{11}H_{12}N_2O$	188,095	71	0,085	0,500	0,145	0,592	6	0	0
Antipyrine-4-amino	2151	$C_{11}H_{13}N_3O$	203,106	57	0,105	0,207	0,140	0,702	6	0	0
Apomorphine	2215	$C_{17}H_{17}NO_2$	267,126	16	0,063	0,077	0,069	0,438	1	0	0
Apramycin	71428	$C_{21}H_{41}N_5O_{11}$	539,280	105	0,410	0,139	0,207	0,857	42	40	4
Betaxolol	2369	$C_{18}H_{29}NO_3$	307,215	95	0,400	0,380	0,390	0,674	32	4	0
Boldenone Undecylenate	25702	$C_{30}H_{44}O_3$	452,329	45	0,311	0,132	0,185	0,867	11	0	0
Bumetanide	2471	$C_{17}H_{20}N_2O_5S$	364,109	73	0,068	0,135	0,091	0,808	4	1	0
Buprenorphine	2476	$C_{29}H_{41}NO_4$	467,304	241	0,012	0,030	0,018	0,598	3	1	0
Buspirone	2477	$C_{21}H_{31}N_5O_2$	385,248	39	0,436	0,142	0,214	0,846	14	11	0
Cholesterol	304	$C_{27}H_{46}O$	386,355	25	0,120	0,088	0,102	0,480	2	0	0
Cromolyn	2882	$C_{23}H_{16}O_{11}$	468,069	51	0,333	0,293	0,312	0,824	17	1	0
Cymarin	539061	$C_{30}H_{44}O_9$	548,299	114	0,219	0,116	0,152	0,649	16	5	0
Daunorubicin	2958	$C_{27}H_{29}NO_{10}$	527,179	35	0,200	0,035	0,060	0,943	7	3	0
Dextromethorphan	3008	$C_{18}H_{25}NO$	271,194	62	0,097	0,222	0,135	0,645	6	0	0
Dihydroergotamine	3066	$C_{33}H_{37}N_5O_5$	583,279	51	0,216	0,039	0,066	0,922	11	8	0
Dimefline	3078	$C_{20}H_{21}NO_3$	323,152	16	0,188	0,136	0,158	0,625	1	0	0
Diphenoxylate	13505	$C_{32}H_{32}N_2O_3$	452,246	91	0,176	0,229	0,199	0,593	16	1	0
Dobutamine	36811	$C_{18}H_{23}NO_3$	301,168	16	0,500	0,178	0,262	0,938	8	1	0
Doxorubicin	1691	$C_{27}H_{29}NO_{11}$	543,174	72	0,208	0,068	0,103	0,972	15	7	0
Drofenine	3166	$C_{20}H_{31}NO_3$	317,235	19	0,474	0,143	0,220	0,947	9	0	0
Enalapril	3222	$C_{20}H_{28}N_2O_5$	376,200	22	0,636	0,046	0,085	0,909	14	5	0
Enalaprilat	5362033	$C_{18}H_{24}N_2O_5$	348,169	21	0,619	0,053	0,098	0,952	13	4	0
Ephedrine	5032	$C_{10}H_{15}NO$	165,115	30	0,267	0,348	0,302	0,700	8	0	0
Ergocristine	98255	$C_{35}H_{39}N_5O_5$	609,295	50	0,340	0,059	0,101	0,960	17	15	4
Ergoloid Mesylate	592735	$C_{33}H_{41}N_5O_5$	591,342	16	0,250	0,011	0,022	0,938	4	3	1
Etamiphylline	28329	$C_{17}H_{21}N_5O_2$	279,170	62	0,194	0,203	0,198	0,726	12	3	0
Etodolac	3308	$C_{17}H_{21}NO_3$	287,152	66	0,197	0,151	0,171	0,773	13	7	0
Fenbendazole	3334	$C_{15}H_{13}N_3O_2S$	299,073	38	0,053	0,222	0,085	0,632	2	0	0
Fenoterol	3343	$C_{17}H_{21}NO_4$	303,147	15	0,467	0,117	0,187	0,867	7	1	0
Folic Acid	3405	$C_{19}H_{19}N_7O_6$	441,140	19	0,368	0,040	0,073	1,000	7	7	1
Gallamine	3450	$C_{30}H_{60}N_2O_3$	510,463	24	0,167	0,060	0,088	0,625	4	3	0
Gingerol	3473	$C_{17}H_{26}O_4$	294,183	34	0,265	0,180	0,214	0,794	9	0	0
Hematoporphyrin I	11103	$C_{34}H_{38}N_4O_6$	598,279	79	0,038	0,056	0,045	0,949	3	2	0
Hydrocortisone	3640	$C_{21}H_{30}O_5$	362,209	174	0,161	0,246	0,194	0,477	20	0	0
Hydroxybutorphanol	3064246	$C_{21}H_{29}NO_3$	343,215	101	0,198	0,190	0,194	0,743	19	1	0
Hydroxyphenethylamine	5610	$C_8H_{11}NO$	137,084	26	0,077	0,400	0,129	0,615	2	0	0
Isoxsuprine	3783	$C_{18}H_{23}NO_3$	301,168	51	0,373	0,279	0,319	0,706	18	1	0
Ketorolac	3826	$C_{15}H_{13}NO_3$	255,090	18	0,278	0,125	0,172	0,722	4	0	0
Leucine Enkephalin	3903	$C_{28}H_{37}N_5O_7$	555,269	53	0,811	0,088	0,159	0,943	39	36	0
Mebeverine	4031	$C_{25}H_{35}NO_5$	429,252	12	0,500	0,052	0,094	0,833	6	2	0
Mefenamic Acid	4044	$C_{15}H_{15}NO_2$	241,110	28	0,036	0,071	0,048	0,643	1	0	0
Meprobamate	4064	$C_9H_{18}N_2O_4$	218,127	13	0,154	0,250	0,190	1,000	2	0	0
Methionine Enkephalin	42785	$C_{27}H_{35}N_5O_7S$	573,226	62	0,710	0,085	0,153	0,968	44	42	5
Methotrexate	4112	$C_{20}H_{22}N_8O_5$	454,171	15	0,200	0,024	0,000	0,800	3	3	3
Methylergonovine	4140	$C_{20}H_{25}N_3O_2$	339,195	53	0,302	0,131	0,183	0,698	16	0	0
Morphine-3-Glucuronide	4318740	$C_{23}H_{27}NO_9$	461,169	56	0,071	0,033	0,045	0,732	4	4	0
Naltrexone	4428	$C_{20}H_{23}NO_4$	341,163	138	0,087	0,128	0,103	0,587	12	1	0
Nandrolone	9904	$C_{18}H_{26}O_2$	274,193	80	0,225	0,419	0,293	0,650	13	0	0
Nimesulide	4495	$C_{13}H_{12}N_2O_5S$	308,047	42	0,000	0,000	0,000	0,714	0	0	0
Norpropoxyphene	18804	$C_{21}H_{25}NO_2$	325,204	10	0,500	0,051	0,093	0,600	4	0	0
Noscapine	4544	$C_{22}H_{23}NO_7$	413,147	165	0,055	0,155	0,081	0,600	7	3	0
Ormetoprim	23418	$C_{14}H_{18}N_4O_2$	274,143	94	0,011	0,125	0,020	0,660	1	1	0
Oxaprozin	4614	$C_{18}H_{15}NO_3$	293,105	23	0,087	0,095	0,091	0,609	2	0	0
Oxybutynin	4634	$C_{22}H_{31}NO_3$	357,230	64	0,328	0,206	0,253	0,859	20	5	0
Oxycodone	4635	$C_{18}H_{21}NO_4$	315,147	146	0,068	0,169	0,098	0,541	9	1	0
Oxytetracycline	5280972	$C_{22}H_{24}N_2O_9$	460,148	152	0,092	0,125	0,106	0,914	13	7	3
Perindopril	107807	$C_{19}H_{32}N_2O_5$	368,231	17	0,706	0,042	0,079	0,882	11	2	0
Piperacetazine	19675	$C_{24}H_{29}N_3O_2S$	410,203	22	0,409	0,184	0,254	0,909	9	8	0
Poldine	11018	$C_{22}H_{25}NO_3$	340,191	34	0,118	0,125	0,121	0,471	4	0	0
Prazosin	4893	$C_{19}H_{21}N_5O_4$	383,159	71	0,169	0,375	0,233	0,915	12	12	0
Prednisolone	4894	$C_{21}H_{28}O_5$	360,194	172	0,140	0,253	0,180	0,483	17	0	0
Prednisolone Tebutate	4898	$C_{27}H_{36}O_6$	458,267	161	0,106	0,106	0,106	0,516	12	0	0
Prednisone	4900	$C_{21}H_{26}O_5$	358,178	194	0,124	0,235	0,162	0,500	20	0	0
Prolintane	14592	$C_{13}H_{29}N$	217,183	8	0,500	0,129	0,205	0,875	4	0	0
Pyrilamine	4992	$C_{17}H_{23}N_3O$	285,184	11	0,182	0,105	0,133	0,909	2	1	0
Remifentanil	60815	$C_{20}H_{28}N_2O_5$	376,200	55	0,400	0,125	0,190	0,891	22	6	0
Reserpine	5052	$C_{33}H_{40}N_2O_9$	608,273	122	0,164	0,096	0,121	0,877	19	19	0
Rolitetracycline	6420073	$C_{27}H_{33}N_3O_8$	527,227	17	0,294	0,029	0,053	1,000	5	5	0
Salmeterol	5152	$C_{25}H_{37}NO_4$	415,272	71	0,282	0,171	0,213	0,789	19	2	0
Spectinomycin	2021	$C_{14}H_{24}N_2O_7$	332,158	122	0,393	0,251	0,307	0,672	33	18	0
Streptomycin	19649	$C_{21}H_{39}N_7O_{12}$	581,266	147	0,184	0,088	0,119	0,755	25	24	0
Strychnine	5304	$C_{21}H_{22}N_2O_2$	334,168	148	0,014	0,051	0,021	0,520	2	0	0
Strychnine N-oxide	73393	$C_{21}H_{22}N_2O_3$	350,163	181	0,011	0,069	0,019	0,630	2	0	0
Sufentanil	41693	$C_{22}H_{30}N_2O_2S$	386,203	34	0,441	0,097	0,160	0,882	14	5	0
Sulfadimethoxine	5323	$C_{12}H_{14}N_4O_4S$	310,074	54	0,037	0,333	0,067	0,778	1	1	0
Sulfasalazine	5384001	$C_{18}H_{14}N_4O_5S$	398,068	76	0,053	0,364	0,092	0,908	4	3	1
Taurocholate	8959	$C_{26}H_{45}NO_7S$	515,292	134	0,060	0,052	0,055	0,806	7	3	0
Tenoxicam	5282194	$C_{13}H_{11}N_3O_4S_2$	337,019	30	0,233	0,318	0,269	0,900	7	7	0
Terbutaline	5403	$C_{12}H_{19}NO_3$	225,136	35	0,229	0,242	0,235	0,771	6	1	0
Terfenadine	5405	$C_{32}H_{41}NO_2$	471,314	101	0,129	0,171	0,147	0,653	13	0	0
Testosterone Propionate	5701990	$C_{22}H_{32}O_3$	344,235	69	0,232	0,213	0,222	0,826	14	0	0
Tetracaine	5411	$C_{15}H_{24}N_2O_2$	264,184	30	0,267	0,136	0,180	0,667	8	0	0
Tetracycline	5353990	$C_{22}H_{24}N_2O_8$	444,153	149	0,060	0,074	0,066	0,799	8	5	1
Tetramisole	3913	$C_{11}H_{12}N_2S$	204,072	42	0,310	0,481	0,377	0,690	11	5	0
Theobromine	5429	$C_7H_8N_4O_2$	180,065	45	0,111	0,333	0,167	0,556	3	1	0
Thiethylperazine	5440	$C_{22}H_{29}N_3S_2$	399,180	33	0,212	0,175	0,192	0,576	8	7	0
Thioridazine	5452	$C_{21}H_{26}N_2S_2$	370,154	25	0,280	0,241	0,259	0,600	7	1	0
Thiothixene	941651	$C_{23}H_{30}N_4O_2S_2$	443,170	89	0,213	0,613	0,317	0,764	18	10	0
Thonzide	5456	$C_{25}H_{19}N_3O$	511,438	6	0,333	0,080	0,129	0,667	2	1	0
Tripelennamine	5587	$C_{16}H_{21}N_3$	255,174	11	0,364	0,222	0,276	0,727	4	0	0
Vecuronium	39765	$C_{34}H_{57}N_2O_4$	557,432	58	0,155	0,040	0,064	0,966	9	5	0

Table S-3: Results for the Micromass QTOF dataset, evaluation against Mass Frontier: Compound, PubChem ID, molecular formula, mass, and number of measured peaks; sensitivity, specificity, and F-value of the Mass Frontier prediction; sensitivity of our prediction; number of common peaks, number of non-trivial common peaks, and number of non-matching explanations. Zero non-matching explanations indicate perfect agreement.

Supplementary Figures

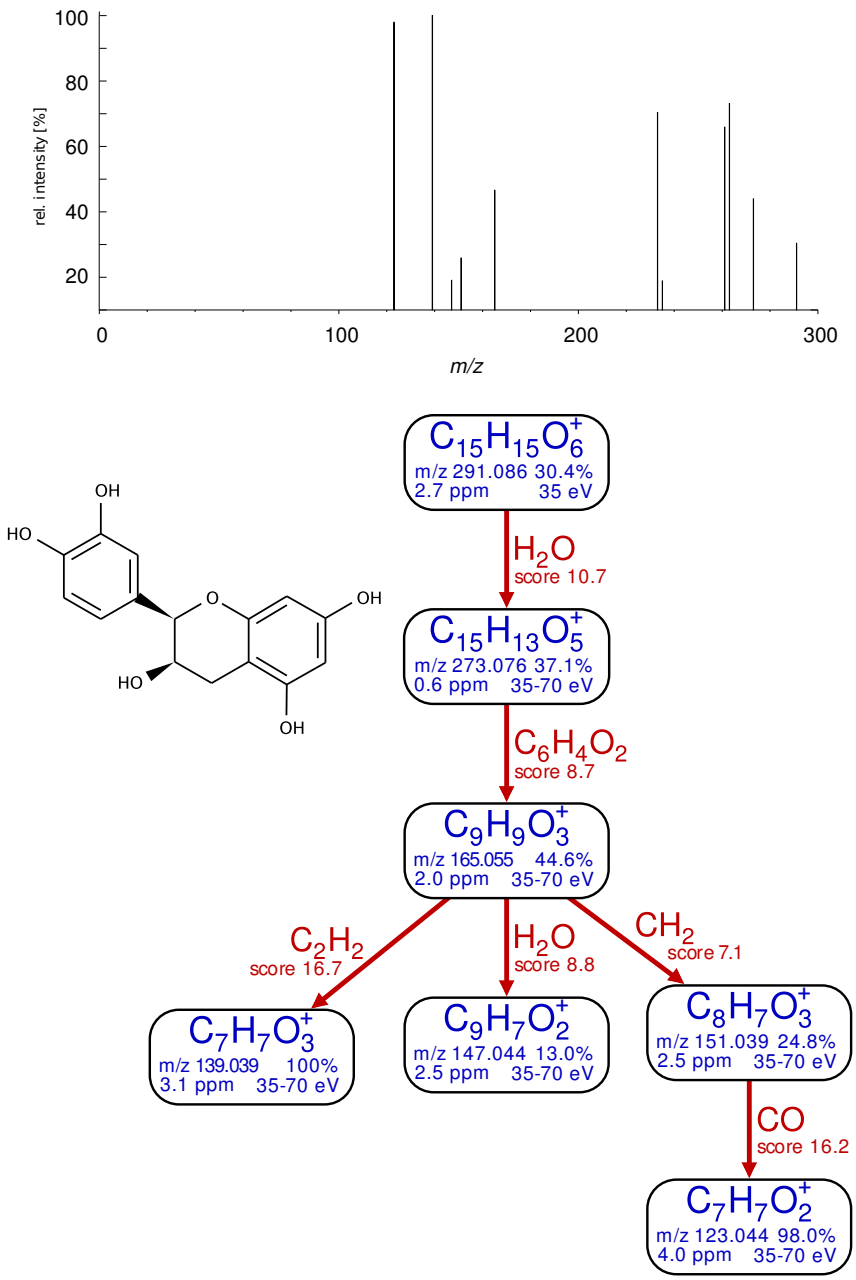

Figure S-1: Hypothetical fragmentation tree of (-)-epicatechine ($C_{15}H_{14}O_6$) computed by our method using Orbitrap data. Nodes (blue) correspond to peaks in the tandem mass spectra and their annotate molecular formula (CE is the range of collision energies), arcs (red) correspond to hypothetical neutral losses.

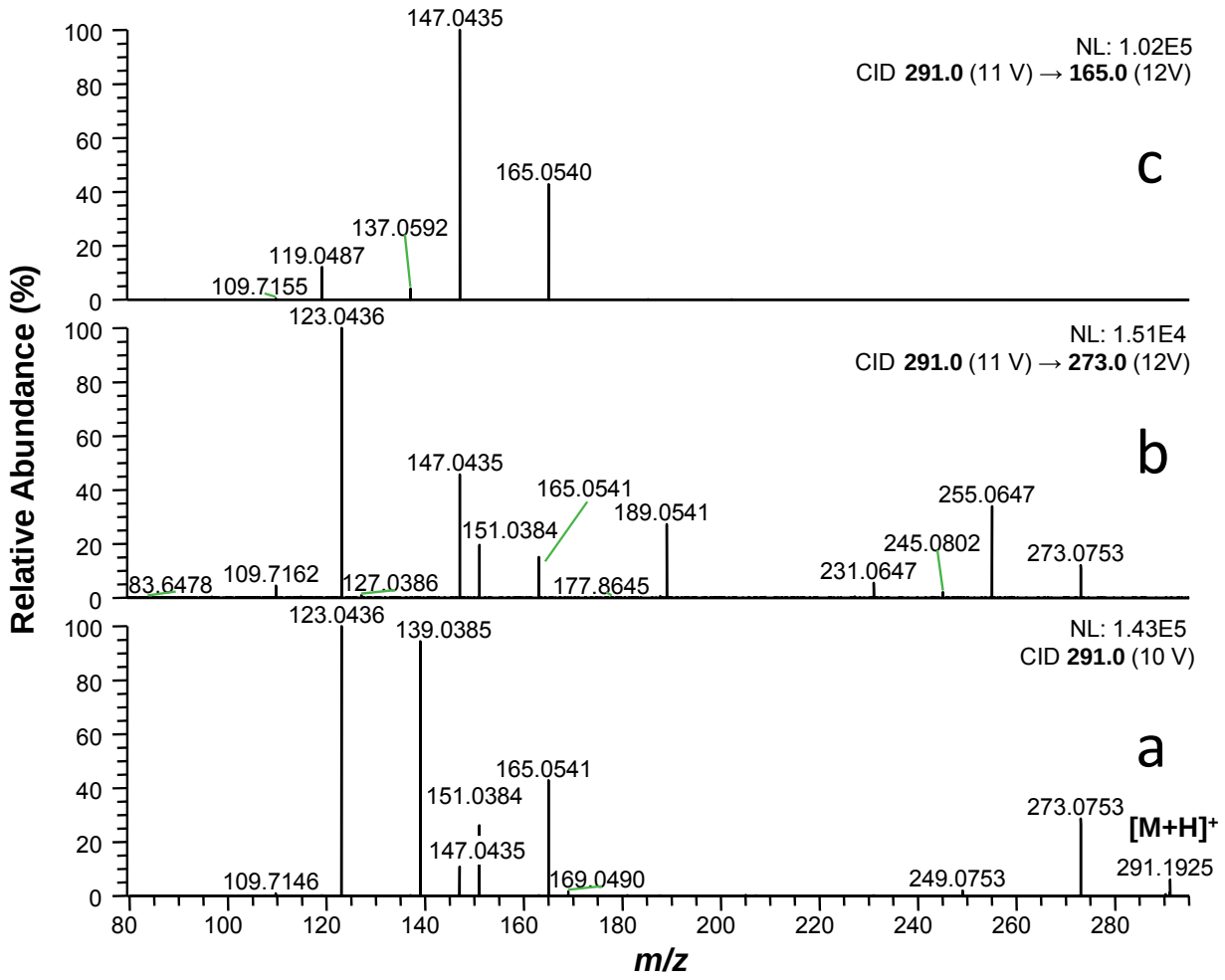

Figure S-2: A series of tandem mass spectrometry experiments performed with protonated (-)-epicatechine generated by electrospray and analyzed with an Orbitrap XL instrument. Fragmentation was realized in linear traps using He as collision gas. (a) CID MS^2 spectra generated from molecular adduct ion [M+H]+ using 10 V in linear trap (other used CID voltages given in brackets). (b,c) MS^3 tandem mass spectra; transitions are given in inserts in bold, used collision energies are indicated in brackets. Intensities (NLs, instrument internal units) are given for all experiments, but spectra are plotted on relative scale (%) for better comparison.

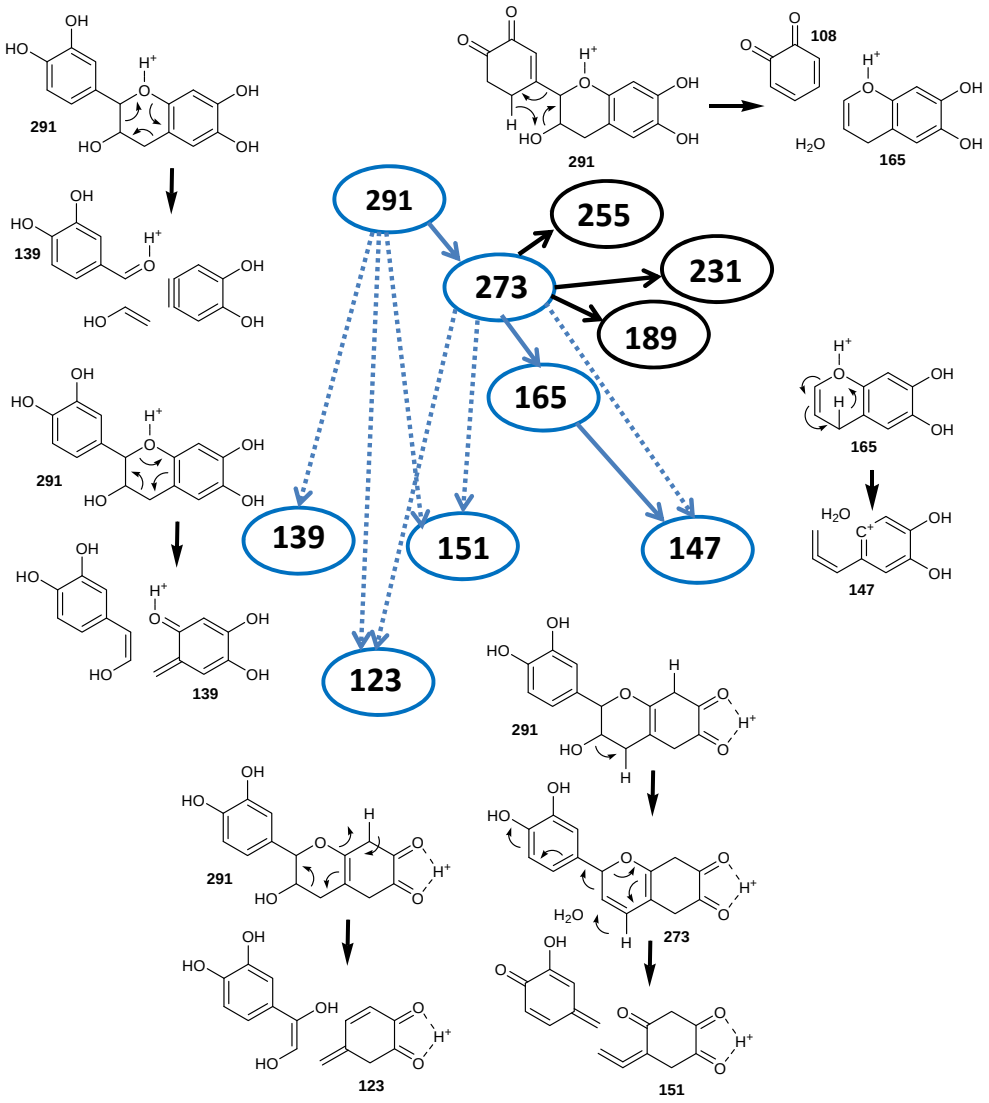

Figure S-3: Fragmentation pathway from (-)-epicatechine MS^3 CID experiments; numbers represent *m/z* ratios. Blue nodes and arcs correspond to the calculated tree (Figure S-1); Black edges correspond to NL not visible in (-)-epicatechine MS^2 CID spectrum. Dashed edges show fragment pathways which differ from the calculated trees. Some of them can be explained by "pull-ups".

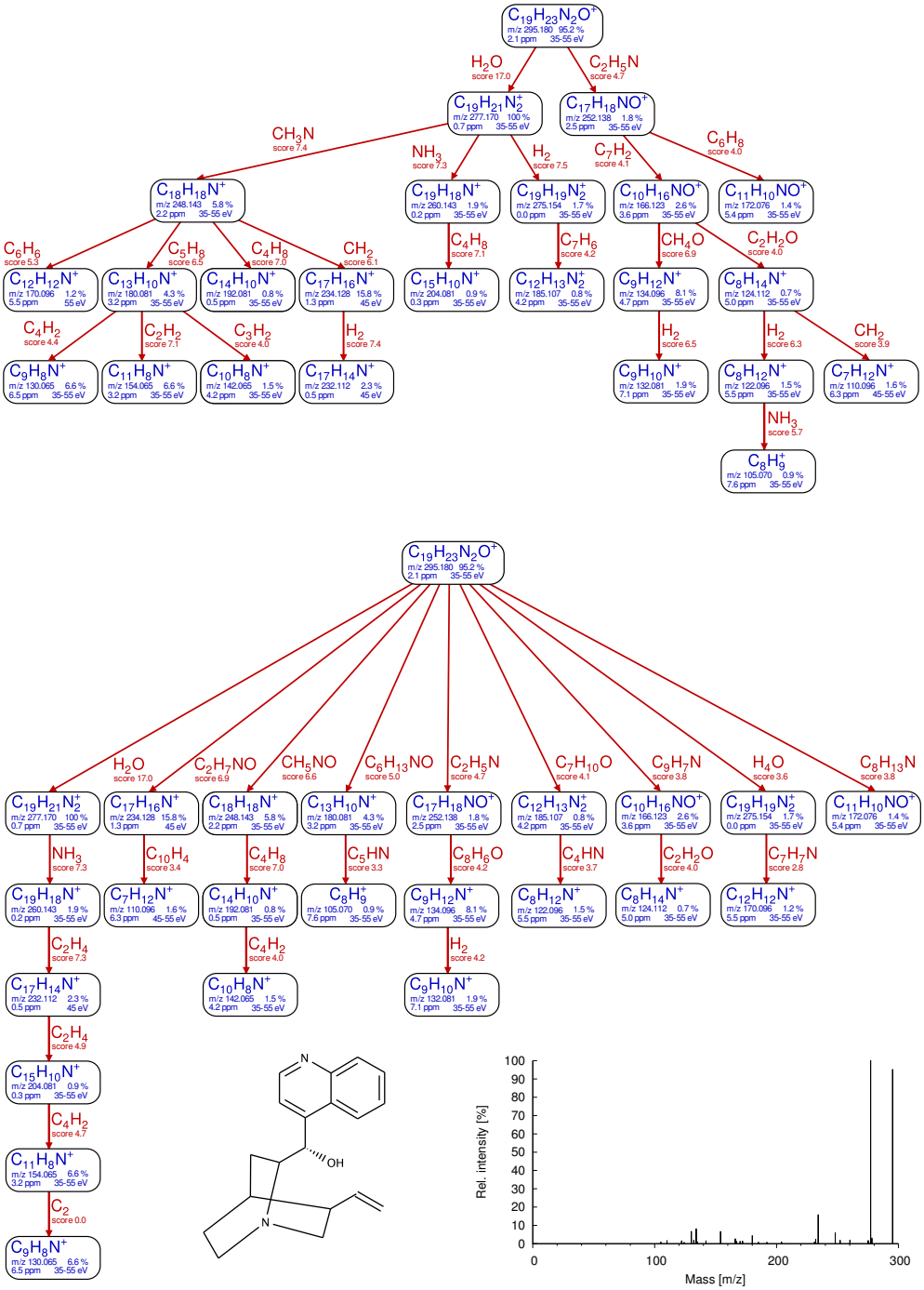

Figure S-4: Example of the failure of heuristic algorithms to reconstruct the correct fragmentation tree, for compound chinchonine ($C_{19}H_{22}N_2O$) from the Orbitrap dataset. For the exact algorithm (top), 19 of 23 NLs (83%) are marked "correct", 1 NL (4%) is marked "unclear", and 3 NLs (13%) are marked "wrong". For the heuristic algorithm (bottom), 9 of 23 NLs (39%) are marked correct, 4 (17%) NLs are marked "unclear", and 10 NLs (44%) are marked "wrong". Both algorithms explain all 24 peaks in the spectrum.

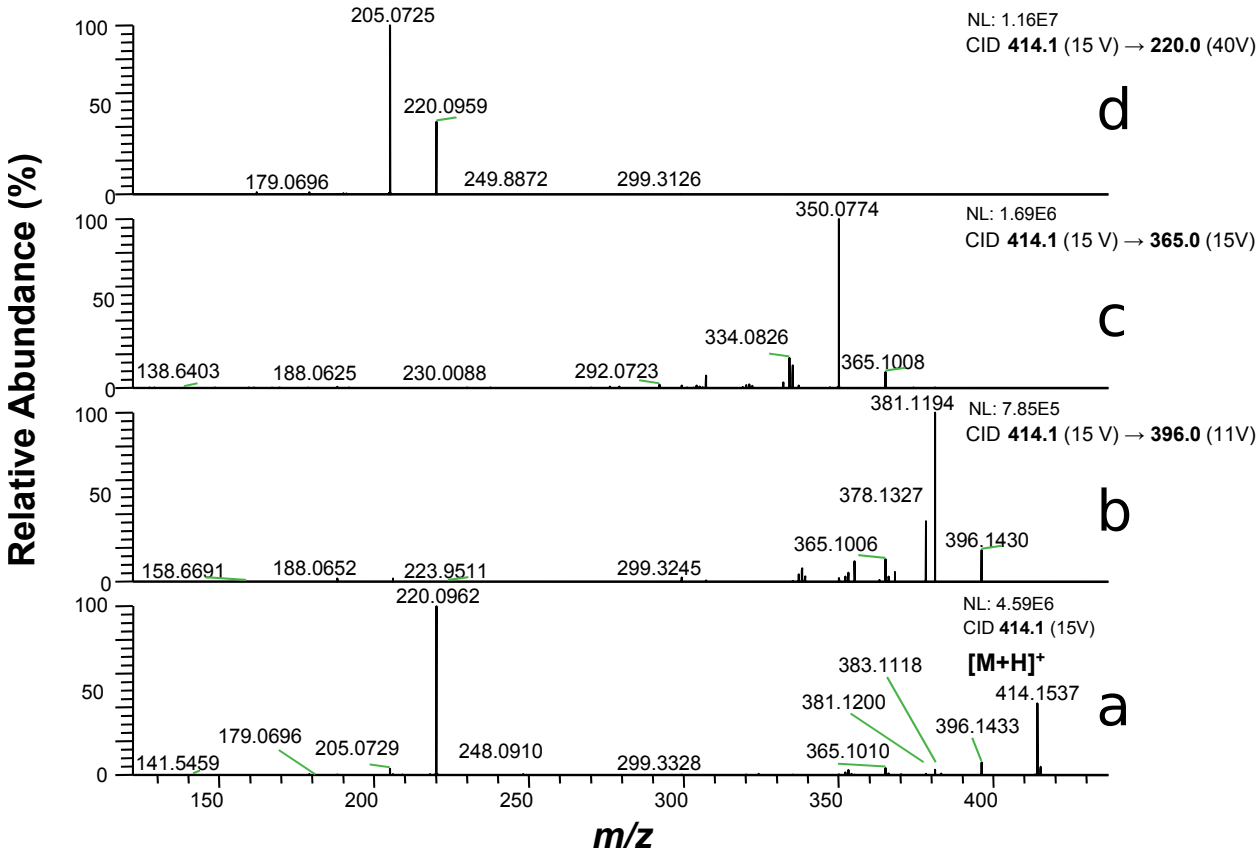

Figure S-5: A series of tandem mass spectrometry experiments performed with protonated (S,R)-noscapine generated by electrospray and analyzed with an Orbitrap XL instrument. Fragmentation was realized in linear trap using He as collision gas. (a) CID MS^2 spectrum generated from molecular adduct ion [M+H]+ using 15 V in linear trap (other used CID voltages given in brackets). (b–d) MS^3 spectra; transitions are given in inserts in bold, used collision energies are indicated in brackets. Intensities (NLs, instrument internal units) are given for all experiments, but spectra are plotted on relative scale (%) for better comparison.

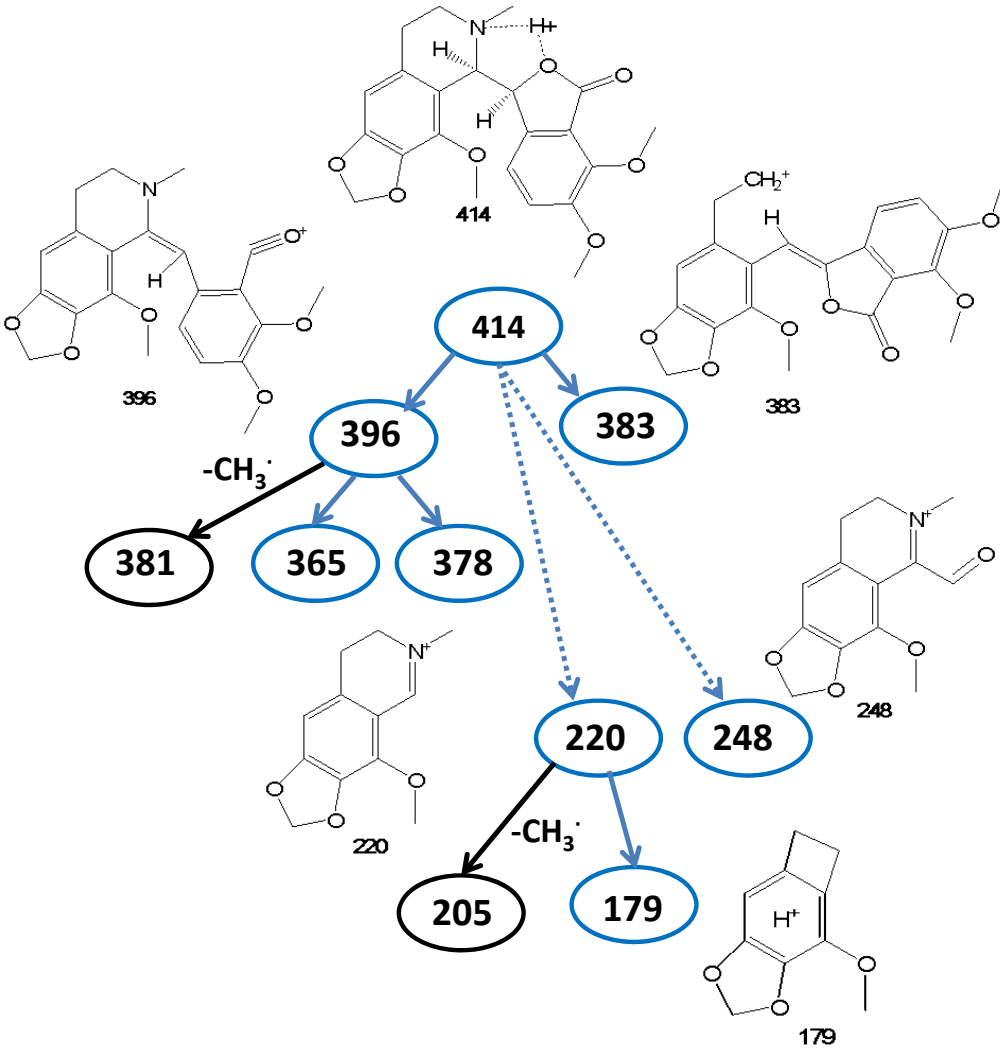

Figure S-6: (S,R)-Noscapine experimental fragmentation pathway; numbers represent *m*/*z* ratios. Arcs and nodes colored in blue are present in the calculated fragmentation tree (Fig. 2); the dotted blue lines represent pull-ups. Black nodes and arcs represent intense ions in the experimental MS^3 spectra which are absent in the tree. Five correct arcs, two pull-ups, and no wrong NL annotations were found by experimental validation and expert evaluation.

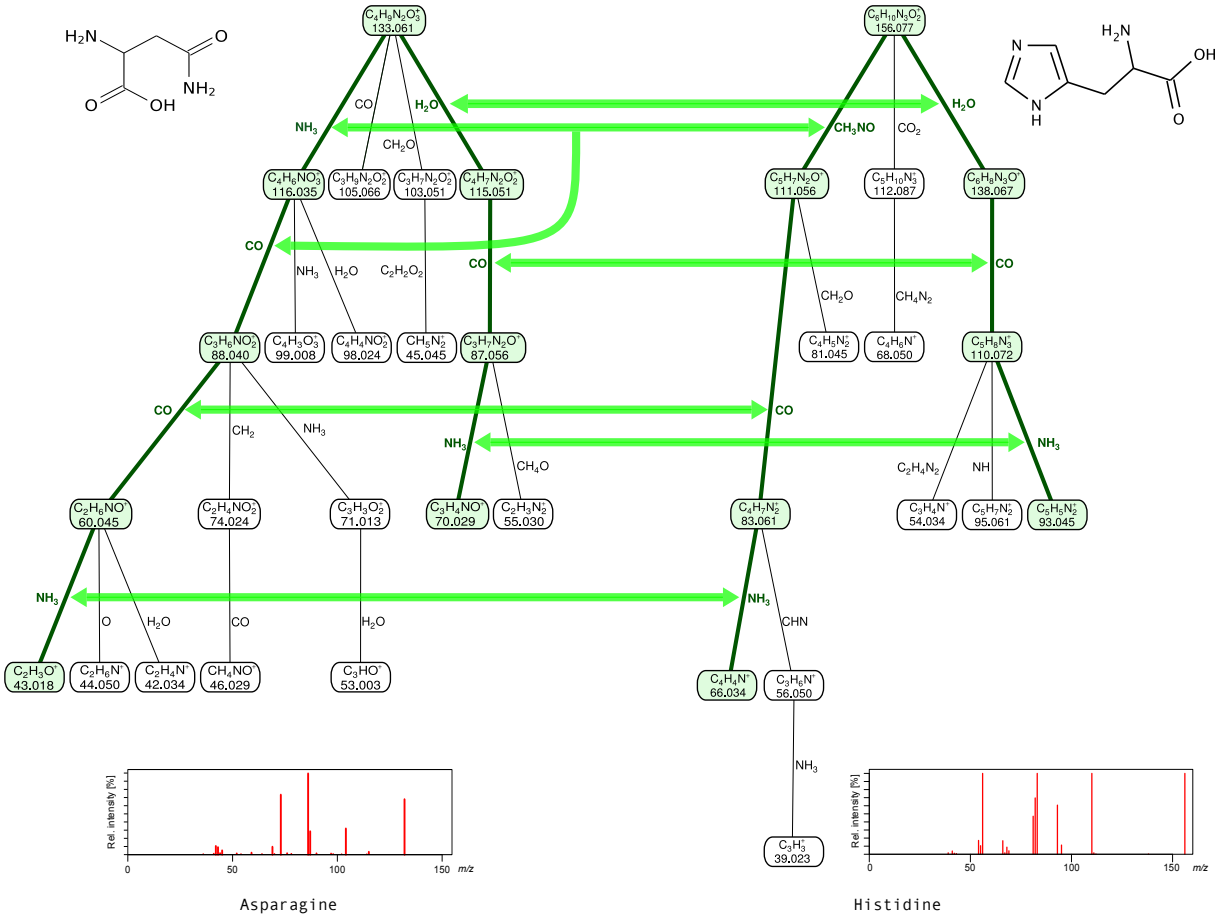

Figure S-7: Comparison between the calculated fragmentation trees of histidine ($C_6H_9N_3O_2$) and asparagine ($C_4H_8N_2O_3$) from the API QSTAR dataset. Uncolored arcs are not part of the common subtree.

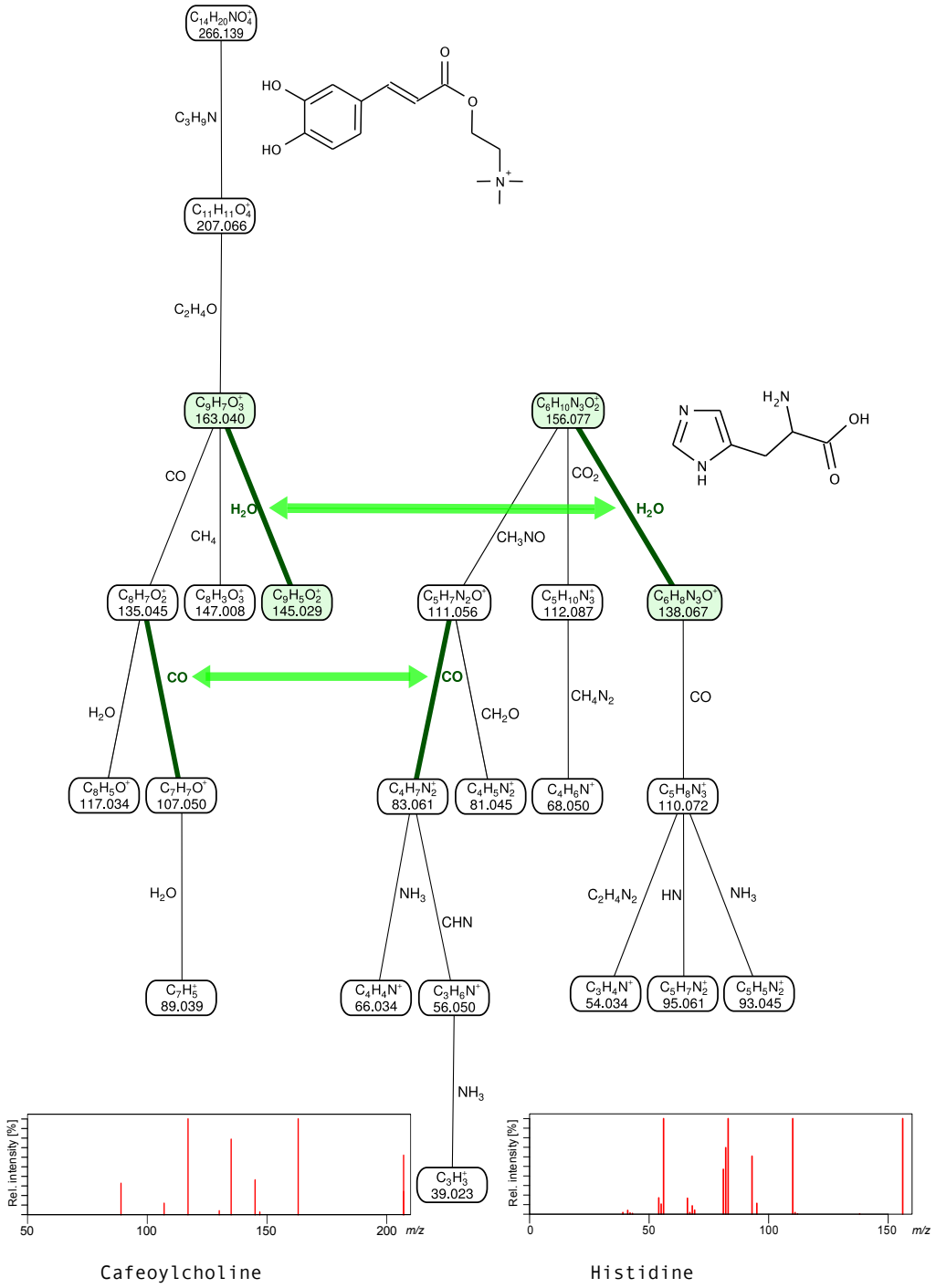

Figure S-8: Exemplary comparison between the fragmentation trees of histidine ($C_6H_9N_3O_2$) and cafeoyl choline ($C_{14}H_{18}NO_4$) from the API QSTAR dataset. The best common subtree contains only one edge (either CO or H2O). A comparison of histidine with other choline esters leads to results of comparable quality, clearly indicating that histidine is not a choline.

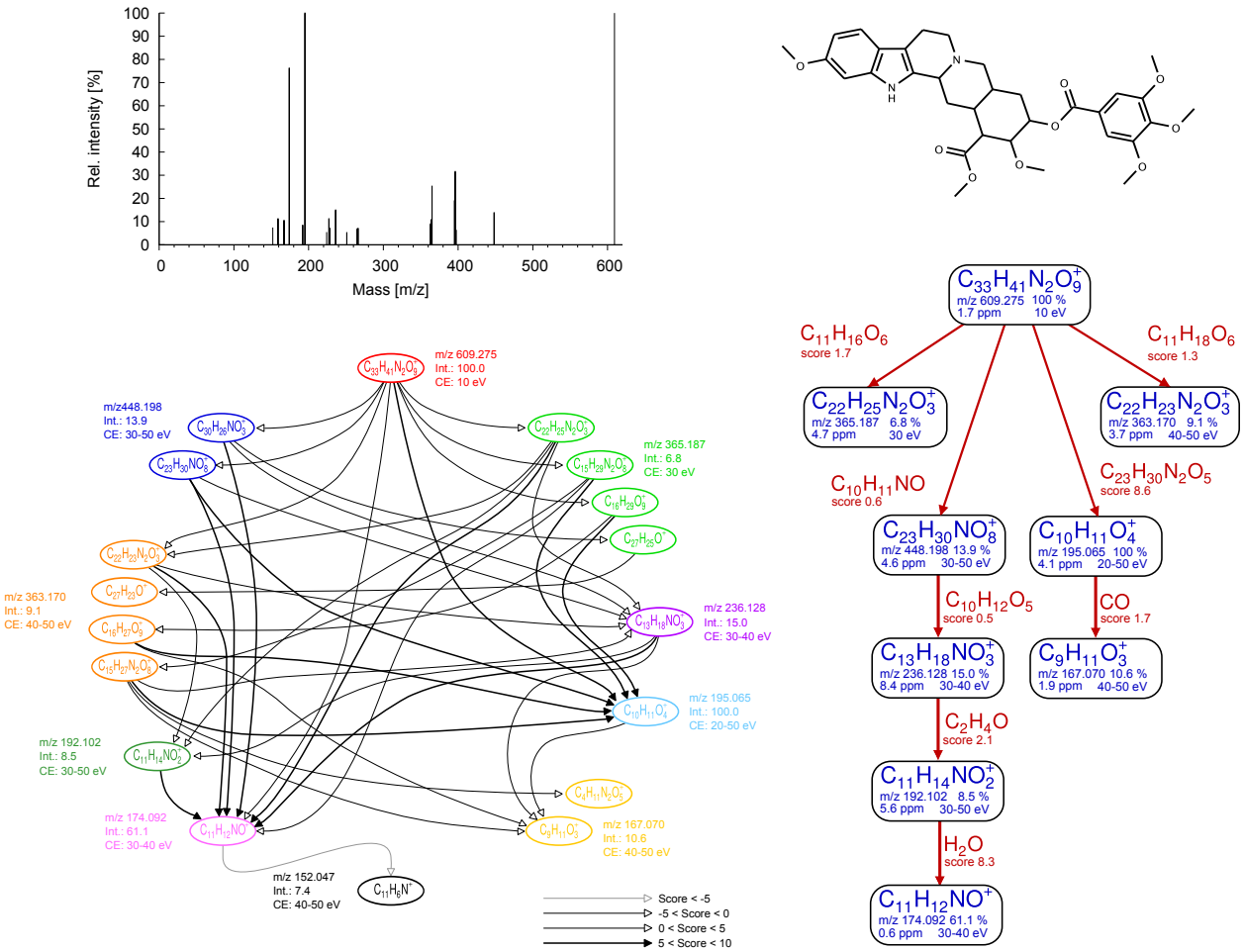

Figure S-9: Left: Fragmentation graph for reserpine ($C_{33}H_{40}N_2O_9$) using Micromass QTOF data. Nodes of the same color correspond to annotations of one measured peak (*m/z*, intensity, and fragmentation energies). Arcs correspond to potential neutral losses. Weight of arcs encoded by different line types, transitive arcs with weight below 0 omitted. NLs can be computed by subtracting molecular formulas for end node and start node. Right: The corresponding hypothetical fragmentation tree of reserpine computed by our method. Nodes (blue) correspond to peaks in the tandem mass spectra and their annotate molecular formula (CE is the range of collision energies), arcs (red) corresponding to hypothetical neutral losses. For clarity, peaks below 5% relative intensity have been excluded from this calculation. See the separate PDF file for the full tree of reserpine.

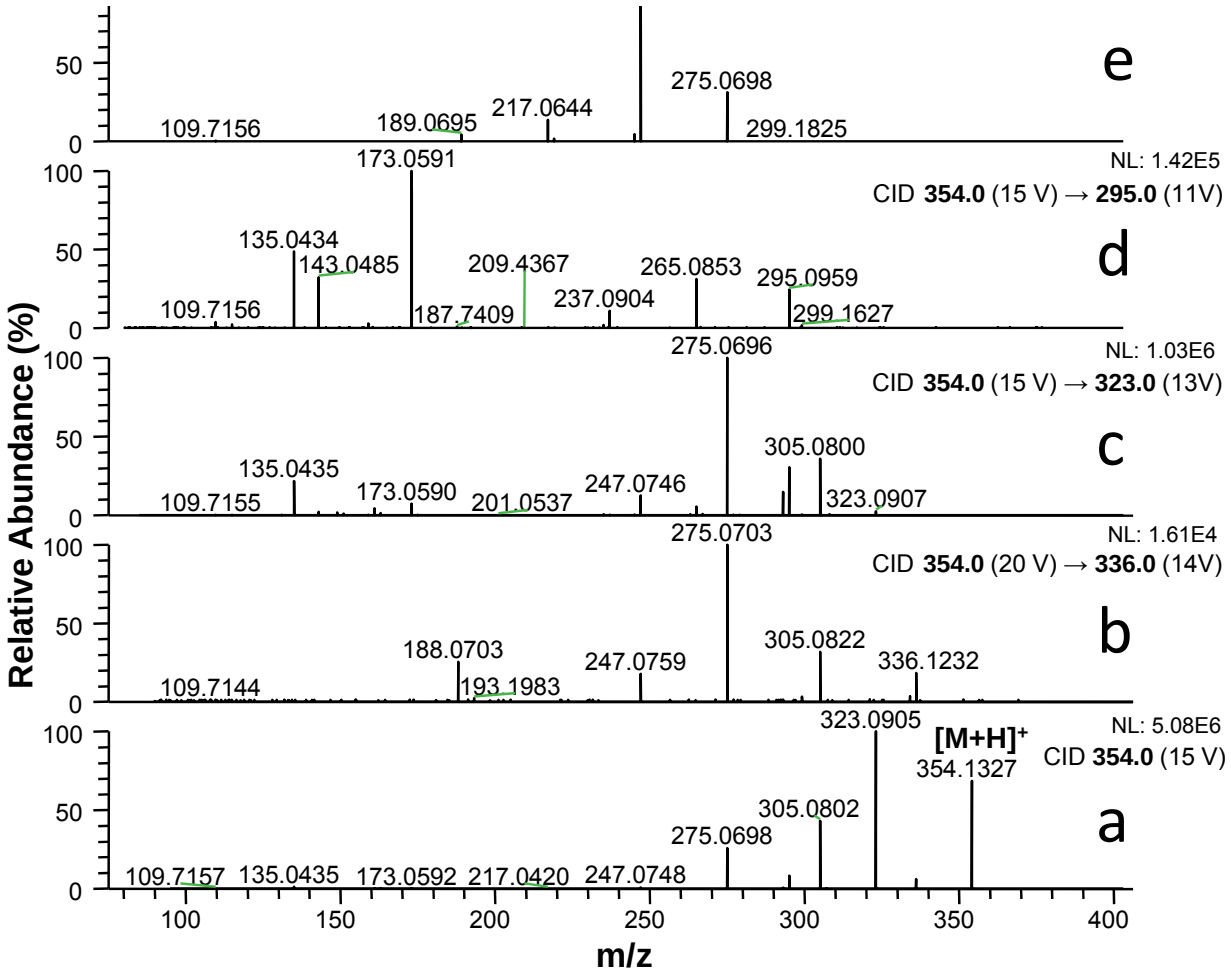

Figure S-10: A series of tandem mass spectrometry experiments performed with protonated chelidonine generated by electrospray and analyzed with an Orbitrap XL instrument. Fragmentation was realized in linear trap using He as collision gas. (a) CID MS^2 spectrum generated from molecular adduct ion [M+H]+ using 15 V in linear trap (other used CID voltages given in brackets). (b–e) MS^3 tandem mass spectra; transitions are given in inserts in bold, used collision energies are indicated in brackets. Intensities (NLs, instrument internal units) are given for all experiments, but spectra are plotted on relative scale (%) for better comparison.

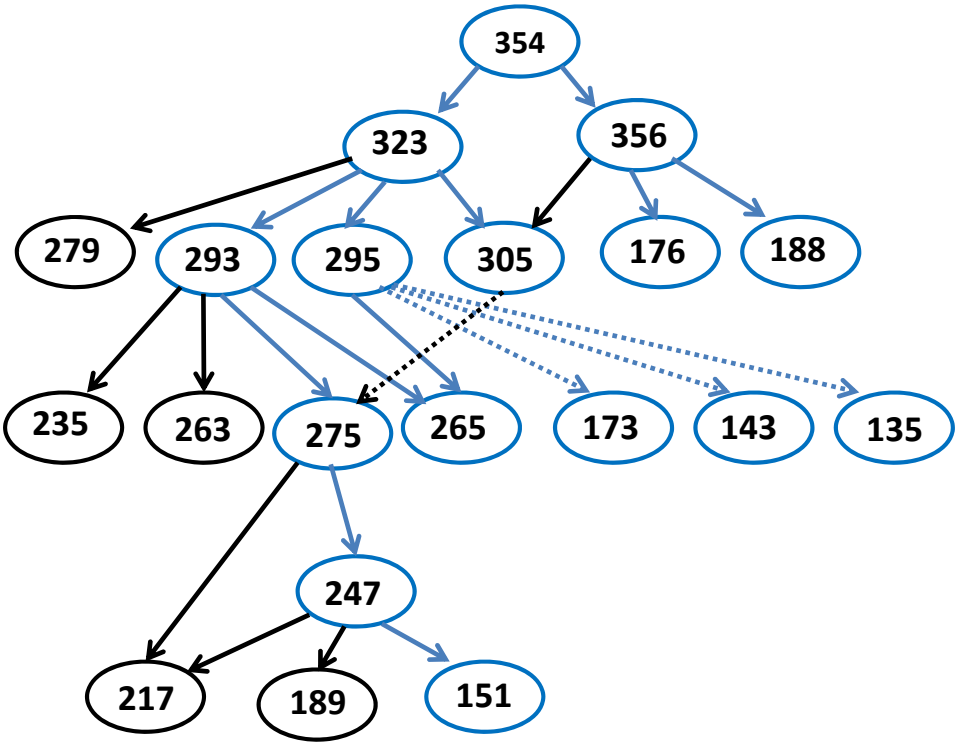

Figure S-11: Fragmentation pathway from chelidonine MS^3 and MS^4 CID experiments; numbers represent *m*/*z* ratios. Blue nodes and arcs correspond to the calculated tree (Fig. 3) Black edges correspond to NL not visible in the chelidonine MS^2 CID spectrum. Dashed edges show fragment pathways which differ from the calculated trees. Some of them can be explained by "pull-ups".