MS/MS DATA IMPROVES AUTOMATED DETERMINATION OF MOLECULAR FORMULAS BY MASS SPECTROMETRY

Markus Meringer[†,1], Stefan Reinker[‡], Juan Zhang[‡], Alban Muller[‡]

[†]German Aerospace Center (DLR),
Münchener Straße 20,
D–82234 Oberpfaffenhofen–Wessling, Germany

[‡]Novartis Institutes for BioMedical Research (NIBR),
Postfach, Novartis Campus,
CH–4002 Basel, Switzerland

Received July 26, 2010

Keywords: Tandem MS, Orbitrap, Structure Elucidation, Algorithm, Biomarker, Natural Product.

Abstract. In theory, the molecular formula of an unknown compound can be calculated from its exact molecular mass. However, even with highly accurate modern mass spectrometers with an accuracy of 1 ppm or lower, it is generally not possible to determine the molecular formula uniquely from measurements. Intensities of isotopic peaks are typically used as additional information to narrow down possible formulas associated with a mass spectrum (MS) peak, but this is not sufficient for larger compounds.

Here, we introduce a method that takes information from fragment peak masses of the MS/MS into account to improve the reliability of formula determination. Matchvalues that reflect the consistency with MS isotope peaks and MS/MS fragment patterns are computed for candidate molecular formulas. We demonstrate that these matchvalues outperform methods based on isotope peak intensities alone. In test cases with medium sized organic molecules (< 1000 u) the true molecular formula achieves the highest matchvalues using the MS/MS data.

1. Introduction

When elucidating the molecular structure of an unknown compound, one of the first steps is usually the determination of the molecular formula. One of the analytical key methods to determine the molecular formula of an unknown is mass spectrometry. In contrast to electron impact MS, soft ionization techniques tend to keep the molecular ion

[1]Corresponding author e–mail markus.meringer@dlr.de, phone +49 8153 281446, fax +49 8153 281446

intact and help to identify the molecular mass [1, 2]. Starting from a measured mass one can compute candidate molecular formulas that match this mass [3].

The number of candidates found mainly depends on the set of chemical elements that are considered [4], the order of magnitude of the molecular mass, and the mass accuracy of the spectrometer. Modern high resolution methods deliver precise mass measurements with only a few parts per million (ppm) deviation from the true mass [5]. Increasing mass accuracy by technical means reduces the number of candidate molecular formulas. However, it is in general not possible yet to allow unambiguous assignment of the molecular formula by high mass accuracy alone [6].

A means to reduce the number of candidates is to compare their theoretical isotopic peak patterns with the measured intensities of isotopic peaks. A ranking of candidates according to the match of theoretical and experimental intensities can be used to exclude some candidates [7, 8, 9, 10, 11, 12, 13, 14].

However, high mass accuracy and isotope matching are often still insufficient for unambiguous assignment of formulas to peak masses, especially at higher masses. Heuristic criteria have been formulated to recognize molecular formulas that are likely to belong to chemical compounds [15, 16] and to exclude formulas with impossible or unlikely combinations of elements and element ratios.

Recently, spectrometers have become available that are able to measure MS/MS data based on collision induced dissociation (CID). While it has often been noted that MS/MS data can be used for determination of the constitution (i.e. structural formula) of unknowns [17] it has almost been overseen that accurate masses from MS/MS data can be used as additional, complementary filters for the determination of the molecular formula. A few studies [18, 19, 20] offer no details on algorithms and present only minimalistic examples, except for [21].

In the following, we outline known algorithmic methods for using accurate masses and isotopic patterns from the MS to calculate an isotopic matchvalue. We describe how to similarly define a matchvalue based on MS/MS data and present several examples that demonstrate how this additional information helps to determine the molecular formula. The main focus is on small and medium–sized molecules with a mass range up to approximately 1000 u.

Applications of this method could lie in drug discovery, metabolomics, molecular diagnostics and environmental chemistry. Examples are the identification of drugs within natural product extracts [22], biomarkers within body fluids [23, 24], or toxic compounds in environmental samples [25].

Natural product libraries are biochemically screened for certain targets. If an active compound is found, mass spectrometry is used to

determine its molecular mass. MS/MS measurements could help to deliver its molecular formula, a first step to identify the structure and possibly synthesize the compound. The key analytical techniques for this approach, liquid chromatography, flow screening and high resolution mass spectrometry have recently been integrated into one platform [22, 26], which also enables the measurement of MS/MS data in parallel.

Further, many metabolites found in blood or other body fluids or tissues have not been identified yet, and determination of their molecular formula through MS based methods would be a first step to identify their structure and role in metabolism. Similarly, biomarkers for severity of diseases and treatment success can be found using MS techniques on body fluids. However these biomarkers need to be identified.

2. Theory

The standard approach to determine the molecular formula of an unknown compound is to measure the compound's molecular mass and then find a formula that fits this mass. This procedure is based on the fact that chemical elements have different atomic masses.

The difference in atomic masses is caused by the fact that atoms of different chemical elements consist of different numbers of elementary particles, in addition to the mass defect which leads to atomic masses differing from the sum of masses of elementary particles.

Tables of atomic masses, including isotopic masses and isotopic distributions are published and updated frequently [27, 28].

Table 1 shows nominal masses $\bar{m}_X$ and exact monoisotopic atomic masses m_X for eleven chemical elements X that are of importance for organic chemistry:

$$\mathcal{E}_{11} := \{\mathsf{H}, \mathsf{C}, \mathsf{N}, \mathsf{O}, \mathsf{F}, \mathsf{Si}, \mathsf{P}, \mathsf{S}, \mathsf{Cl}, \mathsf{Br}, \mathsf{I}\}.$$

The monoisotopic mass of an element is the mass of its most abundant isotope, typically specified in unified atomic mass units (u). Of course the following algorithms can be carried out with any other set of chemical elements. As a second set of elements we will consider elements typically occurring in biochemistry and metabolomics

$$\mathcal{E}_8 := \{\mathsf{H}, \mathsf{C}, \mathsf{N}, \mathsf{O}, \mathsf{S}, \mathsf{Cl}, \mathsf{Br}, \mathsf{I}\},$$

and also a smaller set of the most frequent elements

$$\mathcal{E}_4 := \{\mathsf{H}, \mathsf{C}, \mathsf{N}, \mathsf{O}\}.$$

2.1. Generating molecular formulas.

Mathematically, a molecular formula β can be considered as a mapping

$$\beta : \mathcal{E} \longrightarrow \mathbb{N}, \qquad X \longmapsto \beta(X)$$

X	$\bar{m}_X$	m_X	$I_X(\bar{m}_X)$	$I_X(\bar{m}+1)$	$I(\bar{m}_X+2)$	v_X
H	1	1.007825	1.0000			1
C	12	12.000000	0.9890	0.0110		4
N	14	14.003074	0.9963	0.0037		3
O	16	15.994915	0.9976	0.0004	0.0020	2
F	19	18.998403	1.0000			1
Si	28	27.976928	0.9223	0.0467	0.0310	4
P	31	30.973763	1.0000			3
S	32	31.972072	0.9504	0.0075	0.0421	2
Cl	35	34.968853	0.7577		0.2423	1
Br	79	78.918336	0.5069		0.4931	1
I	126	126.904477	1.0000			1

TABLE 1. Nominal masses $\bar{m}_X$, exact monoisotopic masses m_X, relative isotopic frequencies I_X and element valencies v_X used in our computations.

from a set of chemical elements onto the set of natural numbers. This mapping relates each chemical element X to its multiplicity $\beta(X)$.

2.1.1. *Rules for molecular formulas.* Not every such mapping represents the molecular formula of a chemical compound. There are two types of rules that are able to recognize invalid or unlikely molecular formulas: mathematical and heuristic rules.

Mathematical rules are based on the fact that (uncharged) chemical compounds correspond to (non–ionic) molecular graphs [29]. For example, $\mathsf{CH_2}$, $\mathsf{C_2H_8}$, $\mathsf{C_2H_7O}$ do not correspond to a molecular graph. In order to result in a molecular graph, further restrictions in terms of the atoms' valences v_X have to be fulfilled:

(i) $\sum_{X\in\mathcal{E}} v_X\beta(X) \equiv 0 \bmod 2$,

(ii) $\sum_{X\in\mathcal{E}} v_X\beta(X) - 2\max_{X\in\mathcal{E}}\{v_X \mid \beta(X) > 0\} \geq 0$,

(iii) $\sum_{X\in\mathcal{E}} v_X\beta(X) - 2\sum_{X\in\mathcal{E}} \beta(X) + 2 \geq 0$.

The left-hand side of equation (i) is the sum of all valences, which is required to be an even number in order to avoid dangling bonds. Inequality (ii) specifies that there are sufficient bonds available for the atom of maximum valency. Condition (iii) requires a molecular graph to be connected (one component). The above examples $\mathsf{C_2H_7O}$, $\mathsf{CH_2}$, and $\mathsf{C_2H_8}$ violate restrictions (i), (ii), and (iii), respectively. These restrictions were first formulated in [30], a derivation is found in [31]. For our computations we used valences v_X as shown in Table 1.

In addition to these mathematical rules, there exist several heuristic rules [15, 16], which were obtained by statistical examinations of

large compound databases. For instance, such criteria forbid formulas with unlikely hydrogen to carbon ratios or untypical combinations of different heteroatoms. Although such rules can be very useful tools for reducing large lists of candidate formulas, we will not use them here, as there always is a certain probability of missing the true candidate. In Section 4 it will become apparent that with our method most unlikely formulas achieve only low MS/MS matchvalues and can be eliminated without the need to apply heuristic rules. The only heuristic rule we use is to presume that an organic compound has at least one carbon atom.

2.1.2. *From mass to formula.* The monoisotopic mass of a compound with molecular formula β is defined as the sum of monoisotopic masses of its atoms

$$m_\beta = \sum_{X \in \mathcal{E}} m_X \beta(X),$$

and the calculated mass m' of the molecular ion $[\mathrm{M+H}]^+$ is

$$m' = m_\beta + m_{\mathsf{H}} - m_{\mathrm{e}},$$

where $m_{\mathrm{e}} = 5.485799 \cdot 10^{-4}$ u denotes the mass of an electron. Let m be the measured mass obtained from the MS. The relative deviation of m with respect to m' is

$$\Delta(m) = \frac{1}{m} \cdot |m' - m|.$$

Given an instrument accuracy of δ, the condition $\Delta(m) \leq \delta$ must be fulfilled.

Using these assumptions it is possible to generate candidate molecular formulas of an unknown compound from its measured mass m by solving the inequalities

$$m \cdot (1 - \delta) \leq \sum_{X \in \mathcal{E}} m_X \beta(X) + m_{\mathsf{H}} - m_{\mathrm{e}} \leq m \cdot (1 + \delta).$$

under conditions (i) – (iii). We apply a backtracking algorithm described in [32] to generate solutions. Similar algorithms have been formulated earlier [3, 33].

2.2. **Calculating MS matchvalues.** For each candidate formula generated, a matchvalue can be computed that shows how well the theoretical isotope distribution matches the measured intensities of the isotopic peaks in the MS.

For the low and medium mass region in the focus our studies, it is sufficient to calculate isotopic distributions based on integer masses.

An integer resolution mass spectrum I can be considered as a mapping

$$I : \mathbb{N} \longrightarrow \mathbb{R}_+^0, \qquad \bar{m} \longmapsto I(\bar{m})$$

from the set of natural numbers onto the set of non–negative real numbers. This mapping relates each integer m/z value $\bar{m}$ with an intensity $I(\bar{m})$.

In this manner we can describe experimental spectra as well as theoretical isotope distributions and calculated spectra.

Table 1 shows the natural isotope distributions I_X of the most common organic elements $X \in \mathcal{E}_{11}$ according to [28]. For empty fields we have $I_X(\bar{m}) = 0$. There are four monoisotopic elements listed: $\mathsf{H}, \mathsf{F}, \mathsf{P}$ and I. Hydrogen isotopes Deuterium $^2\mathsf{H}$ and Tritium $^3\mathsf{H}$ are excluded for their extremely low abundance and are not considered in our calculations.

2.2.1. *Calculating theoretical isotope distributions.* Isotope distributions of molecular formulas can be calculated by convolution of element isotope distributions. The convolution $I_1 \cdot I_2$ of two isotope distributions I_1 and I_2 is defined as

$$(I_1 \cdot I_2)(\bar{m}) = \sum_{i=0}^{\bar{m}} I_1(i) I_2(\bar{m} - i). \tag{2.1}$$

Using definition 2.1, the isotope distribution I_β of a molecular formula β can be expressed as composition of convolutions:

$$I_\beta = \prod_{X \in \mathcal{E}} {I_X}^{\beta(X)}.$$

For an example see [34], an early approach for calculation of isotopic distributions was also described in [35].

Isotope distributions can also be calculated based on exact masses. To be precise, isotopic peaks are composed of several different isotopomers. Taking, for example $\mathsf{C_2H_4O}$, there are two isotopomers with nominal mass 46, $^{12}\mathsf{C_2H_4}{}^{18}\mathsf{O}$ and $^{13}\mathsf{C_2H_4}{}^{16}\mathsf{O}$, and even more if hydrogen isotopes are taken into account. Calculation of isotope distributions based on precise masses gains importance for large molecules, where the true mass differs considerably from the nominal mass. Several algorithms have been proposed for calculation of precise mass isotopic distributions [36, 37, 38].

For our purposes, however, it is sufficient to use integer mass isotopic distributions, as differences between nominal and precise mass higher than 0.5 u are simply compensated by shifting the calculated isotope distribution.

2.2.2. *Comparing theoretical isotope distributions and measured intensities.* Comparing computed and measured isotope distributions is a vector comparison between $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$.

There are various methods for comparing two vectors, for instance

- the dot product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_i x_i y_i$,

- the sum of absolute errors $||\mathbf{x}-\mathbf{y}||_1 = \sum_i |x_i - y_i|$, or
- the sum of squared errors $||\mathbf{x}-\mathbf{y}||_2^2 = \sum_i (x_i - y_i)^2$.

Note, that for the latter two, the input vectors should be normalized. Therefore we normalize measured and computed intensities so that the sums $\sum x_i$ and $\sum y_i$ equal 1.

The resulting MS matchvalue should be a normalized, continuous quantity between 0 an 1 with value 1 if the two vectors are identical and 0 if they are entirely distinct. This can be achieved by using

- the normalized dot product

$$NDP(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{y}\rangle}{\sqrt{||x||_2 \cdot ||y||_2}},$$

- the normalized sum of absolute errors [39]

$$NSAE(\mathbf{x},\mathbf{y}) = 1 - \frac{||\mathbf{x}-\mathbf{y}||_1}{||\mathbf{x}+\mathbf{y}||_1}, \text{or}$$

- the normalized sum of squared errors

$$NSSE(\mathbf{x},\mathbf{y}) = 1 - \frac{||\mathbf{x}-\mathbf{y}||_2^2}{||\mathbf{x}+\mathbf{y}||_2^2}.$$

There are numerous other methods to compute matchvalues for isotopic distributions, for instance by using the correlation coefficient [40], or by applying other normalizations, e.g. normalizing the base peak intensity to a certain value.

2.3. **Calculating MS/MS matchvalues.** A fragment of an ion in the MS/MS has to be a subformula β' of its precursor ion β, that is

$$\forall X \in \mathcal{E} : \ \beta'(X) \leq \beta(X).$$

Since ions are charged particles and also can be radicals, condition (i) does not necessarily hold for fragment formulas.[2]

Calculating a matchvalue for molecular formula candidate β with MS/MS data can be achieved by trying to explain every peak in the MS/MS spectrum by a subformula β'. For a false candidate for the mother ion, there might be peaks in the MS/MS spectrum that cannot be explained by any subformula. The idea of taking all possible subformulas of a molecular formula candidate into account for the calculation of a MS related matchvalue was first formulated in [41].

That is, a subformula β' can explain a measured MS/MS peak if the calculated mass m' of β' as a simply charged cation matches the observed peak mass within a certain tolerance, which mainly depends on the spectrometer's accuracy.

[2]In ions and radicals, valencies of atoms might vary from their defaults. Thus also conditions (ii) and (iii) are not fulfilled in general. However, in our computations the filters (ii) and (iii) were always applied in order to avoid very unlikely subformulas.

Similar to Subsection 2.1, let $m_i, i = 1, ..., n$ be the accurate masses of the MS/MS spectrum and $m'_j, j = 1, ..., n'$ the calculated ion masses of the subformulas that fulfill rules (ii) and (iii). We define

$$\Delta(m_i) = \frac{1}{m_i} \cdot \min_{j \leq n} |m'_j - m_i|.$$

Let ε denote the MS/MS mass accuracy. With

$$\omega(m_i) := \begin{cases} 1 & \text{if } \Delta(m_i) \leq \varepsilon, \\ 0 & \text{else.} \end{cases}$$

a MS/MS matchvalue can be defined as

$$MV_1 = \frac{1}{n} \sum_{i=1}^{n} \omega(m_i),$$

that is MV_1 represents the ratio of MS/MS peaks which can be explained by a subformula of the candidate molecular formula.

2.3.1. *Fuzzy logic for peak acceptance or rejection.* Uncertainty of the MS/MS mass accuracy can be reflected by introducing an upper and a lower bound for the precision. Let ε be the accuracy threshold for acceptance and σ the threshold for rejection, $\sigma \geq \varepsilon$. With $\Delta(m_i)$ defined as above, we refine $\omega(m_i)$ to

$$\omega(m_i) := \begin{cases} 1 & \text{if } \Delta(m_i) \leq \varepsilon, \\ \frac{\sigma - \Delta(m_i)}{\sigma - \varepsilon} & \text{if } \varepsilon \leq \Delta(m_i) \leq \sigma, \\ 0 & \text{else.} \end{cases}$$

Then peaks with $\Delta(m_i)$ above ε and below σ are counted with an interpolated value between 0 and 1.

2.3.2. *Weighting by peak intensities and peak masses.* In addition to the mass deviation the peak intensities $I(m_i)$ can be taken into account. Peaks with high intensity should be weighted stronger than small peaks:

$$MV_2 = \frac{\sum_{i=1}^{n} \omega(m_i) \cdot I(m_i)}{\sum_{i=1}^{n} I(m_i)}$$

Another idea is motivated by the fact that peaks of higher mass might be more important for the identification of the molecular formula than peaks of small masses. A matchvalue that takes the peak mass itself into account could be defined as:

$$MV_3 = \frac{\sum_{i=1}^{n} \omega(m_i) \cdot m_i \cdot I(m_i)}{\sum_{i=1}^{n} m_i \cdot I(m_i)}$$

Finally, we can generalize our considerations about weighting peak intensity and mass by introducing

$$MV = \frac{\sum_{i=1}^{n} \omega(m_i) \cdot f(m_i, I(m_i))}{\sum_{i=1}^{n} f(m_i, I(m_i))}$$

with an arbitrary function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, which maps as the first component the peak mass and as the second component the peak intensity onto a combined weight.

In our measurements of test samples (see below) frequently there appears to be one or a few peaks that dominate the fragment spectrum. In this case it can be helpful to use the logarithm of the peak heights for weighting the contribution of the individual peaks to the matchvalue in order not to overemphasize the importance of one fragment peak.

2.3.3. *Criteria for ion formulas.* In Subsection 2.1 we defined criteria for molecular formulas, based on the sum of atom valencies. This concept is also known as the number of double bond equivalents. For a molecular formula β the number of double bond equivalents is defined as

$$DBE(\beta) = 1 + \frac{1}{2} \sum_{X \in \mathcal{E}} \beta(X)(v_X - 2)$$

Simply positively charged radical ions, also known as odd electron ions, have integer DBE values. Since such ions rarely appear in MS/MS spectra we can restrict subformula searches to non–integer DBE values.

Another criterion for ions in a mass spectrometer is that they should not significantly exceed the DBE of the precursor ion. For electron impact MS it has been found that a good value for limiting DBE excess, which could stem from chemical rearrangements after fragmentation, is three [42]. This is also a useful constraint for the calculation of a MS/MS matchvalue in some examples.

2.4. **Assessing the goodness of matchvalues.** Finally, we need a concept to asses the goodness of matchvalues. We are looking for a procedure that is able to decide which method of calculating matchvalues is better suited to distinguish the true candidate from false candidate formulas. For this purpose we calculate matchvalues for all formula candidates and sort them in descending order, resulting in a ranking of candidates.

In order to evaluate the quality of a ranking we can either use the absolute or the relative position of the true formula among formula candidates. We define the absolute ranking position (ARP) simply by the number of better candidates (BC, the number of candidates having higher matchvalue than the true formula) plus 1.

Relative ranking positions are more useful than absolute ranking positions if examples with different numbers of candidate formulas have to be compared. The relative ranking position (RRP) should have a value between 0 and 1, with lower values reflecting better rankings. The RRP should be 0 if the true formula is ranked first and 1 if the true formula is ranked last.

Let WC denote the number of worse candidates, i.e. candidates with a lower MV than the true formula, and let TC be the total number of

Nr	Compound	CID	Formula	$\bar{m}$
1	Creatine	586	$C_4H_9N_3O_2$	131
2	Sinapinic acid	637775	$C_{11}H_{12}O_5$	224
4	Chloropropoxy-9*H*-thioxanthen-9-one	5212856	$C_{16}H_{13}O_2ClS$	304
4	Testosterone acetate	92145	$C_{21}H_{30}O_3$	330
5	Cholesteryl butyrate	101741	$C_{31}H_{52}O_2$	456
6	Peptide MRFA	3565545	$C_{23}H_{37}N_7O_5S$	523
7	Reserpine	5770	$C_{33}H_{40}N_2O_9$	608
8	CHAPS	16211615	$C_{32}H_{58}N_2O_7S$	614
9	Maltopentaose	13489094	$C_{30}H_{52}O_{26}$	828
10	Cyclosporin C	6438160	$C_{62}H_{111}N_{11}O_{13}$	1217

TABLE 2. Overview of the samples; see text for details.

candidates. There are two possibilities to define a relative ranking position:

$$\mathrm{RRP}_0 := \frac{\mathrm{BC}}{\mathrm{TC}-1} \quad \text{and} \quad \mathrm{RRP}_1 := 1 - \frac{\mathrm{WC}}{\mathrm{TC}-1}.$$

Of course RRP_0 and RRP_1 are defined only if at least two candidates exist. Both definitions fulfill the above requirements, but in the case of false candidates having the same MV as the true candidate, RRP_0 and RRP_1 will differ. In order to take such situations into account, we finally define the relative ranking position as the mean of RRP_0 and RRP_1:

$$\mathrm{RRP} := \frac{1}{2}\left(1 + \frac{\mathrm{BC}-\mathrm{WC}}{\mathrm{TC}-1}\right).$$

For instance, if all candidates have the same MV, then $\mathrm{RRP}_0 = 0$, $\mathrm{RRP}_1 = 1$, and $\mathrm{RRP} = 0.5$. Relative ranking positions have already been used in previous studies [34, 43].

3. Experimental

3.1. **Compounds.** We measured MS and MS/MS spectra from ten organic compounds with masses ranging between 132 and 1218 u, as listed in Table 2 together with the PubChem Compound Identifier (CID), molecular formula β and nominal mass $\bar{m}$. Compounds were obtained from Sigma–Aldrich (St. Louis, MO,, USA). All compound solutions were prepared in MeOH/H_2O (1:1).

3.2. **Instrumentation.** MS and MS^n experiments were performed with a LTQ–Orbitrap (Thermo Fisher Scientific Inc., Bremen, San Jose, CA, USA) mass spectrometer equipped with a Triversa Nanomate (Advion Biosciences Inc., Ithaca, NY, USA). Compound solutions were

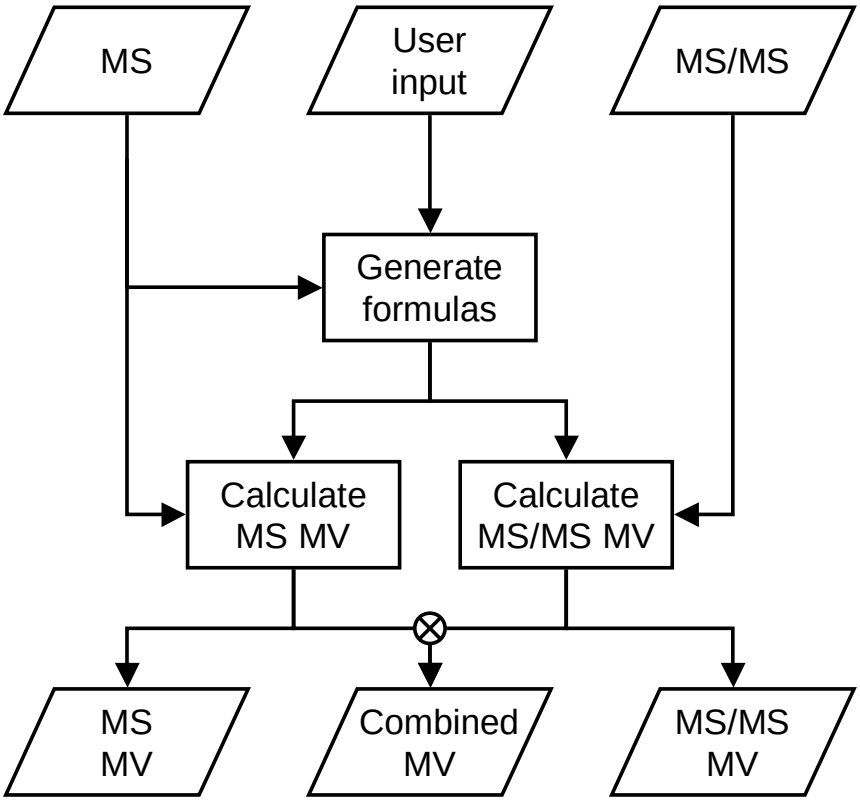

FIGURE 1. Simplified flowchart; see text for details.

infused by a nanospray chip. The mass spectrometer was first calibrated externally with a mixture containing caffeine, L-methionyl-arginyl-phenylalanyl-alanine (MRFA) and Ultramark^{TM}1621 in ACN/MeOH/H_2O/acetic acid. Sub–ppm mass accuracy was finally achieved by using internal calibration (lock mass) in both MS and MS/MS mode. The resolution of Orbitrap MS was set to 100,000 (FWHM) at m/z 400. Protonation was used as ionization mode. For MS/MS experiments, an isolation width of 1.5 u was used. The normalized collision energy was set to the value when the precursor ion was exhausted. Helium was used as the collision gas.

3.3. **Software.** Data analysis with preprocessing, spectrum smoothing and peak detection was performed with in–house software written in Igor 6.0 (Wavemetrics Inc., Portland, OR, USA).

For formula generation, MS and MS/MS matchvalue calculation an advanced version of MOLGEN–MS [44], named MOLGEN–MS/MS [45], was used. MOLGEN–MS/MS is implemented in C++ and is available from www.molgen.de. Scripts for analyzing the results of MOLGEN–MS/MS were written in the programming language R for statistical computing [46].

	MS				MS/MS	
Nr	m	k	$\Delta(m)$	Ion	m_n	n
1	132.07686	2	0.8	$[\mathsf{M}+\mathsf{H}]^+$	90.05521	3
2	225.07578	3	0.1	$[\mathsf{M}+\mathsf{H}]^+$	207.06520	10
3	305.03999	4	0.8	$[\mathsf{M}+\mathsf{H}]^+$	262.99306	2
4	331.22670	3	-0.2	$[\mathsf{M}+\mathsf{H}]^+$	271.20570	20
5	369.35160	3	0.1	$[\mathsf{M}-\mathsf{C_4H_7O_2}]^+$	243.21080	30
6	524.26460	4	-0.7	$[\mathsf{M}+\mathsf{H}]^+$	271.12240	20
7	609.27979	4	-1.4	$[\mathsf{M}+\mathsf{H}]^+$	448.19640	8
8	615.40394	4	0.3	$[\mathsf{M}+\mathsf{H}]^+$	561.37338	7
9	851.26503	4	1.3	$[\mathsf{M}+\mathsf{Na}]^+$	527.15762	11
10	1240.83252	5	5.7	$[\mathsf{M}+\mathsf{Na}]^+$	1200.83337	20

TABLE 3. Overview of the spectra; m: m/z of the MS basepeak; k: number of MS peaks; $\Delta(m)$: relative mass deviation in ppm; Ion: type of ion; m_n: highest m/z in the MS/MS; n: number of MS/MS peaks.

Figure 1 sketches a simplified flowchart of MOLGEN-MS/MS. MS and MS/MS data are passed to the program as peak lists. The user can specify various parameters that were introduced in Section 2, as well as ionization type, adducts, etc. (see [45] for details). If not specified otherwise by the user, the mass of the base peak in the MS is used as starting point for the generation of candidate formulas as described in Subsection 2.1. For each formula generated, MS and MS/MS matchvalues are calculated as described in Subsections 2.2 and 2.3. In addition to these two matchvalues, a multiplicatively combined matchvalue is sent to the output for each candidate formula.

4. Results and Discussion

After calibration, the observed peak masses generally differed from theoretical values by a relative deviation of less than 1 ppm. Only with the larger mass compounds reserpine, maltopentaose and cyclosporin C larger mass errors were observed, but still below 2 ppm for reserpine and maltopentaose. Only for the molecule with the highest molecular mass, cyclosporin C, the relative deviation was relatively high at 5.7 ppm.

In most instances we observed the molecule with a proton adduct in the MS spectrum. In two instances, maltopentaose and cyclosporin C, only the Na^+ adduct appeared in the spectrum. With cholesteryl butyrate, spontaneous fragmentation occurred with a $\mathsf{C_4H_7O_2}$ split–off, so that no intensity at the mass of the original molecule was observed. Consequently, we continued the fragmentation with this peak.

Nr	Candi-dates	NDP		NSAE		NSSE	
		ARP	RRP	ARP	RRP	ARP	RRP
1	1	1	–	1	–	1	–
2	5	2	0.25000	2	0.25000	2	0.25000
3	45	20	0.43182	24	0.52273	28	0.61364
4	8	2	0.14286	2	0.14286	2	0.14286
5	2	1	0.00000	1	0.00000	1	0.00000
6	123	30	0.23771	34	0.27049	34	0.27049
7	318	3	0.00631	6	0.01577	2	0.00316
8	112	40	0.35135	39	0.34234	38	0.33333
9	2699	3	0.00074	4	0.00111	4	0.00111
10	135611	1693	0.01248	1255	0.00925	1347	0.00993
Mean	–	–	0.15926	–	0.17273	–	0.18050

TABLE 4. Numbers of candidates and ranking results for various MS matchvalues; NDP: normalized dot product; NSAE: normalized sum of absolute errors; NSSE: normalized sum of squared errors; ARP: Absolute ranking position; RRP: relative ranking position.

Table 3 offers an overview of these initial results: the observed mass m in the MS and the number of peaks k in the corresponding peak group, relative mass deviations $\Delta(m)$, the type of ion used for the MS/MS, as well as the highest mass m_n in the MS/MS and the number of peaks n in the MS/MS.

Due to the observed mass deviations, we executed calculations with two mass accuracies δ: 2 ppm for samples with molecular mass below 1000 u and 10 ppm for all samples including cyclosporin C.

4.1. **Results obtained by accurate mass and isotopic peaks.** Next, we examined which of the matchvalues for MS isotope peak comparison yields best results. Table 4 shows numbers of candidates bearing at least one carbon atom and absolute ranking positions of the true formulas. Calculations are based on $\mathcal{E}_8$ and MS accuracy $\delta = 10$ ppm. Numbers in the row headers refer to Table 2. We see that all three methods have insufficiencies, especially for samples chloropropoxy-9H--thioxanthen-9-one and those of molecular weight above 500 u. But also for the other samples, there are just two cases where ARP 1 was reached. For the sample of lowest mass, creatine, there is just one candidate in the particular mass interval, and thus ARP 1 is a trivial result. For cholesteryl butyrate there are only two candidates and that the correct one is ranked best is not surprising.

	$\varepsilon = 2$ ppm			$\varepsilon = 5$ ppm			$\varepsilon = 10$ ppm		
Nr	BC	EC	RRP	BC	EC	RRP	BC	EC	RRP
2	0	0	0.00000	0	0	0.00000	0	1	0.12500
3	0	6	0.06818	0	15	0.17046	0	28	0.31818
4	0	0	0.00000	0	0	0.00000	0	0	0.00000
5	0	1	0.50000	0	1	0.50000	0	1	0.50000
6	0	0	0.00000	0	12	0.04918	0	48	0.19672
7	0	15	0.02366	0	68	0.10726	0	167	0.26341
8	52	20	0.55856	0	11	0.04955	0	46	0.20721
9	0	355	0.06579	0	1051	0.19477	0	2049	0.37973
10	40	619	0.00258	0	547	0.00201	0	1153	0.00425
Mean	–	–	0.13542	–	–	0.11924	–	–	0.22161

TABLE 5. Ranking results for MS/MS matchvalues and various accuracy thresholds ε; BC: number of better candidates; EC: number of equal candidates; RRP: see Table 4.

Starting from the ARP, it is not possible to choose the best method. NDP works best for most of the samples, but NSSE yields the best result for maltopentaose, and NSAE is best for cyclosporin C.

In Table 4 we also listed relative ranking positions for the same settings ($\mathcal{E}_8$, $\delta = 10$ ppm). The smallest sample, creatine, is excluded, because the RRP is not defined if only one candidate exists. The arithmetic mean gives a meaningful value for the RRP. It indicates that on average NDP performs best for these ten samples. However, it also shows that the other methods deliver quite similar mean results, and considering the small number of samples, this conclusion is by no means a statistically reliable result. But, when comparing matchvalues based on the MS/MS with matchvalues from MS isotopic peaks, it is a good suggestion to choose the best method on average as reference.

4.2. **Improvements achieved by using MS/MS data.** Next, we computed simple MS/MS matchvalues MV_1 for $\varepsilon = 2, 5$ and 10 ppm (cf. Subsection 2.3). Table 5 shows ranking results in terms of better candidates (BC) and equal candidates (EC). Because MV_1 is just a fraction of integer values with fixed denominator, it happens quite often that EC > 0. However, it is remarkable that for $\varepsilon = 5$ and 10 ppm, in none of the examples a false candidate yields a better MS/MS MV than the true formula. The best mean RRP (0.11924) is obtained for $\varepsilon = 5$ ppm, and it is clearly better than the best mean RRP obtained by MS match values (0.15926 – 0.18050).

Further improvement is gained by combining the matchvalues. We combined them in three ways, by multiplying the MS MV (NDP) and

Nr	$\varepsilon = 2$ ppm			$\varepsilon = 5$ ppm			$\varepsilon = 10$ ppm		
	BC	EC	RRP	BC	EC	RRP	BC	EC	RRP
2	0	0	0.00000	0	0	0.00000	0	0	0.00000
3	2	0	0.04546	6	0	0.13636	12	0	0.27273
4	0	0	0.00000	0	0	0.00000	0	0	0.00000
5	0	0	0.00000	0	0	0.00000	0	0	0.00000
6	0	0	0.00000	4	0	0.03279	15	0	0.12295
7	0	0	0.00000	2	0	0.00631	2	0	0.00631
8	59	0	0.53153	4	0	0.03604	20	0	0.18018
9	1	0	0.00037	2	0	0.00074	2	2	0.00111
10	111	0	0.00082	94	0	0.00069	207	0	0.00153
Mean	–	–	0.06424	–	–	0.02366	–	–	0.06498

TABLE 6. Ranking results for combined matchvalues and various accuracy thresholds ε; see Table 5 for the meanings of abbreviations.

the MS/MS MV, and by taking their arithmetic and geometric mean. These three ways of calculating a combined MV performed virtually identical. Table 6 shows the ranking results for the multiplicatively combined MV. Note that the obtained mean RRP of 0.02366 is again better than using either MS or MS/MS data exclusively. Compared to the NDP this is an improvement by a factor 8 in terms of the RRP. Figure 2 illustrates these improvements in the RRP.

4.3. **Detailed results and refined calculations.** In the following section we discuss results obtained from the ten samples in detail. Unless mentioned otherwise, candidate formulas are generated using an MS accuracy of $\delta = 2$ ppm, for MS MV we used NDP, MS/MS MV is MV_1 with $\varepsilon = 5$ ppm and the combined MV is calculated multiplicatively.

For creatine, there exists only one possible formula within the mass window even when including all eleven elements of $\mathcal{E}_{11}$. Here, mass accuracy alone is sufficient to determine the molecular formula and no additional examinations of isotopic abundances or MS/MS data would be required. The table below shows the DBE, the calculated mass m' of the molecular ion, the mass deviation $\Delta(m)$ in ppm, the MS and MS/MS MV, as well as the combined MV in % for the only candidate.

Candidate	DBE	m'	$\Delta(m)$	Matchvalue in %		
				MS	MS/MS	combined
$C_4H_9N_3O_2$	2.0	132.07675	0.8	99.956	100.000	99.956

This small example is well suited to demonstrate the calculation of the MS/MS MV. For each of the three peaks in the MS/MS there exists

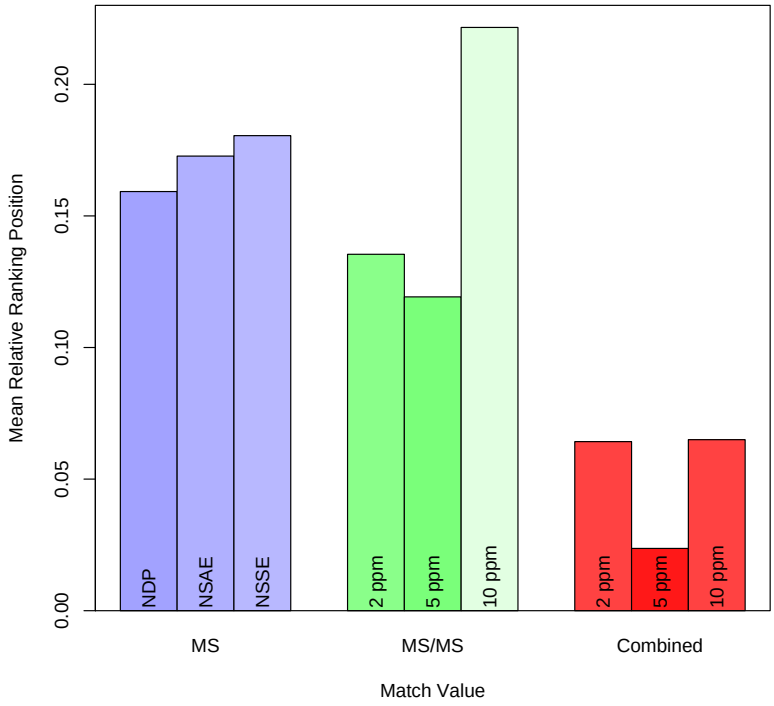

FIGURE 2. Bar chart of mean RRP for different match-values. Combined values are obtained by multiplying NDP with the MS/MS matchvalues. See text for details.

m_i	Subformula	m'_j	$\Delta(m_i)$
90.05521	$C_3H_8NO_2$	90.05496	-2.8
114.06631	$C_4H_8N_3O$	114.06619	-1.1
132.07677	$C_4H_{10}N_3O_2$	132.07675	-0.1

TABLE 7. Calculation of the MS/MS MV for creatine; m_i: m/z of the MS/MS peaks; m'_j: calculated ion masses for subfomulas; $\Delta(m_i)$: mass deviations.

a subformula of $C_4H_9N_3O_2$ with mass deviation $|\Delta(m_i)| \leq 5$ ppm, see Table 7. Here, all MS/MS peaks can be explained and the MS/MS MV has the maximum value of 100%.

For the second sample, sinapinic acid, Table 8 shows all candidate formulas based on elements of $\mathcal{E}_{11}$. There are four candidates with higher MS MV than the true formula $C_{11}H_{12}O_5$. According to the MS/MS MV the true formula is best among these candidates. Only one false candidate, $C_{10}H_{16}O_2Si_2$, yields the same MS/MS MV of 70%. However, this false candidate has a lower MS MV, and accordingly the true formula is ranked first in terms of the combined MV.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{11}H_{12}O_5$	6	0.1	99.838	70.000	69.886
$C_{10}H_{16}O_2Si_2$	5	-1.7	98.647	70.000	69.053
$C_3H_{12}N_6O_4Si$	2	-1.9	99.844	60.000	59.906
$C_9H_{14}O_2F_2S$	2	1.1	99.807	60.000	59.884
$C_7H_{18}F_2Si_3$	1	0.3	98.108	30.000	29.432
$C_7H_8N_4F_4$	4	-0.0	99.969	20.000	19.994
$C_5H_5N_{10}F$	8	1.1	99.969	20.000	19.994
$C_6H_9N_8P$	7	-1.2	99.959	10.000	9.996
$C_5H_{17}N_4SiPS$	1	1.9	99.516	10.000	9.952

TABLE 8. Molecular formula candidates for sinapinic acid; see text for details.

m_i	Candidate $C_{11}H_{12}O_5$		Candidate $C_{10}H_{16}O_2Si_2$	
	Subformula	$\Delta(m_i)$	Subformula	$\Delta(m_i)$
147.0442	$C_9H_7O_2$	-1.0	$C_9H_7O_2$	-1.0
155.0704	$C_8H_{11}O_3$	-0.8	$C_7H_{15}Si_2$	1.8
175.0390	$C_{10}H_7O_3$	-0.2	$C_9H_{11}Si_2$	2.2
181.0860	$C_{10}H_{13}O_3$	-0.4	$C_9H_{17}Si_2$	1.8
183.0653	$C9H_{11}O_4$	-0.6	$C_8H_{15}OSi_2$	1.6
207.0652	$C_{11}H_{11}O_4$	-0.1	$C_{10}H_{15}OSi_2$	1.9
225.0758	$C_{11}H_{13}O_5$	-0.2	$C_{10}H_{17}O_2Si_2$	1.6

TABLE 9. Calculation of the MS/MS MV for sinapinic acid; see text for details.

In order to confirm the true candidate using the MS/MS data alone, it is instructive to observe the peak explanations for the two most likely candidates. Table 9 shows the peak positions m_i of the MS/MS and the subformulas of candidates $C_{11}H_{12}O_5$ and $C_{10}H_{16}O_2Si_2$ explaining the MS/MS peaks. Additionally relative deviations $\Delta(m_i)$ are given in ppm.

It is obvious that deviations for subformulas of $C_{11}H_{12}O_5$ are in general smaller than those for subformulas of $C_{10}H_{16}O_2Si_2$. With smaller mass deviations $\varepsilon = 1$ ppm and $\sigma = 2$ ppm and a fuzzy MS/MS MV as proposed in Subsection 2.3.1, the true formula ranks even better: MV is still 70% for $C_{11}H_{12}O_5$, but only 38.654% for the false candidate $C_{10}H_{16}O_2Si_2$.

In this example however, three unexplained MS/MS peaks remain, at m/z 178.0581, 210.0264 and 224.0635. Even if we allowed a mass

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS^3	combined
$C_{16}H_{13}O_2SCl$	10	0.8	98.356	100.000	98.356
$C_8H_9N_6O_5Cl$	7	1.4	99.336	95.032	94.402
$C_5H_{12}N_4O_9S$	2	0.7	98.942	60.000	59.365
$C_9H_{13}N_6S_2Cl$	6	-1.5	98.157	60.000	58.894
$C_{13}H_8N_2O_7$	11	-1.4	98.102	60.000	58.861
$C_{12}H_{17}OI$	4	1.0	97.919	60.000	58.752
$C_{13}H_{13}N_4Br$	9	1.2	84.174	20.365	17.142
$CH_9N_{12}O_3SCl$	3	-0.9	98.884	0.000	0.000

TABLE 10. Molecular formula candidates for chloropropoxy-9H-thioxanthen-9-one. MS^3 data was used for the combined MV. See text for details.

deviation of 20 ppm for the MS/MS, only one more peak could be explained, the peak at 224.0635 u by $C_{11}H_{12}O_5$ with a deviation of 19.8 ppm. The three unexplained MS/MS peaks might stem from instrumental noise, leaking of molecules or poor isolation of the compound in the ion trap before fragmentation.

For chloropropoxy-9H-thioxanthen-9-one there are 111 candidate formulas based on $\mathcal{E}_{11}$, of which 107 have at least one C atom. If we consider elements of $\mathcal{E}_8$ only, eight candidate formulas remain. Since there are only two peaks in the MS/MS, matchvalues calculated from this spectrum are hardly selective: seven out of the eight candidates achieve 100%. Because of this poor fragmentation profile, we also measured the MS^3 of the most intense MS/MS peak at m/z 263, which fragments into five peaks. With this data only three candidates, $C_8H_9N_6O_5Cl$, $C_9H_{13}N_6S_2Cl$ and $C_{16}H_{13}O_2SCl$, have a maximum MV of 100%. If we choose $\varepsilon = 2$ ppm and $\sigma = 4$ ppm, the true candidate $C_{16}H_{13}O_2SCl$ is ranked at first position. Table 10 shows the results for this setup and the additional restriction of a maximum DBE excess of 3 (according to Subsection 2.3.3). MS MV only, ranking the true formula at topmost position would not have been possible.

For the fourth sample, testosterone acetate, there are only two formula candidates within the assumed mass region, if we consider elements of $\mathcal{E}_8$. Even with the more extensive set $\mathcal{E}_{11}$, the true formula is ranked at first position using the combined MV. Results are shown in Table 11 where the MS/MS MV are calculated in two ways, with and without weighting by peak intensities (see Subsection 2.3.2). With weighting by peak intensities, the separation of the true formula from false candidates is even more evident.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS not weighted	MS/MS weighted
$C_{21}H_{30}O_3$	7	-0.2	99.955	100.000	100.000
$C_{13}H_{30}N_6O_2Si$	3	-1.6	99.931	55.000	59.018
$C_{18}H_{32}OF_2Si$	3	1.1	99.822	95.000	57.796
$C_{11}H_{28}N_6O_3F_2$	0	1.0	99.836	90.000	51.771
$C_{19}H_{32}F_2S$	3	0.4	99.899	85.000	50.495
$C_{20}H_{34}Si_2$	6	-1.5	98.992	85.000	50.495
$C_6H_{26}N_{12}O_4$	0	-1.7	99.598	20.000	34.480

TABLE 11. Molecular formula candidates for testosterone acetate; see text for details.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{23}H_{37}N_7O_5S$	9	-0.7	99.616	100.000	99.616
$C_8H_{33}N_{19}O_6S$	2	-1.7	99.705	94.144	93.867
$C_{15}H_{33}N_{13}O_8$	6	-0.4	99.978	90.000	89.980
$C_{18}H_{38}N_{11}O_3SCl$	5	0.9	93.706	92.275	86.467
$C_{31}H_{33}N_5O_3$	18	-1.9	99.194	79.096	78.458
$C_{23}H_{45}N_3O_4S_3$	3	0.2	98.666	75.000	73.999
$C_{10}H_{34}N_{17}O_6Cl$	2	1.3	95.103	77.160	73.381
$C_{15}H_{41}N_9O_7S_2$	0	0.5	99.488	58.775	58.474
$C_{11}H_{38}N_{17}OS_2Cl$	1	-0.4	92.711	53.516	49.615
$C_{11}H_{30}N_{21}O_2Cl$	7	-1.3	95.301	50.396	48.028
$C_{26}H_{34}N_9OCl$	14	-0.3	94.285	50.009	47.151
$C_{22}H_{41}N_3O_9S$	4	1.9	99.699	47.027	46.886

TABLE 12. Molecular formula candidates for peptide MRFA; see text for details.

Due to the loss of the $C_4H_7O_2$ in the MS, cholesteryl butyrate needs some special treatment. There is only one formula within the considered mass window. This formula represents the remaining fragment $C_{27}H_{44}$. All peaks in the MS/MS can be explained, and we have a MV of 100% (see table below).

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{27}H_{44}$	6	0.1	99.933	100.000	99.933

The sixth sample is a peptide, Met–Arg–Phe–Ala (MRFA). There are 318 formula candidates based on $\mathcal{E}_{11}$, with all but one having at

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS with OEI	MS/MS without OEI
$C_{33}H_{40}N_2O_9$	15	-1.4	99.959	100.000	100.000
$C_{15}H_{28}N_{24}O_4$	14	-0.0	99.215	100.000	59.443
$C_{30}H_{32}N_{12}O_3$	21	0.8	99.982	81.559	47.567
$C_{17}H_{40}N_{10}O_{14}$	3	-0.1	98.904	6.418	6.418
$C_{45}H_{36}O_2$	28	1.6	99.452	4.768	4.768

TABLE 13. Molecular formula candidates for reserpine; see text for details.

least one C atom. 22 remain if we restrict our computations to $\mathcal{E}_8$. With the standard setup, the true formula yields MS/MS MV 100%, along with six false candidates. If we set $\varepsilon = 2$ ppm and $\sigma = 4$ ppm, the true formula is the only candidate ranked at first position with respect to the MS/MS MV. Table 12 shows the best twelve candidates in terms of the combined MV, where the MS/MS MV was computed with a maximum excess of 3 for DBE of ions, and without radical ions. Note how well the true formula can be distinguished from the false candidate $C_{22}H_{41}N_3O_9S$, which has even a better MS MV. In this example all MS/MS peaks can be explained for the true formula, even if we use an accuracy of $\varepsilon = 1$ ppm. The separation of the true formula from false candidates by means of the MS/MS MV is even more convincing in this case.

For the reserpinesample, we get 318 candidate formulas based on $\mathcal{E}_8$, reduced to five if we consider elements of $\mathcal{E}_4$ only. MS/MS results in Table 13 are obtained with $\varepsilon = 2$ ppm, $\sigma = 4$ ppm and weighting by peak intensity. Once radical ions (OEI) are taken into account, and once radical ions are ignored. Excluding radical ions is quite useful in this example, because this way the false candidates $C_{15}H_{28}N_{24}O_4$ and $C_{30}H_{32}N_{12}O_3$ achieve clearly lower MS/MS MV, while for the true formula ignoring radical ions does not affect the MS/MS MV.

In Table 14 the subformulas of $C_{33}H_{40}N_2O_9$ and $C_{30}H_{32}N_{12}O_3$ explaining the peaks in the MS/MS are shown, together with their charge and electron configuration. For the true formula, all listed subformulas belong to even electron ions, whereas several subformulas of the false candidate are odd electron ions, i.e radical ions. When radical ions are rejected, the subformulas $C_{22}H_{18}N_5O$, $C_{19}H_{22}N_{11}O_2$ and $C_{20}H_{22}N_{11}O_2$ cause a worse MS/MS MV because they have higher deviations from the measured masses.

Again, it is noteworthy that the false candidate $C_{30}H_{32}N_{12}O_3$ has a higher MS MV than the true formula. Here, the MS/MS MV is an important tool to exclude false candidates.

	Candidate $C_{33}H_{40}N_2O_9$		Candidate $C_{30}H_{32}N_{12}O_3$	
m_i	Subformula	$\Delta(m_i)$	Subformula	$\Delta(m_i)$
236.1282	$C_{13}H_{18}NO_3^+$	-0.3	$C_{13}H_{18}NO_3^+$	-0.3
365.1860	$C_{22}H_{25}N_2O_3^+$	-0.1	$C_{22}H_{25}N_2O_3^+$	-0.1
			$C_{20}H_{23}N_5O_2^{+\bullet}$	-3.8
368.1493	$C_{21}H_{22}NO_5^+$	-0.1	$C_{20}H_{16}N_8^{+\bullet}$	-0.2
			$C_{22}H_{18}N_5O^+$	3.5
397.2120	$C_{23}H_{29}N_2O_4^+$	0.5	$C_{21}H_{27}N_5O_3^{+\bullet}$	-2.9
			$C_{24}H_{25}N_6^+$	3.8
436.1965	$C_{22}H_{30}NO_8^+$	0.2	$C_{21}H_{24}N_8O_3^{+\bullet}$	0.2
			$C_{19}H_{22}N_{11}O_2^+$	-2.9
448.1964	$C_{23}H_{30}NO_8^+$	0.4	$C_{22}H_{24}N_8O_3^{+\bullet}$	0.4
			$C_{20}H_{22}N_{11}O_2^+$	-2.6
577.2537	$C_{32}H_{37}N_2O_8^+$	1.3	$C_{29}H_{29}N_{12}O_2^+$	-1.0
609.2798	$C_{33}H_{41}N_2O_9^+$	1.4	$C_{30}H_{33}N_{12}O_3^+$	-0.8

TABLE 14. Calculation of the MS/MS MV for reserpine; see text for details.

For the eighth sample, CHAPS, 22 formula candidates are generated based on $\mathcal{E}_8$. MS/MS results in Table 15 were computed with $\varepsilon = 4$ ppm, $\sigma = 5$ ppm, peak weighting by intensity, a maximum excess in DBE of 4 and without taking radical ions into account. We see that with these particular settings the true formula is again ranked first. However, in this example it is hard to obtain a clear result. The two best candidates are very similar, and the MS/MS peaks are explained by exactly the same subformulas for these two formula proposals.

Figure 3 shows the measured MS/MS and MS (inset) of CHAPS. Subformulas of the true formula explaining the MS/MS peaks are attached to the peaks. For the peak at m/z=548.0492, no suitable subformula with mass deviation below 5 ppm was found. The gray boxes of the inset represent the calculated isotopic distribution for the true formula (normalized to 100 %). We can see that for the isotopic peaks the measured intensities are below the calculated abundances, which is the reason for the relatively low MS MV. Figure 4 graphically illustrates MS and MS/MS matchvalues as listed in Table 15. The dots each stand for one formula candidate with the MS MV corresponding to the location on the horizontal and the MS/MS MV on the vertical axis, so that the best matches should appear in the upper right corner. The correct formula of CHAPS, indicated by an arrow, is not furthest on right, reflecting its lower MS match, but is closest to the top because of the good MS/MS match. Note that several candidates with higher MS MV can be excluded due their low MS/MS MV, e.g. $C_{25}H_{50}N_{12}O_6$, $C_{24}H_{54}N_8O_{10}$ and $C_{24}H_{59}N_{10}I$. In contrast, excluding $C_{25}H_{58}N_8O_5S_2$ due

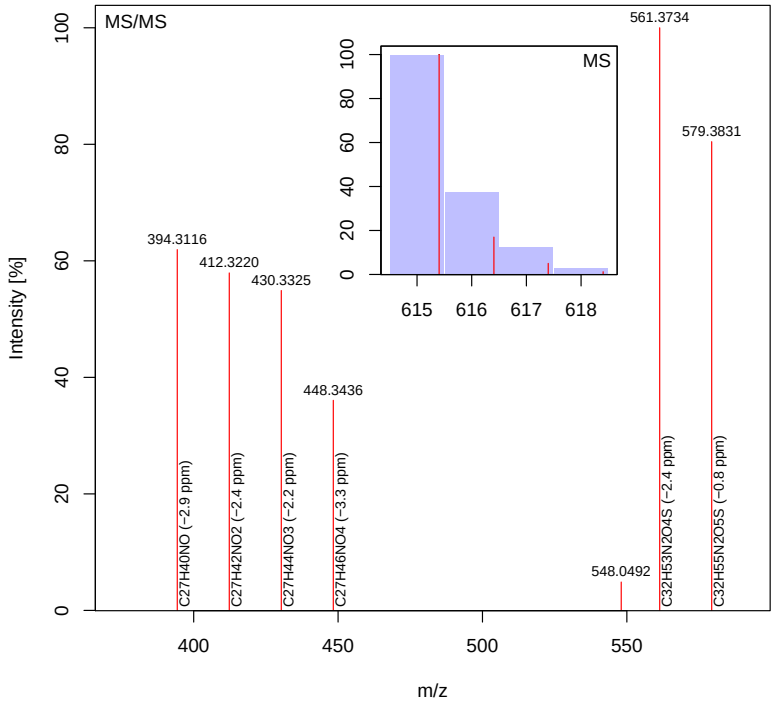

FIGURE 3. MS/MS and MS of CHAPS, together with the calculated isotopic distribution of CHAPS (blue shading).

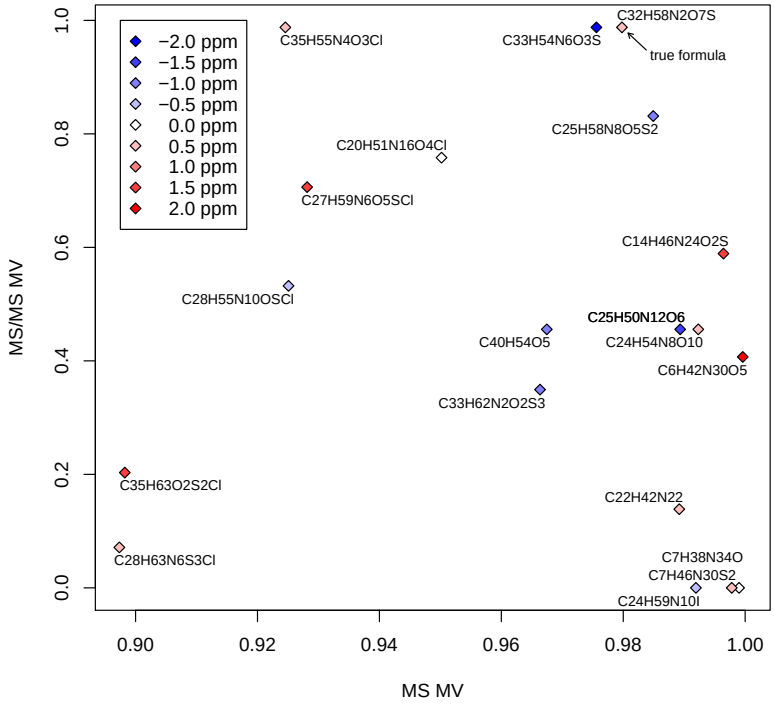

FIGURE 4. Plot of molecular formula candidates for CHAPS; see text for details.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{32}H_{58}N_2O_7S$	5	0.3	97.979	98.771	96.775
$C_{33}H_{54}N_6O_3S$	10	-1.9	97.560	98.771	96.361
$C_{35}H_{55}N_4O_3Cl$	10	0.6	92.457	98.771	91.320
$C_{25}H_{58}N_8O_5S_2$	1	-0.8	98.492	83.139	81.885
$C_{20}H_{51}N_{16}O_4Cl$	3	-0.2	95.019	75.808	72.032
$C_{27}H_{59}N_6O_5SCl$	1	1.7	92.814	70.617	65.542
$C_{14}H_{46}N_{24}O_2S$	4	1.7	99.644	58.904	58.695
$C_{28}H_{55}N_{10}OSCl$	6	-0.5	92.509	53.217	49.230
$C_{24}H_{54}N_8O_{10}$	2	0.6	99.231	45.554	45.204
$C_{25}H_{50}N_{12}O_6$	7	-1.6	98.934	45.554	45.069
$C_{40}H_{54}O_5$	14	-0.8	96.746	45.554	44.072
$C_6H_{42}N_{30}O_5$	1	2.0	99.961	40.693	40.678
$C_{33}H_{62}N_2O_2S_3$	4	-1.1	96.635	34.932	33.756
$C_{24}H_{59}N_{10}O_3Br$	0	1.9	72.905	42.987	31.340
$C_{35}H_{63}O_2S_2Cl$	4	1.4	89.823	20.310	18.243
$C_{22}H_{42}N_{22}$	13	0.6	98.919	13.860	13.710
$C_{29}H_{61}N_6OCl_3$	1	-0.9	70.509	18.275	12.886
$C_{28}H_{63}N_6S_3Cl$	0	0.3	89.734	7.129	6.397
$C_{24}H_{59}N_{10}I$	0	-0.4	99.192	0.000	0.000
$C_{32}H_{63}N_4SBr$	3	1.6	70.094	0.000	0.000
$C_7H_{46}N_{30}S_2$	0	0.6	99.780	0.000	0.000
$C_7H_{38}N_{34}O$	6	-0.2	99.900	0.000	0.000

TABLE 15. Molecular formula candidates for CHAPS; see text for details.

to its MS/MS MV can be considered risky. In the upper right corner, another candidate, $C_{33}H_{54}N_6O_3S$ appears close. For identification of the molecular formula of an unknown, all two (or even three) candidates would have to be considered, but $C_{32}H_{58}N_2O_7S$ matches the measurements marginally better in terms of MS mass.

For the ninth example, maltopentaose, we obtain 538 formula candidates based on $\mathcal{E}_8$. This number is reduced to 19 candidates, if we consider elements of $\mathcal{E}_4$ only. In this example, we obtain a very good MS MV for the true formula. This value is best among the 19 candidates. For the MS/MS MV given in Table 16 we used $\varepsilon = \sigma = 2$ ppm, no radical ions, maximum excess in DBE of 3 and peak weighting by intensity. The true formula also achieves the maximal MS/MS MV of 100%, along with several other candidates. Because of the high MS MV, the true formula is ranked alone at the first position according to the combined MV.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{30}H_{52}O_{26}$	5	1.3	99.995	100.000	99.995
$C_{31}H_{48}N_4O_{22}$	10	-0.3	99.979	100.000	99.979
$C_{28}H_{40}N_{14}O_{16}$	16	1.3	99.965	100.000	99.965
$C_{29}H_{36}N_{18}O_{12}$	21	-0.3	99.901	100.000	99.901
$C_{26}H_{28}N_{28}O_6$	27	1.3	99.871	100.000	99.871
$C_{43}H_{44}N_2O_{15}$	23	1.9	99.225	100.000	99.225
$C_{41}H_{32}N_{16}O_5$	34	2.0	98.973	100.000	98.973
$C_{44}H_{40}N_6O_{11}$	28	0.3	98.966	100.000	98.966
$C_{13}H_{36}N_{26}O_{17}$	9	0.7	99.471	97.806	97.289
$C_{27}H_{24}N_{32}O_2$	32	-0.3	99.763	82.058	81.864
$C_{14}H_{32}N_{30}O_{13}$	14	-0.9	99.642	75.788	75.517
$C_{11}H_{24}N_{40}O_7$	20	0.7	99.588	65.879	65.607
$C_{32}H_{44}N_8O_{18}$	15	-1.9	99.917	60.028	59.978
$C_{30}H_{32}N_{22}O_8$	26	-1.9	99.795	60.028	59.905
$C_{45}H_{36}N_{10}O_7$	33	-1.3	98.676	54.026	53.311
$C_{12}H_{20}N_{44}O_3$	25	-0.9	99.703	43.861	43.730
$C_{42}H_{28}N_{20}O$	39	0.3	98.682	37.652	37.156
$C_{16}H_{44}N_{16}O_{23}$	3	-0.9	99.507	27.178	27.044
$C_{57}H_{32}N_8$	46	0.9	96.654	5.244	5.069

TABLE 16. Molecular formula candidates for maltopentaose; see text for details.

Due to its high molecular mass of 1218 u, cyclosporin C is the most difficult test for the scope of our algorithm. This sample is also an outlier in terms of measurement accuracy. While for all other measurements we had relative mass deviations less than 2 ppm in the MS, here the deviation is as high as 5.7 ppm. Figure 5 shows the MS and the MS/MS for this sample.

In order to have the true formula among the formula candidates, we need to generate formulas with $\delta = 10$ ppm, resulting in 225 formula candidates based on the elements of $\mathcal{E}_4$. It is remarkable that in this example the NSAE works much better as MS MV than the NDP (cf. Subsection 2.2.2). For instance, false candidates $C_{47}H_{107}N_{23}O_{14}$ and $C_{46}H_{107}N_{25}O_{13}$ earn better MS MV than the true formula when applying NDP.

Results in Table 17 were obtained with NSAE and the following setup for the MS/MS MV: $\varepsilon = 2$ ppm, $\sigma = 4$, maximum excess in DBE 4, no radical ions, and weighting by mass and logarithmic intensities, i.e. with $f(m, I(m)) = m \cdot \lg(1 + 10^5 \cdot I(m))$ according to Subsection 2.3.2. In the table, only candidates with combined MV greater than 65% are listed. Figure 6 shows all 225 candidates as plots with MS and

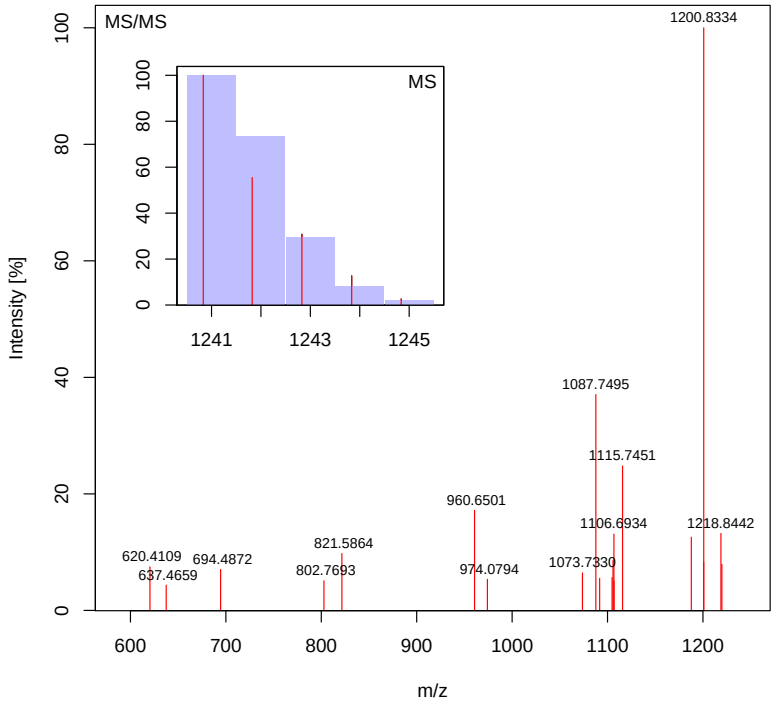

FIGURE 5. MS/MS and MS of cyclosporin C, together with the calculated isotopic distribution of cyclosporin C (blue shading).

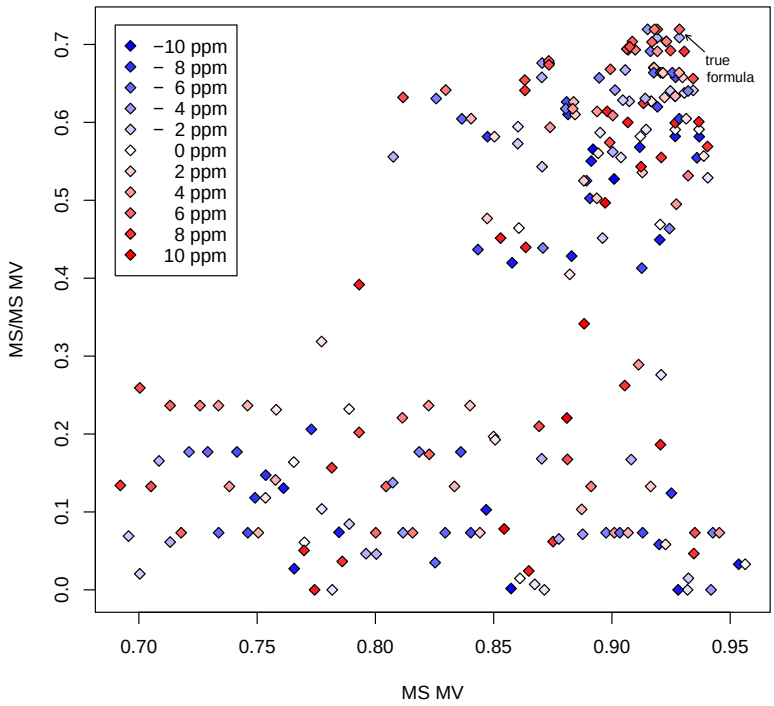

FIGURE 6. Plot of the 225 molecular formula candidates for cyclosporin C; see text for details.

Candidate	DBE	$\Delta(m)$	Matchvalue in %		
			MS	MS/MS	combined
$C_{62}H_{111}N_{11}O_{13}$	13	5.7	92.853	71.938	66.797
$C_{60}H_{99}N_{25}O_3$	24	5.7	91.919	71.938	66.125
$C_{47}H_{107}N_{23}O_{14}$	6	5.3	91.798	71.938	66.038
$C_{46}H_{107}N_{25}O_{13}$	6	-3.9	91.507	71.938	65.829
$C_{61}H_{111}N_{13}O_{12}$	13	-3.5	92.849	70.876	65.808
$C_{59}H_{99}N_{27}O_2$	24	-3.5	91.907	70.834	65.101

TABLE 17. Molecular formula candidates for cyclosporin C; see text for details.

MS/MS MV as coordinates. The true formula is at the top position in Table 17 and in the upper left corner of Figure 6. However, in an experiment for identifying an unknown compound, it would be difficult and somewhat arbitrary to exclude candidate molecule formulas with very similar matchvalues.

4.4. **Scope and Limitations.** A number of difficulties may arise when trying to identify the molecular formula associated with a peak in a spectrum from a biological sample.

Firstly, in contrast to our artificial measurements with pure compounds, real measured samples, for example from body fluids, will be complex mixtures. However, the mass isolation window for the MS/MS stage of current spectrometers is still relatively broad (Orbitrap ≈ 1 u), so that compounds with masses close to the peak of interest will enter the ion trap. Then, during the fragmentation phase these will be fragmented as well, making it impossible to distinguish which fragments originate from which mother ion.

Sensitivity might also become a problem because there is a loss of peak intensity between MS and MS/MS phases. In order to analyze a peak with MS/MS, a certain initial intensity is needed for fragment peaks to appear well above the noise.

Both of these limitations could partly be compensated for by chemical isolation, e.g. through LC separation, to increase signal intensity and prominence. Also, improvements in spectrometer hardware are likely to overcome technical limitations in the foreseeable future.

Compared with NMR spectrometry, which is a powerful method to obtain structural information of molecules, mass spectrometry does not require as high concentrations and purification of the compound of interest. This is unfortunately somewhat offset by the loss of intensity in the MS/MS spectra, but our proposed method delivers structural information complementary to NMR, so that it is useful to apply both spectrometry methods in conjunction.

In theory, our method could be viable for medium-throughput screening for automatic determination of compounds in a sample.

Naturally, our method could also be extended to recursively identify molecular formulas of fragments from MS^n measurements, however, again problems with loss of signal intensity limit this approach for impure compounds.

5. Conclusion

Results presented in Section 4.2 have shown that MS and MS/MS MV provide complimentary information on a compound's molecular formula. Joining these types of information improves the determination of the molecular formula by a factor of 8 (in terms of RRP) compared with ranking by MS MV alone.

In Section 4.3 we have seen that modifications of the MS/MS MV can improve the results further and help to separate the true formula from false candidates. However, more test cases would be required to find which settings are optimal in general.

6. Outlook

It is foreseeable that new spectrometers with higher mass accuracy, resolution and isolation properties will be developed. This would likely overcome some of the limitations of our method and make it possible to narrow down competing molecular formulas candidates to one most likely one.

The big challenge of the future is the identification of the structural formula (constitution) from MS/MS and MS^n. However, the numbers of candidate structural formulas will be many magnitudes higher than candidate molecular formulas [32, 47], and ranking entire constitutional spaces by MS data has been recognized as a serious problem [32, 34, 43]. Nevertheless, encouraging progress in identifying structural formulas and fragments from MS/MS has been made recently [48, 49] and raises hope for further progress.

Acknowledgement

The authors would like to thank Emma Schymanski and Christoph Rücker for reading and improving the manuscript.

References

[1] V. G. Zaikin and J. M. Halket. *Derivatization in mass spectrometry — 8. Soft ionization mass spectrometry of small molecules.* Eur. J. Mass Spectrom., 12(2):79–115, 2006.

[2] A. El-Aneed, A. Cohen, and J. Banoub. *Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers.* Appl. Spectrosc. Rev., 44(3):210–230, 2009.

[3] J. Lederberg. *Rapid Calculation of Molecular Formulas from Mass Values*. J. Chem. Educ., 49(9):613, 1972.

[4] B. V. Ioffe, D. A. Vitenberg, and I. G. Zenkevich. *Calculation of Ion Composition in Organic High-Resolution Mass Spectrometry*. Org. Mass. Spectrom., 28(8):907–913, 1993.

[5] A. G. Marshall and Hendrickson C. L. *High–Resolution Mass Spectrometers*. Ann. Rev. Anal. Chem., 1:579–599, 2008.

[6] T. Kind and O. Fiehn. *Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm*. BMC Bioinf., 7:234–, 2006.

[7] K. F. Blom. *Enhanced Selectivity in Determining Elemental Composition: Concerted Precise Mass and Isotope Pattern Moment Analysis*. Org. Mass. Spectrom., 23(11):783–788, 1988.

[8] A. Tenhosaari. *Computer–Assisted Composition Analysis of Unknown Compounds by Simultaneous Analysis of the Intensity Ratios of Isotope Patterns of the Molecular Ion and Daughter Ions in Low–Resolution Mass Spectra*. Org. Mass. Spectrom., 23:236–239, 1988.

[9] A. Tenhosaari. *Microcomputer Program for Determining Elemental Compositions of Unknown Organic compounds from Low-Resolution Electron Impact Mass Spectrometry*. Chemom. Intell. Lab. Syst., 8:167–171, 1990.

[10] A. Tenhosaari. *Determination of Molecular Formulae from Low-Resolution Mass Spectral Data by Matching Experimental and Calculated Isotope Patterns of Logical Sets of Daughter Ion Candidates*. Anal. Chim. Acta, 248:71–75, 1991.

[11] K. Kumar and A. G. Menon. *Computer–Assisted Determination of Elemental Composition of Fragments in Mass Spectra*. Rapid Comm. Mass Spec., 6:585–591, 1992.

[12] S. G. Roussis and R. Prouix. *Reduction of Chemical Formulas from the Isotopic Peak Distributions of High-Resolution Mass Spectra*. Anal. Chem., 75:1470–1482, 2003.

[13] S. Ojanperä, A. Pelander, M. Pelzing, I. Krebs, E. Vuori, and I. Ojanperä. *Isotopic pattern and accurate mass determination in urine drug screening by liquid chromatography/time-of-flight mass spectrometry*. Rapid Comm. Mass Spec., 20(7):1161–1167, 2006.

[14] S. Böcker, M. Letzel, Z. Liptàk, and A. Pervukhin. *SIRIUS: Decomposing isotope patterns for metabolite identification*. Bioinf., 25(2):218–224, 2009.

[15] S. Heuerding and T. Clerc. *Simple Tools for the Computer–Aided Interpretation of Mass Spectra*. Chemometr. Intell. Lab. Syst., 20(1):57–69, 1993.

[16] T. Kind and O. Fiehn. *Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry*. BMC Bioinf., 8(1):105–, 2007.

[17] Y. Konishi, T. Kiyota, C. Draghici, J.-M. Gao, F. Yeboah, S. Acoca, S. Jarussophon, and E. Purisima. *Molecular Formula Analysis by an MS/MS/MS Technique To Expedite Dereplication of Natural Products*. Anal. Chem., 79(3):1187–1197, 2007.

[18] S. Ashton, R. Gallagher, J. Warrander, N. Loftus, I. Hirano, S. Yamaguchi, N. Mukai, and Y. Inohana. *Isotope modeling routines applied to empirical formula prediction using high mass accuracy MS^n data*. In *23rd LC/MS Montreux Symposium*, Montreux, Switzerland, Nov 8 - 10 2006.

[19] H. M. Shackman, J. M. Ginter, J. P. Fox, and M. Nishimura. *Structural Elucidation by Composition Formula Predictor Software Using MS^n Data*. In *ASMS 2006*, 2006.

[20] J. M. Ginter, J. P. Fox, H. M. Shackman, and R. J. Classon. *Empirical Formula Prediction Using MS and MS^n Spectra and Isotope Modeling*. Chromatography Online.com, 2007.

[21] S. Böcker and F. Rasche. *Towards de novo identification of metabolites by analyzing tandem mass spectra*. Bioinf., 24:I49–I55, 2008.

[22] H. Irth, S. Long, and T. Schenk. *High–Resolution Screening in an Expanded Chemical Space*. Current Drug Discovery, pages 19–23, January 2004.

[23] D. B. Kell. *Metabolomic biomarkers: search, discovery and validation*. Expert Rev. Mol. Diagn., 7(4):329–333, 2007.

[24] X. Zhang, D. Wei, Y. Yap, L. Li, S. Guo, and F. Chen. *Mass spectrometry-based "omics" technologies in cancer diagnostics*. Mass Spectrom. Rev., 26(3):403–431, 2007.

[25] W. Brack, M. Schmitt-Jansen, M. Machala, R. Brix, D. Barceló, E. Schymanski, G. Streck, and T. Schulze. *How to confirm identified toxicants in effect-directed analysis*. Anal. Bioanal. Chem., 390(8):1959–1973, 2006.

[26] H. Irth. *Continuous–Flow Systems for Ligand Binding and Enzyme Inhibition Assays Based on Mass Spectrometry*, pages 185–246. Mass Spectrometry in Medicinal Chemistry. Wiley, 2007.

[27] J. R. De Laeter, J. K. Böhlke, P. De Bièvre, H. Hidaka, H. S. Peiser, K. J. R. Rosman, and P. D. P. Taylor. *Atomic weights of the elements. Review 2000 (IUPAC Technical Report)*. Pure Appl. Chem., 75(6):683–799, 2003.

[28] Scientific Instrument Services, Inc. *Exact Masses and Isotopic Abundances*. http://www.sisweb.com/referenc/source/exactmaa.htm, checked May 2009.

[29] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: The Generation of Structural Formulae and its Applications*. J. Comput. Chem. Jpn., 3:85–96, 2004.

[30] J. K. Senior. *Partitions and Their Representative Graphs*. American Journal of Mathematics, 73(3):663–689, 1951.

[31] R. Grund. *Construction of Molecular Graphs with Given Hybridizations and Non–overlapping Fragments*. Bayr. Math. Schriften, 49:1–113, 1995. In German.

[32] M. Meringer. *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation*. Logos–Verlag Berlin, 2004. In German.

[33] A. Fürst, J. T. Clerc, and E. Pretsch. *A Computer Program for the Computation of the Molecular Formula*. Chemom. Intell. Lab. Syst., 5:329–334, 1989.

[34] A. Kerber, M. Meringer, and C. Rücker. *CASE via MS: Ranking Structure Candidates by Mass Spectra*. Croat. Chem. Acta, 79(3):449–464, 2006.

[35] H. Kubinyi. *Calculation of Isotope Distributions in Mass Spectrometry. A Trivial Solution for a Non-Trivial Problem*. Anal. Chim. Acta, 247:107–109, 1991.

[36] A. L. Rockwood, S. L. Van Orden, and R. D. Smith. *Rapid Calculation of Isotope Distributions*. Anal. Chem., 67:2699–2704, 1995.

[37] A. L. Rockwood and S. L. Van Orden. *Ultrahigh-Speed Calculation of Isotope Distributions*. J. Am. Soc. Mass Spectrom., 68:2027–2030, 1996.

[38] A. L. Rockwood and P. Haimi. *Efficient Calculation of Accurate Masses of Isotopic Peaks*. J. Am. Soc. Mass. Spectr., 17(3):415–419, 2006.

[39] J. Kwiatkowski. *Computergestützte Identifizierung von Isotopenmustern in niederaufgelösten Massenspektren*. Org. Mass. Spectrom., 13(8):513–517, 1978.

[40] K. Varmuza. *Automatische Erkennung von Isotopenpeakmustern in Massenspektren*. Fresenius' Journal of Analytical Chemistry, 322(2):170–174, 1985.

[41] T. Grüner, A. Kerber, R. Laue, M. Liepelt, M. Meringer, K. Varmuza, and W. Werther. *Bestimmung von Summenformeln aus Massenspektren durch*

Erkennung überlagerter Isotopenmuster. MATCH Commun. Math. Comput. Chem., 37:163–177, 1998.

[42] B. Seebass and E. Pretsch. *Automated Compatibility Tests of the Molecular Formulas or Structures of Organic Compounds with Their Mass Spectra*. J. Chem. Inf. Comput. Sci., 39(4):705–717, 1999.

[43] E. L. Schymanski, M. Meringer, and W. Brack. *Matching Structures to Mass Spectra Using Fragmentation Patterns: Are the Results As Good As They Look?* Anal. Chem., 81(9):3608–3617, 2009.

[44] A. Kerber, R. Laue, M. Meringer, and K. Varmuza. *MOLGEN–MS: Evaluation of Low Resolution Electron Impact Mass Spectra with MS Classification and Exhaustive Structure Generation*, volume 15 of *Advances in Mass Spectrometry*, pages 939–940. Wiley, 2001.

[45] M. Meringer. *MOLGEN–MS/MS Software User Manual*. Available from www.molgen.de, March 2009.

[46] R. Ihaka and R. Gentleman. *R: A Language for Data Analysis and Graphics*. J. Comput. Graph. Stat., 5:299–314, 1996.

[47] A. Kerber, R. Laue, M. Meringer, and C. Rücker. *Molecules in Silico: Potential versus Known Organic Compounds*. MATCH Commun. Math. Comput. Chem., 54(2):301–312, 2005.

[48] D. W. Hill, T. M. Kertesz, D. Fontaine, R. Friedman, and D. F. Grant. *Mass Spectral Metabonomics beyond Elemental Formula: Chemical Database Querying by Matching Experimental with Computational Fragmentation Spectra*. Anal. Chem., 80(14):5574–5582, 2008.

[49] M. Heinonen, A. Rantanen, T. Mielikäinen, J. Kokkonen, J. Kiuru, R. A. Ketola, and J. Rousu. *FiD: A software for ab initio structural identification of product ions from tandem mass spectrometric data*. Rapid Comm. Mass Spec., 22(19):3043–3052, 2008.