analytical chemistry

Article

Cite This: *Anal. Chem.* 2019, 91, 3350−3358

pubs.acs.org/ac

Mass Difference Maps and Their Application for the Recalibration of Mass Spectrometric Data in Nontargeted Metabolomics

Kirill S. Smirnov,[†,∥] Sara Forcisi,[†,‡,∥] Franco Moritz,[†] Marianna Lucio,[†] and Philippe Schmitt-Kopplin*[,†,‡,§]

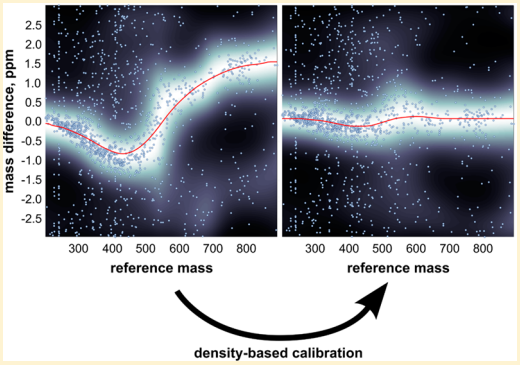
[†]Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

[‡]German Center for Diabetes Research (DZD), Ingolstädter Landstraße 1, 85764 Neuherberg, Germany

[§]Chair of Analytical Food Chemistry, Technische Universität München, Alte Akademie 10, 85354 Freising, Germany

Supporting Information

ABSTRACT: Modern high-resolution mass spectrometry provides the great potential to analyze exact masses of thousands of molecules in one run. In addition, the high instrumental mass accuracy allows for high-precision formula assignments narrowing down tremendously the chemical space of unknown compounds. The adequate values for a mass accuracy are normally achieved by a proper calibration procedure that usually implies using known internal or external standards. This approach might not always be sufficient in cases when systematic error is highly prevalent. Therefore, additional recalibration steps are required. In this work, the concept of mass difference maps (MDiMs) is introduced with a focus on the visualization and investigation of all the pairwise differences between considered masses. Given an adequate reference list of sufficient size, MDiMs can facilitate the detection of a systematic error component. Such a property can be potentially applied for spectral recalibration. Consequently, a novel approach to describe the process of the correction of experimentally derived masses is presented. The method is based on the estimation of the density of data points on MDiMs using Gaussian kernels followed by a curve fitting with an adapted version of the particle swarm optimization algorithm. The described recalibration procedure is examined on simulated as well as real mass spectrometric data. For the latter case, blood plasma samples were analyzed by Fourier transform ion cyclotron resonance mass spectrometry. Nevertheless, due to its inherent flexibility, the method can be easily extended to other low- and high-resolution platforms and/or sample types.



Small molecule profiling by means of high-resolution mass spectrometry (HRMS) is a very active area of research.[1] Thousands of signals can be measured simultaneously, contributing to the sample typology characterization. Moreover, the corresponding MS instruments (comprising time-of-flight and Fourier transformation based mass spectrometers) are often characterized by high mass accuracy which significantly narrows down the search space of putative molecular formula assignments to the experimentally derived masses. In Fourier transform ion cyclotron resonance-mass spectrometry (FTICR-MS), mass accuracy greatly depends upon the measurement physics. Therefore, proper calibration procedures are required in order to obtain spectral signals of high quality.[2] A common solution includes finding constants of the Ledford equation by internal or external calibration scheme. The former procedure assumes the presence of calibrant ions within a sample of interest, whereas in the latter case a separate mixture of known molecules is used. Although this approach can define the form of the calibration function, subtle systematic shifts can still be present impairing future interpretation.[3−5] Therefore, additional recalibration procedures are frequently required.

In nontargeted metabolomics studies there is no *a priori* knowledge of the measured compounds.[6] As a consequence, recalibration may represent a challenging task.[4] Several reports have emphasized that information-rich profiles obtained in HRMS experiments can themselves be a sufficient source for recalibration.[3−5] The corresponding methods depend on the construction of a sufficiently long reference list based on the nature of experimentally derived masses. In this work, an extended strategy is described through the introduction of mass difference maps (MDiMs). This approach is based on the fact that ratios of mass to charge m/z can take on only certain values, which has already been described as a valuable property in developing tools for chemical and biological HRMS-based research.[7] For example, Kendrick mass analysis is often used to reveal sets of compositionally similar molecules by plotting the data in a form of Kendrick mass defect against Kendrick mass, providing an opportunity to assign molecular formulas to each

Received: October 4, 2018
Accepted: February 1, 2019
Published: February 1, 2019

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

member within these sets.[8,9] Another example includes building van Krevelen diagrams, where elemental compositions are represented as ratios (normally, H/C, O/C, and N/C) with their subsequent projection onto the Cartesian coordinate system.[10−13] In this way, it is possible to estimate the degree of saturation of molecules as well as their classification according to the chemical taxonomy (e.g., alcohol, carboxylic acid, etc.). The discretization property is also involved in building so-called mass difference networks (MDiNs) for data derived from FTICR-MS.[14,15] High-resolution and exceptional mass accuracy of this analytical technique allow measuring the distances between detected peaks which, in turn, can correspond to exact masses of certain elemental compositions (i.e., mass differences). Thus, given a set of predefined mass differences, a network can be created for the purpose of molecular formula assignment. From the biological perspective, this concept can be very useful, since the mass differences can represent loss or gain of elements, occurring during enzymatic reactions. Therefore, the network can be interpreted as the map of metabolite transformations.

In turn, MDiM construction involves the calculation of all the pairwise differences between considered masses and plotting them against reference entries. Such a visual representation can lead to highly observable patterns, which can be relatively consistent with the presence of noise. In this work it was shown that MDiMs have an inherent potential to be used for recalibration of mass spectra. No strict prior knowledge on the reference compounds, normally used for this purpose, is necessary. To uncover this potential, a computational workflow for finding the corresponding recalibration curve was developed. It is based on locating a maximum density path using an adapted version of the particle swarm optimization algorithm. The method was tested on a real sample via a nontargeted metabolite profiling of blood plasma using a direct infusion (DI) FTICR-MS. The proposed guideline is not restricted to any particular platform since it is solely based on the discrete nature of exact masses. Therefore, it can be easily applied to any nontargeted metabolomics study using HRMS.

METHODS

Sample Preparation and Analysis Using DI FTICR-MS. The preparation of plasma samples, spiked with standard compounds (Table S-1), and their subsequent analysis by FTICR-MS is described in detail in the Supporting Information.

Theory Behind MDiM Construction. Since the current work concerns nontargeted metabolomics profiling of a biological sample, the focus was narrowed down to compounds containing only the most prevalent elements in biological structures, namely, C (carbon), H (hydrogen), O (oxygen), N (nitrogen), S (sulfur), and P (phosphorus). Despite these limitations, mainly used for descriptive purposes, the workflow is generalizable to any kinds of elements.

The exact/theoretical mass M of a molecule, obeying the aforementioned restrictions, can be calculated by summing up the exact masses m_e of each of the constituent elements:[16]

$$M = \sum_e n_e m_e \tag{1}$$

where $e \in \{C,H,O,N,S,P\}$, $n_e \geq 0$, $m_e > 0$

In eq 1, index e iterates through all the considered elements, whereas n_e stays for the amount of a certain element in the composition of the molecule. As can be seen, eq 1 represents a linear sum, where the values m_e and n_e play roles of constant and variable terms, respectively. Since n_e are non-negative integer numbers and m_e are positive real-valued numbers, the set of possible outcomes for M is discrete and consists of non-negative real-valued numbers. The same applies for m/z values, normally measured in mass spectrometric studies. Therefore, for further considerations, the charge z will be omitted for simplicity.

Considering all the possible combinations of the elements CHONSP, together with computing the values for M (meaning, with no limitations on the upper bound for M or n_e in eq 1), it is possible to calculate all the pairwise differences between the exact masses using "Dalton (Da)" or "parts-per-million (ppm)" metrics:

$$\Delta M_{ij} = M_i - M_j \text{ (in Da)} \tag{2a}$$

$$\Delta M_{ij} = 10^6 \cdot \frac{M_i - M_j}{M_j} \text{ (in ppm)} \tag{2b}$$

In mass spectrometric research, these calculations are normally used while finding the closest match of accurate/measured masses to the exact/theoretical ones.[17] In addition, the estimation of the accuracy of a single measurement as well as mass accuracy (that is an average of accuracies over n measurements) can be done in such a way when reference compounds are available (or when the composition of an experimental mass is known). Such estimations are usually done at different mass ranges, since mass accuracy can vary depending on the value of a measured mass.[18] Equations 2a and 2b can be rewritten as (using eq 1):

$$\Delta M_{ij} = \sum_e (n_{ie} - n_{je}) \cdot m_e \text{ (in Da)} \tag{3a}$$

$$\Delta M_{ij} = \frac{\sum_e (n_{ie} - n_{je}) \cdot m_e}{\sum_e n_{je} m_e} \text{ (in ppm)} \tag{3b}$$

Every ΔM_{ij} describes the difference in the element amounts in the corresponding compositions. Therefore, plotting ΔM_{ij} versus M_j (or M_i) makes it possible to observe the patterns of separated horizontal lines in the case of "Da" metrics or converging curves in the case of "ppm" metrics, built up of data points arranged in a regular manner. This is the basis for the MDiM construction: for each given reference mass, plotting the difference between examined masses and the reference mass. Therefore, the construction of a MDiM requires two arguments, namely, the list of examined masses and the list of reference masses (that can coincide as well).

For all further considerations, the reference list, used for MDiM construction, represented unique masses retrieved from the databases KEGG, HMDB, LIPIDMAPS, and METACYC. The corresponding molecular formulas were restricted to contain only the six aforementioned elements. For the rest of the manuscript, this list will be referred to as the exact mass list. With respect to MDiM construction, we will skip mentioning reference/exact mass list to which other lists of interest (e.g., consisting of experimental masses) were compared (e.g., we will write MDiM was built for experimental masses).

Constructing MDiMs for Simulated and Real Data Sets. To check the consistency of MDiMs, a simulated mass list, corresponding to a fictional mass spectrometric experi-

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

ment, was created in three stages. First, the fraction of entries was randomly chosen from the exact mass list without repetitions. Second, the list was additionally complemented by a fraction of exact masses that were not part of the exact mass list. To create a pool of such additional entries, a similar approach described by Matsuda et al. was followed.[19] The exact mass list was modified and extended by adding and subtracting the exact masses of CH_2 and O, followed by leaving only those entries that were not found in the original nonmodified exact mass list. Thus, these masses can be considered as signals that were detected in a mass spectrometric experiment but were not assigned to any compound in a database. Third, the simulated list was additionally complemented by a certain amount of random numbers following a uniform distribution in the interval from 0 to 1000 Da to represent noise. It is important to mention that in the simulation, the masses in the noncharged form M are considered, whereas a real mass spectrometric study deals with ratios of mass to charge, m/z.[16] Nevertheless, the corresponding modifications can be done without losing consistency. This also applies for multiple-charged ions.

In a mass spectrometric experiment, the error associated with a measurement can be mainly determined by three constituents, namely, random error, systematic error, and gross error:[17]

$$E(M) = E_r(M) + E_s(M) + E_g(M) \tag{4}$$

Further considerations will be focused on the first two terms, since the gross error is associated with undetectable failures or mistakes. The presence of a random noise was simulated by adding a normally distributed term to the entries in the simulated list (except for the uniformly distributed random numbers). Since several mass spectrometric scans can be performed, the standard deviation of an average experimental mass will be equal to a standard deviation of a single measurement divided by the square root of the number of runs:

$$M_i^{\mathrm{mod}} = M_i + N\left(0, \frac{{\sigma_1}^2(M_i)}{N_S}\right) = M_i + \frac{\sigma_1(M_i)}{\sqrt{N_S}} N(0, 1) \tag{5}$$

Here M_i stays for an exact mass from the simulated list and M_i^{mod} stays for the modified version of this mass. We assumed that σ_1 depends on M_i in such a way that the mass measurement error, corresponding to a "ppm" metrics, is normally distributed along the entire considered mass range:

$$\sigma_1(M_i) = a_1 \cdot 10^{-6} \cdot M_i \tag{6}$$

The factor a_1 denotes the standard deviation of a mass measurement error in "ppm" metrics. The presence of a systematic noise was simulated by adding a normally distributed term (with respect to eq 5) centered at $\mu(M_i)$:

$$M_i^{\mathrm{mod}} = M_i + N\left(\mu(M_i), \frac{{\sigma_1}^2(M_i) + {\sigma_2}^2(M_i)}{N_S}\right) \tag{7}$$

The form for σ_2 was chosen to be the same as for σ_1. For both cases (the presence of the random and systematic noise) the corresponding MDiMs for simulated lists were constructed.

In case of a real mass spectrometric experiment, the exact mass list was modified to represent protonated ($[M + H]^+$) as well as sodiated ($[M + Na]^+$) adducts, since the original entries in this list correspond to molecules in the noncharged form. After these modifications, MDiMs were constructed.

Spectral Recalibration with the Maximum Density Path. After plotting a MDiM for the simulated or experimental mass list, the kernel density estimation approach was used to evaluate the closeness of the data points. Gaussian kernels were chosen for this purpose. Initially, in order to apply kernel density estimation, a grid has to be defined. Assuming that (x_i, y_i) represents a coordinate of a data point on a MDiM (x_i is a reference mass and y_i is the corresponding mass difference), the set of points defining a grid of dimensions $H \times V$ were chosen to be equal to

$$x(h) = \min(\{x_i\}) + \frac{\max(\{x_i\}) - \min(\{x_i\})}{H} \cdot h \tag{8a}$$

$$y(v) = \min(\{y_i\}) + \frac{\max(\{y_i\}) - \min(\{y_i\})}{V} \cdot v \tag{8b}$$

In this representation, $1 \le h \le H$ and $1 \le v \le V$. Therefore, the density corresponding to the coordinate ($x(h)$, $y(v)$) is equal to

$$f(x(h), y(v)) = \frac{1}{2\pi n} \sum_{i=1}^{n} e^{\left(-1/2\left(\begin{bmatrix} x(h)-x_i \\ y(v)-y_i \end{bmatrix}^{\mathrm{T}} \mathbf{H} \begin{bmatrix} x(h)-x_i \\ y(v)-y_i \end{bmatrix}\right)\right)} \tag{9}$$

In this equation, $\mathbf{H}$ represents a bandwidth symmetric matrix that was chosen according to the rule of thumb.[20] The density function f was additionally normalized in order to compensate for unequal distribution of data points on the MDiM:

$$f_n(x(h), y(v)) = \frac{f(x(h), y(v))}{\sum_{i=1}^{V} f(x(h), y(i))} \tag{10}$$

After normalization, a curve was fitted by finding a maximum density path. An adapted version of a particle optimization algorithm was used to solve this task.[21] Initially, P particles are created on the grid representing, at the beginning, equally distanced horizontal lines. Therefore, each particle is assigned to its own set of coordinates, e.g., $\{(x(h_{1p}), y(v_{1p})), \ldots, (x(h_{Hp}), y(v_{Vp}))\}$, where p is the particle index ($1 \le p \le P$). The indices h_{ip} were chosen to be equal i, meaning that each particle always goes monotonically through all possible $x(h)$ leaving only v_{ip} indices to vary. At the beginning and every further iteration of the algorithm, a particle/curve was scored by

$$S_p = \sum_{i=1}^{H} f_n(x(i), y(v_{ip})) \tag{11}$$

After scoring, the particle p with the highest S_p was chosen followed by selecting the ($x(i)$, $y(v_{ip})$) with the highest $f_n(x(i), y(v_{ip}))$ corresponding to this curve. This location ($x(i)$, $y(v_{ip})$) served as a trigger toward the change of coordinates for other particles k according to

$$\begin{aligned} &(h_i, v_{ik}) \rightarrow (h_i, v_{ik} + 1) \text{ if } v_{ik} < v_{ip} \\ &(h_i, v_{ik}) \rightarrow (h_i, v_{ik} - 1) \text{ if } v_{ik} > v_{ip} \\ &(h_i, v_{ik}) \rightarrow (h_i, v_{ik}) \text{ if } v_{ik} = v_{ip} \end{aligned} \tag{12}$$

These changes followed the rule that the absolute distance between adjacent v-coordinates of a particle (e.g., $|v_{ik} - v_{(i+1)k}|$) cannot exceed 1, implying that there should not be any breaks.

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

Therefore, the adjacent indices modify their values in the direction of the main change if necessary. Afterward, the particles were rescored and the aforementioned steps repeated until convergence or until reaching a predefined number of iterations.

After the final arrangement of particle data points, the particle with the highest score was chosen. For every corresponding $x(i)$, the associated $y(v_i)$ were recalculated by using a simple running average algorithm to smooth the data. The generated set of points served as a recalibrator for adjusting the spectral data. At any given M_{test} between the adjacent x-coordinates $x(i)$ and $x(i + 1)$ an offset ε was estimated by fitting a local line through $(x(i), y(v_i))$ and $(x(i + 1), y(v_i + 1))$ and calculating the following measure:

$$\varepsilon = \frac{y(v_{i+1}) - y(v_i)}{x(i+1) - x(i)} \cdot M_{test} + \frac{y(v_i)x(i+1) - y(v_{i+1})x(i)}{x(i+1) - x(i)} \quad (13)$$

Afterward, the value for M_{test} could be adjusted accordingly:

$$M_{test}^{adj} = M_{test} - \varepsilon \text{ (in Da)} \quad (14a)$$

$$M_{test}^{adj} = \frac{10^6}{10^6 + \varepsilon} \cdot M_{test} \text{ (in ppm)} \quad (14b)$$

Knowing reference compounds, it was possible to evaluate the quality of recalibration by calculating the root mean squared error (RMSE) between observed (simulated or experimental) and exact masses using "ppm" metrics:

$$\text{RMSE} = \sqrt{\frac{1}{N-1} \sum_i \left(10^6 \frac{M_i^{obs} - M_i}{M_i}\right)^2} \quad (15)$$

This choice was determined by similarity of RMSE to the form of standard deviation when the expected value of the error is assumed to be zero. To emphasize this similarity, $N - 1$ was put in the denominator rather than N. In case of simulated data set, RMSE was calculated by comparing the exact mass list to its modified version (after introducing random and systematic noise). In case of a real mass spectrometric study, RMSE was calculated by comparing experimental m/z of the standard compounds to their theoretical values. It is important to mention that the standard compounds (Table S-1) were not considered during the recalibration process in order to avoid possible bias.

The main parameters involved in the aforementioned recalibration procedure include the grid size, the particle amount, and the number of iterations. All these values were selected empirically. It is worth to point out that the adapted version of the particle swarm optimization algorithm does not implement any random behavior as well as does not take into account the memory effect. Therefore, the solution will be the same for identical starting conditions. All the calculations for the aforementioned procedures were done in MATLAB R2015b supplemented with a Curve Fitting Toolbox (The MathWorks Inc., Natick, MA).

RESULTS AND DISCUSSION

In the current work a perspective look on the evaluation of patterns is presented, generated by constructing all the pairwise differences between considered masses. The corresponding MDiMs serve as the basic tool for such examination. Despite the simplicity to construct MDiMs, the observed patterns can possess the information on the structure and quality of mass spectrometric measurements as well as possible pitfalls to be avoided. As a consequence, the aim of the work focused on showing the potential of using MDiMs for recalibration of mass spectrometric data. The corresponding implementation involved a new approach based on kernel density estimation and an adapted version of the particle swarm optimization algorithm.

Mining unique molecular formulas from KEGG, HMDB, LIPIDMAPS, and METACYC databases resulted in 15 677 elemental compositions. Considering only exact masses up to 1000 Da, the space was diminished until 13 763 molecular formulas. This list served as an exact mass list/reference list in constructing MDiMs. Of course, the use of such a list will give only approximate results, since many more compounds with other elemental compositions (and, as a consequence, exact masses) can be added. However, for the current investigation these approximations were sufficient.

First, a MDiM for the exact mass list itself was investigated by using both "Da" as well as "ppm" metrics (Figure 1). As

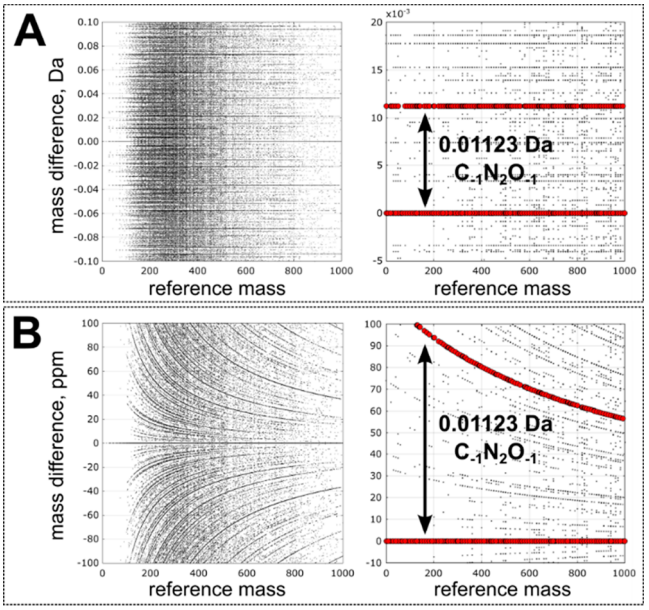


Figure 1. MDiMs for the exact mass list using "Da" (A) and "ppm" (B) metrics. The list of exact masses is obtained by mining unique molecular formulas from KEGG, HMDB, LIPIDMAPS, and METACYC databases.

expected, it is possible to observe a pattern consisting of horizontal lines (Figure 1A) or converging curves (Figure 1B), each representing a certain mass difference determined by the linear combination of the exact masses of considered elements. Both representations (either using "Da" or "ppm" metrics) can be used interchangeably according to the question of interest. It is possible to notice that every line has a different "degree of saturation" with the central line being the most distinct. Due to the way of MDiM construction, this degree emphasizes the prevalence of certain mass differences at specific mass ranges within the exact mass list. Naturally, the central line has the highest degree, since the exact masses are compared to themselves.

Figure 1 suggests that similar patterns can be seen while constructing a MDiM for an experimental mass list. As a partial proof of the concept, a simulation was performed by creating a

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

list of 10 000 entries. In total, 25% (2500 entries) corresponded to randomly selected masses from the exact mass list, 50% (5000 entries) corresponded to randomly selected masses from the modified exact mass list, and 25% (2500 entries) corresponded to random numbers. All the values in the simulated mass list lied within the interval from 0 to 1000 Da. A random noise was added to the exact masses within the list according to the eq 5. The number of scans N_S was assumed to be equal to 100, whereas the standard deviation for a single measurement was set to 10 ppm (factor a_1 in eq 6). The MDiM for the simulated mass list with the presence of the random noise is shown on Figure 2. Although

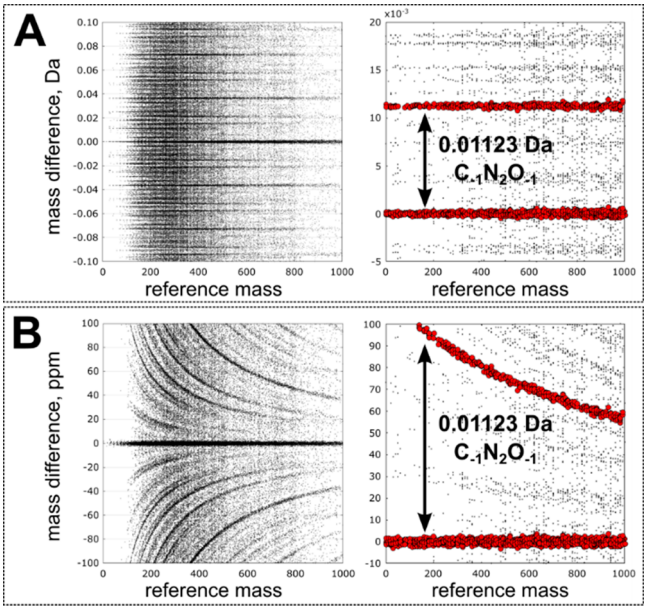


Figure 2. MDiM for the simulated mass list using "Da" (A) and "ppm" (B) metrics. The simulation assumes the presence of a random noise with a standard deviation for one single measurement equal to 10 ppm.

the picture is partially distorted, it is still possible to observe distinct lines/curves. Higher prevalence of a certain mass difference implies better contrast of the corresponding line/curve. It is possible to suggest that MDiMs can be relatively consistent with the presence of a random noise in a mass spectrometric experiment.

However, such experiments can be often characterized by the presence of a systematic noise component.[5] In the simulation experiment this component is characterized by $\mu(M_i)$ and σ_2 from eq 7. The former term was chosen to be

$$\mu(M_i) = -8 \times 10^{-8} \cdot (10^{-3} \cdot (M_i)^3 - (M_i)^2) \quad (16)$$

This corresponds to a parabola for a MDiM using the "ppm" metrics. The curve goes through the coordinates (0 Da, 0 ppm) and (1000 Da, 0 ppm) with a vertex located at (500 Da, 20 ppm). The parabola was chosen only for descriptive purposes rather to simulate a real physics behind mass spectrometric measurements. Moreover, it is a nonlinear function that is simple to portray. The number of scans N_S was assumed to be equal to 100, whereas the standard deviation for a single measurement as well as the parameter a_2 for σ_2 was set to 10 ppm. The MDiM for the simulated mass list with the presence of the random as well as systematic noise is shown on Figure 3.

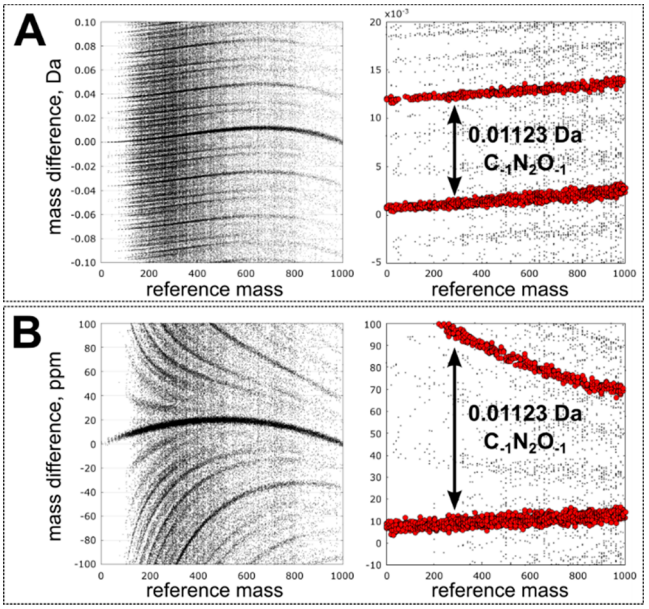


Figure 3. MDiM for the simulated mass list using "Da" (A) and "ppm" (B) metrics. The simulation assumes the presence of the random as well as systematic noise.

The essential difference between Figure 2 and Figure 3 is the changes in the line/curve shapes. However, these changes are coherent among all the lines/curves, meaning that the relative distances between them are preserved. Therefore, it is possible to suggest that any of these lines/curves, when extracted, contains all the necessary information to eliminate the systematic interfering component. Such an observation implies the possibility to use MDiMs for the recalibration of the data obtained in nontargeted metabolomics experiments. Without prior knowledge on specific compounds, present in a sample, it is possible to use a large list of exact masses to construct a MDiM for an experimental mass list, to investigate if there is a presence of a systematic noise component, and to recalibrate the spectral data correspondingly. The only requirement is having an adequate list of exact masses and a sufficient amount of spectral features that can be compared to this list. Similar approaches have been described in proteomics research,[3,5,22] where the spectrometric data was compared to the reference list built of exact masses found by assigning experimental m/z ratios to the theoretical ones. Therefore, the search of a recalibration curve is performed with respect to the proportion of correct assignments. The idea of representing some of the experimental peaks as internal standards for the purpose of recalibration has been also reflected in petroleomics intended to measure molecules of low masses.[4] Compared to the aforementioned approaches, there is no necessity to enrich the assignments by the amount of true positives while constructing MDiMs. The aim is rather to investigate the whole surrounding space and its behavior. Any of the generated lines/curves, presented on MDiMs, can be potentially used for recalibrating spectral data by adding or subtracting the corresponding mass difference and to perform the opposite operation after. It can be especially useful in cases when the central line/curve is not observable, whereas some of the side bends are distinct enough.

Using MDiMs as the basis for recalibration, a novel approach is proposed for finding the corresponding offsets. It is based on the estimation of the density of data points

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

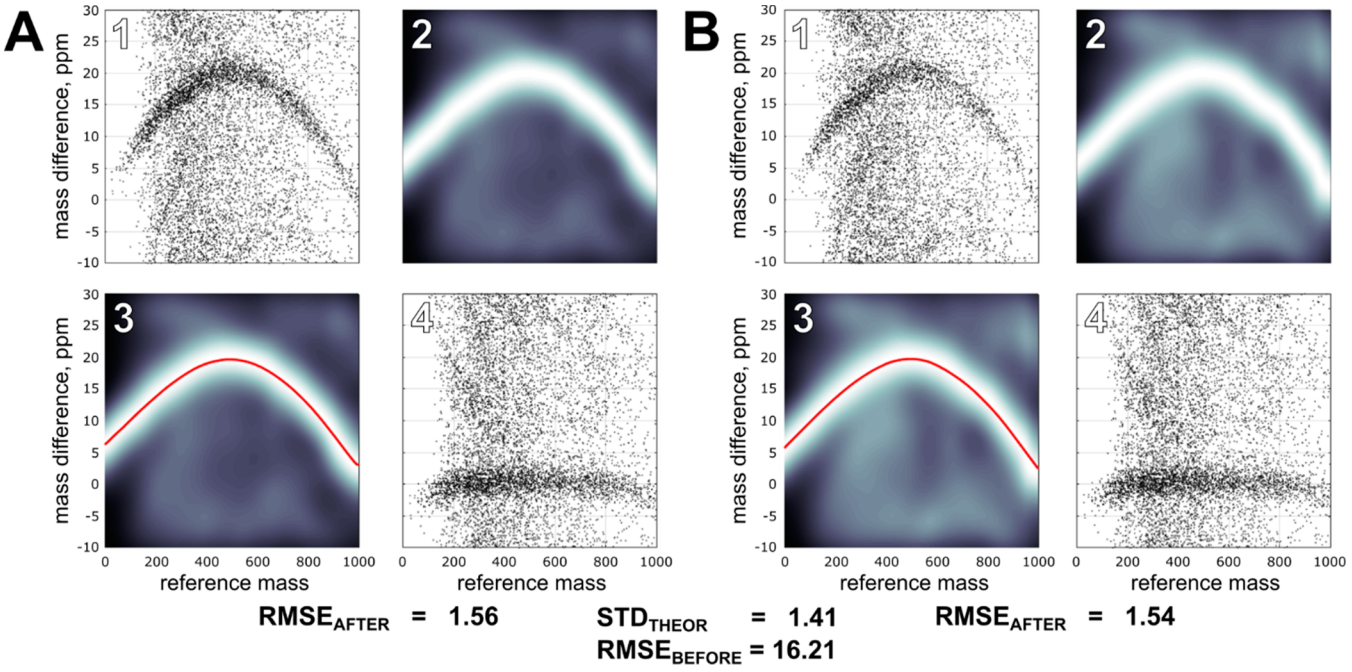


Figure 4. Recalibration workflow for the simulated data (A) in its original form and (B) shifted by the exact mass of H_2. (1) The MDiM for the simulated mass list assuming the presence of the random as well as systematic noise. (2) Kernel density estimation on a grid of points. (3) Resulting curve obtained by running a particle swarm optimization algorithm. (4) The simulated data after recalibration procedure shifted back, if necessary, by the exact mass of H_2. STD_{THEOR} is a theoretical standard deviation. $RMSE_{BEFORE}$ and $RMSE_{AFTER}$ represent root mean squared error before and after calibration.

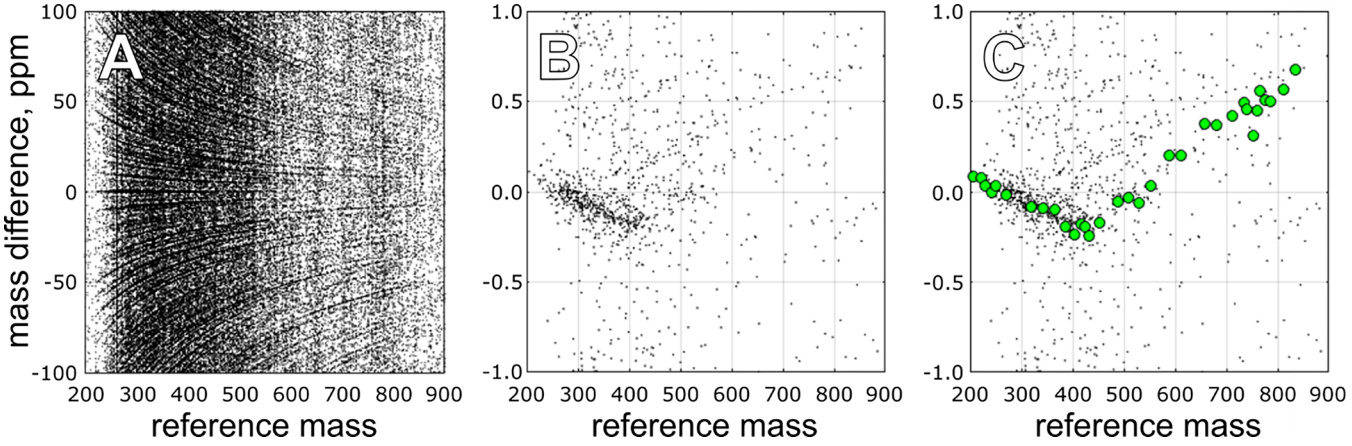


Figure 5. MDiM for the experimental mass list derived from the FTICR-MS analysis of plasma. The "ppm" metrics is used for the representation. (A) Overview of the MDiM in the interval from −100 to 100 ppm. (B) Smaller interval from −1 to 1 ppm emphasizing the behavior of the central line. (C) Projection of the mass error, corresponding to the standard compounds, onto the MDiM.

followed by searching a maximum density path corresponding to the systematic component to be eliminated. Before applying it to a real mass spectrometric experiment, the method is depicted on the simulated data corresponding to Figure 3B (Figure 4).

First, the line/curve that will be used for recalibration needs to be identified. The recalibration was performed using the original MDiM for the simulated mass list (Figure 4A) and the MDiM for the same list but shifted in a positive direction by an exact mass of H_2 (Figure 4B). The latter approach assumes recalibration using the line/curve of the corresponding mass difference. The shift was chosen arbitrary and used for descriptive purposes only. As can be seen, both constructed MDiMs (Figure 4A-1 and Figure 4B-1) are very similar. However, the Figure 4B-1 has a less pronounced pattern. For both scenarios, the kernel density estimation was performed on a 100 × 75 grid (Figure 4A-2 and Figure 4B-2). The region of higher density in the middle of the plots emphasizes the systematic component to be eliminated. An adapted version of the particle swarm optimization algorithm was utilized to create a recalibration curve. For both scenarios, 9 particles were chosen and the number of iterations was set to 1500. Running the algorithm resulted in a curve going through the maximum density path (Figure 4A-3 and Figure 4B-3). This curve was used to shift the masses in the simulated list by the corresponding offset and, if necessary, subtract the initially added exact mass of H_2 (Figure 4A-4 and Figure 4B-4). It is possible to see that the corresponding adjustments resulted in

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

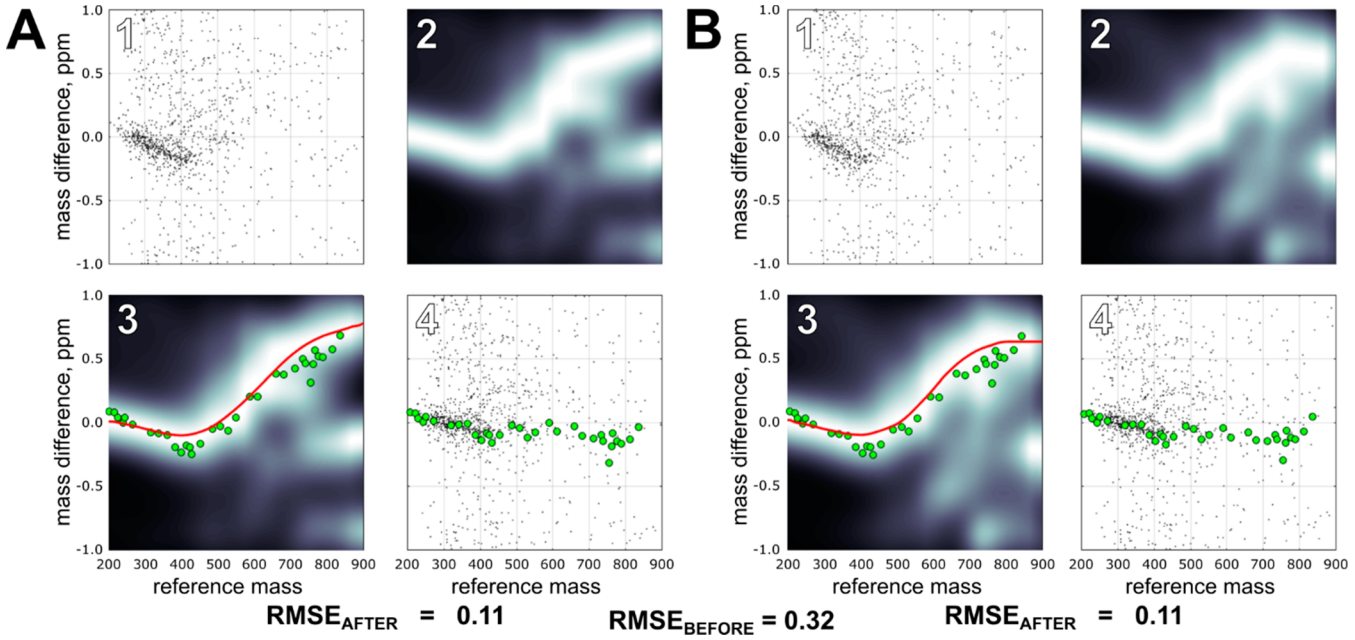


Figure 6. Recalibration workflow for the experimental data (A) in its original form and (B) shifted by the exact mass of H_2. (1) The MDiM for the experimental mass list. (2) Kernel density estimation on a grid of points. (3) Resulting curve obtained by running a particle swarm optimization algorithm together with the data points corresponding to the standard compounds. (4) The experimental data after recalibration procedure shifted back, if necessary, by the exact mass of H_2. $RMSE_{BEFORE}$ and $RMSE_{AFTER}$ represent root mean squared error before and after calibration.

data of better quality with the central line being straighter and more centered around zero. Empirically calculated values of RMSE after the calibration are much closer to the theoretical value of standard deviation (1.56 and 1.54, respectively, versus 1.41) compared to the initial RMSE values. Therefore, it can be concluded that the described method achieved an adequate agreement between the theory and the simulation.

The construction of MDiMs together with the recalibration procedure was performed for the case of a real mass spectrometric nontargeted metabolomics experiment. The analysis of blood plasma samples (spiked with standard compounds, Table S-1) by means of DI FTICR-MS followed by spectral processing resulted in 11 704 signals, 33 of which corresponded to the standard compounds in the protonated and sodiated forms. The signals less than 3×10^6 relative intensity units were excluded from further consideration as well as the signals out of the interval from 200 to 900 *m/z*. Such a filtering resulted in 7442 peaks. The MDiM for the experimental mass list is shown on the Figure 5 in "ppm" metrics.

It is possible to observe a similar pattern of converging curves, seen previously for the simulated experiment (Figure 5A). However, there is much more noise in the lower *m/z* range, thereby the corresponding MDiM is more dense in that region. It is important to mention that, in terms of mass spectrometric measurements, the observed curves are not only associated with distinct mass differences but can represent isobars, i.e., molecules having the same nominal masses relative to the reference compounds of the corresponding masses.[1] Therefore, by constructing MDiMs, we cover an observable isobaric space to certain extent. The examination of such a space enables to carefully define the pattern corresponding to the genuine molecular formula assignments or to which degree false assignments distinguish from the true ones.

The central curve on the MDiM is scarcely observable after 500 *m/z* (Figure 5B) letting only a short decline to be visible in the beginning. Projecting the mass accuracies associated with the standard compounds (Table S-2) onto the constructed MDiM (Figure 5C), it is possible to observe that the low *m/z* values follow quite well the initial trend defined by the central curve. However, after approximately 400 *m/z*, the error shows an opposite behavior.

Such heterogeneity emphasizes the fact that even the application of the external calibration (done in this study) does not help against some systematic component that can impair mass accuracy.[3] The application of the recalibration procedure using kernel density estimation followed by particle swarm optimization algorithm is shown on Figure 6.

As with the workflow for the simulated data presented on Figure 4, the recalibration was performed for the experimental list in its original form (Figure 6A) and shifted by the exact mass of H_2 (Figure 6B). Figure 6A-1 and Figure 6B-1 show the MDiMs for the corresponding lists in the interval from −1 to 1 ppm. As with simulated data, Figure 6B-1 looks similar to Figure 6A-1 but has less pronounced pattern of the central curve. The kernel density estimation was performed on a 100 × 100 grid (Figure 6A-2) and 100 × 50 grid (Figure 6B-2), respectively. Intriguingly, even when no obvious distribution of data points after 500 *m/z* was observed, this representation allows seeing similar trend shown by the standard compounds on Figure 5C. However, there were other regions of high density arising due to the normalization. For the first scenario (Figure 6A), the recalibration procedure was applied using 5 particles and 2000 iterations, whereas 5 particles and 1500 iterations were used in the second case (Figure 6B). The resulting curves followed the pattern associated with standard compounds (Figure 6A-3 and Figure 6B-3). The consequent adjustments by the corresponding offsets and, if necessary, the subtraction of the exact mass of H_2 resulted in patterns centered more around zero (Figure 6A-4 and Figure 6B-4). The values of mass accuracies associated with the standard compounds (Table S-2) became better for both scenarios,

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

which is additionally emphasized by calculating the values of the corresponding RMSEs before and after applying the recalibration procedure. It is possible to observe the 3-fold decrease of RMSEs from 0.32 to around 0.11 for both cases.

The reported outputs generated from the simulated and experimental data sets suggest that the described recalibration method, using a long list of exact masses for MDiM construction, can represent a valuable tool in nontargeted metabolite profiling using FTICR-MS. Nevertheless, the approach can be easily extended to other MS techniques, since there is no dependence on a specific analytical platform. The recalibration method is nonparametric which represents an advantage in favor to flexibility in fitting complicated functions without introducing restrictions on their exact form. Therefore, it can be applied in cases when parametric approaches can fail, especially in case of noisy data. Since any observed pattern/curve on MDiM can be potentially used for recalibration, there is no strict requirements on precise assignments of experimental mass signals to molecular formulas or database entries. Nevertheless, achieving a high proportion of true positive assignments is certainly beneficial. In turn, it needs to be emphasized that only a sufficiently rich reference list can provide reliable information on the generated patterns that provide the opportunity to detect and correct subtle changes in mass accuracies. For MDiM construction and the recalibration procedure, it is not necessary to be restricted by a certain database. Although presented study was focused on a collection of unique entries from four databases, rough approximation of experimental m/z by molecular formulas can be used as well. However, it is important to remember that the density of feasible assignments increases along the m/z axis.[23,24] Therefore, the density map may not be resolvable in the corresponding region. As a possible solution, the described method can be combined with MDiNs for molecular formula assignment,[14] where no drastic expansion of possible annotations over the considered m/z range takes place.

There is an agreement with other described approaches emphasizing that mass spectrometric profiling experiments with the capability to detect thousands of signals can themselves have enough information for recalibration by using some of the corresponding m/z values as calibrants.[3−5,22] MDiMs represent a necessary tool for investigating the behavior of mass spectrometric data derived from nontargeted metabolomics experiments and can provide the basis for other applications in addition to spectral recalibration. Such a concept can facilitate matching molecular formulas to experimental m/z values because of the possibility to associate an assignment to a specific line/curve on a MDiM. Certainly, the advantages are more pronounced for HRMS, since the mapping involves resolving patterns associated with distinct mass differences that, in turn, provide a basis for capturing isobaric compounds.

CONCLUSION

The concept of MDiMs was described based on the discrete nature of exact masses corresponding to chemical compounds. MDiMs can be a powerful tool in describing samples used in nontargeted metabolomics experiments. A different perspective, dealing with mass differences rather than masses only, is provided. The construction of MDiMs by measuring all the pairwise differences between an experimental and exact mass list can be potentially used for spectral recalibration because of the opportunity to observe the systematic component to be eliminated. Therefore, the current work offers a novel recalibration procedure based on kernel density estimation followed by searching the corresponding function via particle swarm optimization algorithm. It was possible to show that the method is capable to define a calibration curve required for the respective correction of spectral data. Moreover, it was demonstrated that any pattern observed on MDiMs and associated with a specific mass difference can be potentially used for recalibration. The peculiarity of this approach is the flexibility in reference list selection and, consequently, mass differences. Therefore, it is not limited to any particular reference compounds. Furthermore, the approach is not restricted to FTICR-MS users only but can be easily adapted to any other metabolomics and/or proteomics analytical platforms.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b04555.

It provides detailed description of the experimental section, table of standards used in the analysis (Table S-1), and table describing the results of the recalibration (Table S-2) (PDF)

AUTHOR INFORMATION

Corresponding Author

* E-mail: schmitt-kopplin@helmholtz-muenchen.de.

ORCID

Kirill S. Smirnov: 0000-0002-0580-7976

Author Contributions

‖K.S.S. and S.F. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank the German Center for Diabetes Research (DZD; Grants G-501900-482 and G-501901-020).

REFERENCES

(1) Junot, C.; Fenaille, F.; Colsch, B.; Becher, F. *Mass Spectrom. Rev.* **2014**, *33*, 471−500.

(2) Qi, Y.; O'Connor, P. B. *Mass Spectrom. Rev.* **2014**, *33*, 333−352.

(3) Becker, C. H.; Kumar, P.; Jones, T.; Lin, H. *Anal. Chem.* **2007**, *79*, 1702−1707.

(4) Kozhinov, A. N.; Zhurov, K. O.; Tsybin, Y. O. *Anal. Chem.* **2013**, *85*, 6437−6445.

(5) Petyuk, V. A.; Jaitly, N.; Moore, R. J.; Ding, J.; Metz, T. O.; Tang, K.; Monroe, M. E.; Tolmachev, A. V.; Adkins, J. N.; Belov, M. E.; Dabney, A. R.; Qian, W. J.; Camp, D. G., 2nd; Smith, R. D. *Anal. Chem.* **2008**, *80*, 693−706.

(6) Johnson, C. H.; Ivanisevic, J.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451−459.

(7) Ohta, D.; Kanaya, S.; Suzuki, H. *Curr. Opin. Biotechnol.* **2010**, *21*, 35−44.

(8) Kujawinski, E. B.; Hatcher, P. G.; Freitas, M. A. *Anal. Chem.* **2002**, *74*, 413−419.

(9) Hughey, C. A.; Hendrickson, C. L.; Rodgers, R. P.; Marshall, A. G.; Qian, K. *Anal. Chem.* **2001**, *73*, 4676−4681.

(10) Kim, S.; Kramer, R. W.; Hatcher, P. G. *Anal. Chem.* **2003**, *75*, 5336−5344.

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358

(11) Schmitt-Kopplin, P.; Gabelica, Z.; Gougeon, R. D.; Fekete, A.; Kanawati, B.; Harir, M.; Gebefuegi, I.; Eckel, G.; Hertkorn, N. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 2763−2768.
(12) Gougeon, R. D.; Lucio, M.; Frommberger, M.; Peyron, D.; Chassagne, D.; Alexandre, H.; Feuillat, F.; Voilley, A.; Cayot, P.; Gebefugi, I.; Hertkorn, N.; Schmitt-Kopplin, P. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 9174−9179.
(13) Hertkorn, N.; Frommberger, M.; Witt, M.; Koch, B. P.; Schmitt-Kopplin, P.; Perdue, E. M. *Anal. Chem.* **2008**, *80*, 8908−8919.
(14) Moritz, F.; Kaling, M.; Schnitzler, J. P.; Schmitt-Kopplin, P. *Plant, Cell Environ.* **2017**, *40*, 1057−1073.
(15) Forcisi, S.; Moritz, F.; Lucio, M.; Lehmann, R.; Stefan, N.; Schmitt-Kopplin, P. *Anal. Chem.* **2015**, *87*, 8917−8924.
(16) Pleil, J. D.; Isaacs, K. K. *Journal of breath research* **2016**, *10*, 012001.
(17) Brenton, A. G.; Godfrey, A. R. *J. Am. Soc. Mass Spectrom.* **2010**, *21*, 1821−1835.
(18) Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P. *Journal of chromatography. A* **2013**, *1292*, 51−65.
(19) Matsuda, F.; Shinbo, Y.; Oikawa, A.; Hirai, M. Y.; Fiehn, O.; Kanaya, S.; Saito, K. *PLoS One* **2009**, *4*, e7490.
(20) Scott, D. W.; Sain, S. R. *Handbook of Statistics* **2005**, *24*, 229−261.
(21) Poli, R.; Kennedy, J.; Blackwell, T. *Swarm Intelligence* **2007**, *1*, 33−57.
(22) Gibbons, B. C.; Chambers, M. C.; Monroe, M. E.; Tabb, D. L.; Payne, S. H. *Bioinformatics* **2015**, *31*, 3838−3840.
(23) Kumar, S.; Kumar, M.; Stoll, R.; Thurow, K. A mathematical programming for predicting molecular formulas in accurate mass spectrometry. In *IEEE Conference on Automation Science and Engineering*, 2010.
(24) Kind, T.; Fiehn, O. *BMC Bioinf.* **2007**, *8*, 1−20.

DOI: 10.1021/acs.analchem.8b04555
Anal. Chem. 2019, 91, 3350−3358