High-resolution mass spectrometry: basic principles for using exact mass and mass defect for discovery analysis of organic molecules in blood, breath, urine and environmental media

To cite this article: Joachim D Pleil and Kristin K Isaacs 2016 J. Breath Res. 10 012001

View the article online for updates and enhancements.

Related content
- Crowd-based breath analysis: assessing behavior, activity, exposures, and emotional response of people in groups
  Jonathan Williams and Joachim Pleil
- Immunochemistry for high-throughput screening of human exhaled breath condensate (EBC) media: implementation of automated quanterix SIMOA instrumentation
  Joachim D Pleil, Michelle M Angrish and Michael C Madden
- Adapting biomarker technologies to adverse outcome pathways (AOPs) research: current thoughts on using in vivo discovery for developing in vitro target methods
  Joachim D Pleil, Jonathan D Beauchamp, Wolfram Miekisch et al.

Recent citations
- Exhaled Volatile Organic Compounds of Infection: A Systematic Review
  Waqar M. Ahmed et al
- Calibration and performance of synchronous SIM/scan mode for simultaneous targeted and discovery (non-targeted) analysis of exhaled breath samples from firefighters
  M. Ariel Geer Wallace et al
PERSPECTIVE

High-resolution mass spectrometry: basic principles for using exact mass and mass defect for discovery analysis of organic molecules in blood, breath, urine and environmental media

Joachim D Pleil and Kristin K Isaacs
Exposure Methods and Measurements Division, National Exposure Research Laboratory, US Environmental Protection Agency, Research Triangle Park, NC, USA
E-mail: pleil.joachim@epa.gov
Keywords: DSSTox database, discovery analysis, toxicity testing
Online supplementary data available from online

Premise

The evolution of state of the art analytical instrumentation now embraces pragmatically available high-resolution mass spectrometry (MS) capable of separating mass fragments at the fourth or fifth decimal place (exact mass) where previous instrumentation was limited to single-digit mass units (integer mass). This perspective explains how the concept of ‘exact mass’ (or ‘accurate mass’) arises from basic physics principles, how different atoms contribute to the difference between integer mass and exact mass, and how these distinctions can serve as an additional dimension in the identification of unknown molecules in environmental or biological samples. The focus here is only on exact mass of ‘mono-isotopic features’. More complex data reduction techniques involving molecular fragmentation, isotopic abundance, and chromatographic separation are left for future articles.

Overview

Toxicity testing and disease diagnosis have evolved from simple target analysis with modest instrumentation to embrace discovery (or non-targeted) analysis. Therein, as many compounds are measured as possible, and then the resulting complex data stream is interpreted to deduce probative changes or perturbations (Reichenbach et al 2012, Pleil and Stiegel 2013). Furthermore, the traditional in vivo animal and human biological media approaches are being supplemented by high-throughput in vitro analyses (Pereira and Williams 2007, Andersen et al 2010, Tice et al 2013, Pleil et al 2015a, Judson et al 2014). The ultimate goal is to link environmental exposures to internal dose to ultimate health outcome (Pleil and Sheldon 2011, Sobus et al 2011, Pleil 2015, Bean et al 2015), which requires a variety of mathematical, statistical, and database interpretation techniques (Sobus et al 2010, Pleil et al 2015b, Sobus et al 2015). Additionally, interpretation of the in vivo to in vitro linkage is becoming a crucial component for streamlining environmental and public health research as implemented through adverse outcome pathways (Vinken 2013, Angrish et al 2015). Regardless of the specific application, current research relies primarily on instrumentation using some form of mass spectrometry as the analytical finish.

State of the art instrumentation

Recent developments in such new analytical instrumentation have coupled time-resolved separation (gas and liquid chromatography) with high-resolution mass spectrometry; the new technology is adding complexity to analyses, especially with respect to data interpretation (Glish and Vachet 2003, Schymanski et al 2014, 2015). We can now routinely discern up to thousands of molecular ‘features’ in individual samples of blood, breath and urine, in cell lines, as well as in environmental matrices of air, water, soil, dust, and food (Chance et al 2003, Muir and Howard 2006, Hernandez et al 2012, Forcisi et al 2013, Kalluri et al 2014, Marzinke et al 2014, Gago-Ferrero et al 2015, Moritz et al 2015).

This increasing ability to detect ultra-trace levels of many thousands of chemicals is no longer relegated to a few ‘super-labs’, but is commercially available for standard analysis in research and production laboratories. The two basic technological improvements are in multi-dimensional gas or liquid chromatography (GC or LC) commonly referred to as GC × GC, or LC × LC, and in high-resolution mass spectrometry detection based on sector, time of flight (ToF), or Fourier-transform ion cyclotron resonance (FT-ICR) trapping methods (Marshall et al 1998,}

© 2016 IOP Publishing Ltd
Mondello et al. 2008, Liang et al. 2011, Wu et al. 2012). To date, most of the applications have been with LC-MS focused on blood and plasma; we note that gas-phase analytics, especially with GC-MS, and GC-ToF-MS are now gaining ground especially for in vitro measurements of cellular respiration that will ultimately gain advantage with the implementation of high resolution MS (Zhu et al. 2013, Amann et al. 2014, Bean et al. 2014, Schivo et al. 2014, Baranska et al. 2015). Many of these new applications are directly geared towards disease and infection diagnostics, especially for pneumonia and cystic fibrosis (Filipiak et al. 2015, Shestivska et al. 2015).

Data interpretation

Given the complexity of the resulting data stream from the new generation of analytical instrumentation, bioinformatics mining is becoming routinely applied both for ‘targeted’ analysis and for ‘discovery’ (non-targeted) analysis. The underling mathematics are complex, and specific to the particular instrumentation. As such, the interpretation of results is often subject to the current revision of software/firmware from a manufacturer and increasingly opaque to the user. We note that discrimination based on chromatographic elution time is still relatively straightforward; where the high complexity appears is in the interpretation of molecular weight which is measured to the 4th or 5th decimal place in standard atomic mass units such as amu, u, or Dalton (Da) discussed later. As such, the question arises as to what the total mass and tiny fractions of nucleon mass actually mean, and how differences should be interpreted (Cristoni and Bernardi 2004, Sleno 2012, Sugimoto et al. 2012, Barbara et al. 2011).

This perspective focuses on the concept of ‘exact mass’ which is the fundamental physics principle for high-resolution mass spectrometry. Certainly most of us are aware of the general ideas of mass spectrometry (ions are separated by their mass/charge ratio in an electromagnetic field, ionization creates a pattern of charged mass fragments, some atoms have isotopes, etc), but the subtleties of how the masses of individual atoms and the combinations of masses in a molecule are expressed at the atomic level are sometimes a bit vague. Herein, we hope to explain at a fundamental level how exact mass comes about, and how to exploit this new level of discrimination. The focus is on traditional organic molecules plus those containing at least one carbon atom as might be found in organic based media. The ultimate goals are to improve identification of targeted compounds by their exact mass determination, and to use exact mass to help classify unknown compounds found in discovery (non-targeted) analyses of complex organic matrices. There are certainly additional techniques for helping identify unknowns with mass spectrometry including isomer structure and fragmentation patterns; here we stay with the basics of mono-isotopic mass as the analytical discriminator. We further note that we focus exclusively with singly charged ions as is the basic assumption for MS in general.

What are atomic number and integer mass?

An element is defined by the number of protons in the atomic nucleus; this is the ‘atomic number’ as found in the periodic table of the elements. The ‘integer mass’ is defined as the sum of the number of protons and the number of neutrons in the atomic nucleus; an element can have more than one integer mass, but only one distinct atomic number. The number of electrons in a neutral atom is equal to the number of protons, but they do not contribute to the integer mass. Protons and neutrons are collectively referred to as nucleons.

What is the atomic (nuclear) mass?

We are much more attuned to measuring mass (or weight on the earth’s surface) in units like kilograms or pounds. At the sub-atomic (proton, neutron and electron) level, this is a bit unwieldy as their masses are $1.6726 \times 10^{-27}$ kg, $1.6749 \times 10^{-27}$ kg, and $9.1094 \times 10^{-31}$ kg, respectively. Initially, relative masses of individual atoms were assigned as the integer sum of the protons and neutrons (nucleons), and the contributions from the electrons were considered negligible. The nucleon counts reflected the initial designation of atomic mass, and a mole of an atom was defined as weighing as many grams as the nucleon number. For standard single mass unit resolution spectrometry, this remains a sufficient and elegant identifier. Many elements have isotopes, that is, the same number of protons, but different numbers of neutrons, and so the molar mass of an element accounts for the typical composition of isotopes; this is designated as ‘atomic weight’, and should not be confused with the ‘mono-isotopic mass’ which is discussed later.

The problem of fractions arose when instruments became capable of measuring mass at a finer resolution. The nucleons in different sized atoms have different observed masses. From a basic physics perspective, this is accounted for by Einstein’s energy versus mass equation, $E = mc^2$. The premise is that it requires energy to overcome the nuclear strong-force to pry neutrons and protons apart from their nucleus. As such, a free proton or neutron has slightly more mass than one bound inside an atom; the atomic mass is not just the sum of the masses of free protons and neutrons, but actually a slightly lower number.

Important: The atomic weight as found in the periodic table of the elements is comprised of a weighted average of all stable isotopes of that element. It is not the same as atomic mass, which is defined as the mass of an individual isotope. The atomic weight is used in thermodynamics equations using pressure, volume, temperature, density, etc. The atomic mass is the salient value for mass spectrometry.
What is an atomic mass unit (amu)?

The concept of ‘amu’ or atomic mass unit, was implemented to avoid using kg numbers and to make the mass of different elements (nuclei) more easily comparable. The popular web literature and Wikipedia searches attribute the concept of indexing atomic weight to hydrogen in 1803 to John Dalton (1766–1844) although his contemporaries including William Prout (1785–1850) and Jons Jakob Berzelius (1779–1848) were also involved in the early work for coming to consensus regarding atoms and their masses. Later on, Wilhelm Ostwald (1853–1932) is credited with suggesting 1/16 of the mass of the oxygen atom \((^{16}\text{O})\), with 8 protons and 8 neutrons) as the index. Anecdotally, this pragmatic change was adopted because, at the time, it was more accurate to weigh mass in general. As the relative abundance of \(^{16}\text{O}\) is 99.757%, and \(^{18}\text{O}\) is only ~0.205%, it was not considered an important problem and the common use of the ‘amu’ based on 1/16 of an oxygen atom mass continued. However, mass measurement and chemistry instrumentation continued to improve and the study of organic compounds dominated chemistry. In 1961, the International Union of Pure and Applied Chemistry (IUPAC) revisited the issue and it was decided to designate 1/12 of a carbon atom comprised of 6 protons and 6 neutrons \((^{12}\text{C})\) as the index mass as carbon defines what is an organic molecule; this was designated as the ‘unified atomic mass unit’ or ‘u’. To date, the acronym ‘amu’ is commonly used to represent u based on carbon, and more recently, the Dalton, or Da, was chosen as the official name of the unit. In some applications, the mass to charge ratio is used as an identifier; the mass defect of an atom as the difference between the exact mass and the integer mass.

The discovery of isotopes in 1912, and especially that of the \(^{18}\text{O}\) isotope (8 protons and 10 neutrons), caused some re-evaluation of how to interpret atomic mass in general. As the relative abundance of \(^{18}\text{O}\) is 99.757%, and \(^{16}\text{O}\) is only ~0.205%, it was not considered an important problem and the common use of the ‘amu’ based on 1/16 of an oxygen atom mass continued. However, mass measurement and chemistry instrumentation continued to improve and the study of organic compounds dominated chemistry. In 1961, the International Union of Pure and Applied Chemistry (IUPAC) revisited the issue and it was decided to designate 1/12 of a carbon atom comprised of 6 protons and 6 neutrons \((^{12}\text{C})\) as the index mass as carbon defines what is an organic molecule; this was designated as the ‘unified atomic mass unit’ or ‘u’. To date, the acronym ‘amu’ is commonly used to represent u based on carbon, and more recently, the Dalton, or Da, was chosen as the official name of the unit. In some applications, the mass to charge ratio is used as an identifier; the mass defect of an atom as the difference between the exact mass and the integer mass.

An important problem and the common use of the ‘amu’ based on 1/16 of an oxygen atom mass continued. However, mass measurement and chemistry instrumentation continued to improve and the study of organic compounds dominated chemistry. In 1961, the International Union of Pure and Applied Chemistry (IUPAC) revised the issue and it was decided to designate 1/12 of a carbon atom comprised of 6 protons and 6 neutrons \((^{12}\text{C})\) as the index mass as carbon defines what is an organic molecule; this was designated as the ‘unified atomic mass unit’ or ‘u’. To date, the acronym ‘amu’ is commonly used to represent u based on carbon, and more recently, the Dalton, or Da, was chosen as the official name of the unit. In some applications, the mass to charge ratio is used as an identifier; the mass defect of an atom as the difference between the exact mass and the integer mass.

What is exact mass?

The ‘exact mass’ of a molecule is stated in u or Da and is measured empirically; nowadays, it is typically determined by the path a singly ionized molecule takes in a precisely defined electromagnetic field, and represents the underlying physics of mass spectrometry. If we know the chemical formula of the molecule, then the exact mass can also be calculated by summing the exact masses of the individual atoms making up the molecule. In mass spectrometry, we usually index each atom and molecule by its ‘mono-isotopic mass’, that is, we do not allow mixtures of isotopes to contribute to the exact mass of an individual molecule. This becomes a bit difficult to define for atoms that have numerous isotopes with similar percentages of naturally occurring stable isotopes, especially for those in metals. For organic species the six most common atoms are carbon, hydrogen, nitrogen, oxygen, phosphorus and sulfur (C, H, N, O, P, S). We are also concerned with exogenous organic compounds containing halogens (e.g. fluorine, chlorine, bromine and iodine). In either grouping, the mono-isotope choice is straightforward. For example, carbon (6 protons) has isotopes with 6, 7, and 8 neutrons, but the 6-neutron isotope dominates with 98.9% abundance. Chlorine (17 protons) has two major isotopes with 18 and 20 neutrons (75.8% and 24.2% abundance, respectively) so we use the 18-neutron isotope. Phosphorus (15 protons) has 23 different isotopes ranging from 9 to 31 neutrons, but only the 16-neutron isotope is stable, and so it is the proper choice. Even when there are lower isotopes, we generally choose the one most likely to be present in the real-world; for some metals, this is not always obvious, but fortunately, organic compounds with such atoms are rare.

Table 1 shows the number of nucleons and mass data for the atoms mentioned above, plus some others that are found in organic compounds. We define the mono-isotopic atom as the most prevalent stable isotope and the mass defect of an atom as the difference between the exact mass and the integer mass.

<table>
<thead>
<tr>
<th>Atom</th>
<th>Mass (u)</th>
<th>Mass Defect (u)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>12.0111</td>
<td>0.0111</td>
</tr>
<tr>
<td>H</td>
<td>1.00794</td>
<td>0.00794</td>
</tr>
<tr>
<td>N</td>
<td>14.0031</td>
<td>0.0031</td>
</tr>
<tr>
<td>O</td>
<td>15.9949</td>
<td>0.9949</td>
</tr>
<tr>
<td>P</td>
<td>30.9737</td>
<td>0.9737</td>
</tr>
<tr>
<td>S</td>
<td>32.065</td>
<td>0.065</td>
</tr>
</tbody>
</table>

What is the exact mass of a molecule?

As mentioned above, the exact mass for a molecule (where we do not know the chemical formula) can be measured empirically using basic physics principles of a charged object moving in an electromagnetic field. If we know the chemical formula, then the exact mass can be calculated using table 1 entries for the exact masses of the constituent atoms. For example, the exact mass of ethane, \(\text{C}_2\text{H}_4\), is calculated as:

\[
\text{C}_2\text{H}_4 : (2 \times 12.0000) + (4 \times 1.007 825) = 28.0313 \text{ Da}
\]

and the exact mass of dichloromethane (the common solvent \(\text{CH}_2\text{Cl}_2\)) is calculated as:

\[
\text{CH}_2\text{Cl}_2 : (1 \times 12.0000) + (2 \times 1.007 825) + (2 \times 34.9689) = 83.9534 \text{ Da}
\]

We note that the integer masses of ethane and dichloromethane are 28 and 84, respectively, and that the exact masses can be above or below the integer masses depending on the atomic constituents.

How does measuring exact mass help tell compounds apart?

Traditional mass spectrometers, either linear quadrupole or standard ion trap configurations, have 1 Da resolution; for example, if a fragment lights up as 73 Da, then we can tell it apart from 72 or 74 Da, but not at any finer discriminator level. (We note that some higher performance quadrupole and triple
quadrupole instruments may have finer resolution with about 0.1 to 0.5 Da). In general, however, if we have a chromatogram where compounds with 73 Da come out close together in time, we can only guess as to which compound, or mixture of compounds, is present. As a hypothetical example, suppose we run a breath sample using GC-MS and we find extracted ion profiles for 73 and for 86 Da. We could not discern the difference among the specific compounds as shown in the example if we only knew the integer mass. Table 2 shows different possible compounds (isobars) with 73 Da and 86 Da integer masses.

Here we see that instruments with high mass resolution could distinguish among the specific compounds as shown in the example if we only knew the integer mass. Table 2 shows different possible compounds (isobars) with 73 Da and 86 Da integer masses.

Table 1. Mono-isotopic masses and nucleon numbers for common atoms in organic molecules.

<table>
<thead>
<tr>
<th>Atom</th>
<th>Protons (#)</th>
<th>Neutrons (#)</th>
<th>Integer mass (Da)</th>
<th>Exact mass (Da)</th>
<th>Mass defect (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>12.000 00</td>
<td>0.000 00</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.007 83</td>
<td>0.007 83</td>
</tr>
<tr>
<td>Nitrogen</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td>14.003 07</td>
<td>0.003 07</td>
</tr>
<tr>
<td>Oxygen</td>
<td>8</td>
<td>8</td>
<td>16</td>
<td>15.994 92</td>
<td>−0.005 08</td>
</tr>
<tr>
<td>Phosphorus</td>
<td>15</td>
<td>16</td>
<td>31</td>
<td>30.973 76</td>
<td>−0.026 24</td>
</tr>
<tr>
<td>Sulfur</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>31.972 07</td>
<td>−0.027 93</td>
</tr>
<tr>
<td>Fluorine</td>
<td>9</td>
<td>10</td>
<td>19</td>
<td>18.998 40</td>
<td>−0.001 60</td>
</tr>
<tr>
<td>Chlorine</td>
<td>17</td>
<td>18</td>
<td>35</td>
<td>34.968 85</td>
<td>−0.031 15</td>
</tr>
<tr>
<td>Bromine</td>
<td>35</td>
<td>45</td>
<td>79</td>
<td>78.918 34</td>
<td>−0.081 66</td>
</tr>
<tr>
<td>Iodine</td>
<td>53</td>
<td>74</td>
<td>127</td>
<td>126.904 48</td>
<td>−0.095 52</td>
</tr>
</tbody>
</table>


Table 2. Two examples of compounds with same integer mass and different chemical formulas; knowledge of exact mass allows discrimination.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Formula</th>
<th>Integer mass</th>
<th>Exact mass (Da)</th>
<th>Mass defect (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methylisocyanate</td>
<td>C₃H₅NS</td>
<td>73</td>
<td>72.998 62</td>
<td>−0.001 38</td>
</tr>
<tr>
<td>Ethene, nitro-</td>
<td>C₂H₃NO₂</td>
<td>73</td>
<td>73.016 38</td>
<td>0.016 38</td>
</tr>
<tr>
<td>N,N-Dimethylformamide</td>
<td>C₂H₅NO</td>
<td>73</td>
<td>73.052 76</td>
<td>0.052 76</td>
</tr>
<tr>
<td>1-methylguanidine</td>
<td>C₅H₃N₂</td>
<td>73</td>
<td>73.064 00</td>
<td>0.064 00</td>
</tr>
<tr>
<td>Isobutylamine</td>
<td>C₄H₉N</td>
<td>73</td>
<td>73.089 15</td>
<td>0.089 15</td>
</tr>
<tr>
<td>Chlorodifluoromethane</td>
<td>CH₂Cl₂F₂</td>
<td>86</td>
<td>85.973 48</td>
<td>−0.026 52</td>
</tr>
<tr>
<td>Methoxy, trifluoro-</td>
<td>CHF₂O</td>
<td>86</td>
<td>85.997 93</td>
<td>−0.002 05</td>
</tr>
<tr>
<td>Vinyl acetate</td>
<td>C₄H₆O₂</td>
<td>86</td>
<td>86.036 78</td>
<td>0.036 78</td>
</tr>
<tr>
<td>Ethylene urea</td>
<td>C₅H₈N₂O</td>
<td>86</td>
<td>86.048 01</td>
<td>0.048 01</td>
</tr>
<tr>
<td>Pentanal</td>
<td>C₅H₁₀O</td>
<td>86</td>
<td>86.073 16</td>
<td>0.073 16</td>
</tr>
</tbody>
</table>


What is the mass defect of an atom or molecule?

The difference between the integer mass and the exact mass of an atom or molecule is termed the mass defect as shown for atoms in the last column of table 1 and for molecules in the last column of table 2. In effect, it is the mass equivalent to the energy required to separate each nucleon away from each nucleus. One might wonder why the mass defect for ethane appears positive, and the mass defect for methylene chloride is negative as we expect all mass defects to be negative under the physics interpretation. This discrepancy is an artifact of the definition of the Dalton as 1/12 of the carbon 12C atom; as such, the mass defect of carbon is defined as the zero set-point, rather than using a separated ‘free’ nucleon.

When the integer mass is known via the chemical formula, the mass defect is calculated directly by subtraction. However, if we measure the exact mass of an unknown ‘feature’ in a high-resolution mass spectrometer analysis, we cannot calculate the mass defect directly, and need to infer the integer mass somehow.

What is the mass defect of a molecule with unknown formula?

The mass spectrometer can provide the exact mass for many thousands of molecules and fragments in
any individual sample analysis; the problem is how to determine the mass defect if we do not know the integer mass from the formula. This has been resolved to some extent with an approximation technique. For any given exact mass, the integer mass is defined as the closest whole unit. So, for the two examples above, if ethane and dichloromethane were considered unknowns of exact mass 28.0313 and 83.9534 Da, respectively, then the integer masses would be (correctly) assigned as 28 and 84 Da.

This is a reasonable approach as the deviations of atomic exact mass from the atomic integer mass are small (see table 1). However, we must acknowledge that this limits the range of assigned mass defect from $\geq -0.5$ to $< 0.5$. What if the true mass defect is outside of this imposed range?

For example, a simple molecule like C$_{4}$Br$_{8}$ (a catalyst used in making oligomers) has a true mono-isotopic integer mass of 680 Da and exact mass of:

$$C_{4}Br_{8} : (4 \times 12.0000) + (8 \times 78.9183) = 679.3467 Da$$

If we already know that the true formula is C$_{4}$Br$_{8}$, then the true mass defect is calculated as $-0.6533$; if we treat this compound as an unknown, then we would assign the integer mass as 679 Da, and erroneously calculate the mass defect as $+0.3467$.

Similarly, larger hydrocarbons (HC), especially saturated compounds, can have large positive mass defect. The n-alkanes with more than 31 carbons found in greases, waxes and asphalts, all have true mass defect greater than $+0.5000$. For example, C$_{31}$H$_{64}$ is a major alkane constituent of tobacco leaves and beeswax. It has a true integer mass of 436 Da and an exact mass of:

$$C_{31}H_{64} : (31 \times 12.0000) + (64 \times 1.0078) = 436.5008 Da$$

As for the previous case with C$_{4}$Br$_{8}$, if we know the formula C$_{31}$H$_{64}$ and thus the true integer mass, then the true mass defect can be calculated as $+0.5008$ Da. If however, this compound were treated as an unknown, we would assign the integer mass to 437 Da and erroneously assign the mass defect to $-0.4992$ Da. All higher n-alkanes would similarly be incorrectly assigned.

**How can mass defect be used to classify unknown compounds?**

In general, the exact mass can be used to better narrow down the identity of compounds that have already been placed into a library, and it can assist in suggesting specific chemical formulas that would add up to a specific exact mass within that instruments’ particular resolution window. As such, the combination of exact mass and mass defect gives an indication of compound class and size according to the information in table 1. If mass defect is positive, then we expect a preponderance of oxygen, sulfur, phosphorus and halogens. Recall that the carbon content of organic molecules remains neutral in these calculations as it is defined as having zero mass defect.

Various classification schemes have been proposed for specific biological matrices. Most often, they are based on a specific range of related compounds of interest, such as perfluorinated compounds (PFCs), organophosphorus pesticides (OPP), or groups of protein fragments for which there is already some knowledge. For example, if we have exact mass and true mass defect for a series 5 or 10 PFCs ranging from 175 to 450 Da with mass defect ranging from $-0.015$ to $-0.066$, then we could exclude unknowns outside of these ranges for detecting candidates of similar, but as yet undocumented, PFCs (Strynar et al 2015).

**Testing mass defect classification with DSSTox**

As indicated above, there are some caveats to using mass defect for classifying unknown compounds into groups. To test the value of using mass defect based on integer mass estimates, we need to use a database of known structures, and then determine how well we could classify them under the assumption that we do not know the true formula. To this end, we use the Distributed Structure—Searchable Toxicity (DSSTox) database developed by U.S. Environmental Protection Agency as a baseline. As defined by US EPA:

“…The DSSTox Database incorporates state-of-the-art cheminformatics workflows, provides the chemical infrastructure for EPA’s Safer Chemicals Research, including the ToxCast and Tox21 high-throughput toxicology efforts. EPA will continue to add content to DSSTox and users will be able to directly access the updated and expanded DSSTox database, as well as chemical exploration tools and capabilities currently under development.”

(www2.epa.gov/chemical-research/distributed-structure-searchable-toxicity-dsstox-database).

Briefly, we use the DSSTox database to represent a subset of the chemical space of environmental and biological, and commercial chemicals that includes “…standardized, high quality chemical structure files for chemical substances of interest to the U.S. EPA, and the larger environmental health community.” (Rager et al 2016). The purpose is to have a highly accurate and well-curated set of known chemicals against which methods for assessing unknowns could be tested.

The mass range for the organic entries in the DSSTox Database is from 15 to 2284 Da; these compound masses are lognormally distributed with geometric mean, $GM = 244.73$ and geometric standard deviation, $GSD = 1.70$. Figure 1 shows the theoretical mass frequency distribution. Although the total range goes up to 2284 Da, the 90th, 95th, 99th and 99.9th percentiles are much lower at 475, 586, 841, and 1261 Da, respectively.
Perspective

Distribution of mass defect for DSSTox entries

When we use mass defect for classifying unknowns, it is important to have an underlying concept of how these values are distributed; that is, how probative are differences in mass defect for classifying different unknowns? The calculated (true) mass defect is approximately normally distributed with mean \( \mu = 0.102 \) and standard deviation \( \sigma = 0.128 \). Figure 2 shows this distribution.

Figure 2 demonstrates that about 95% of all mass defect values lie between \(-0.156\) and \(0.347\) Da, and that only 22.4% are negative. This shows that there is a preponderance of hydrogen and nitrogen (positive mass defect) in organic molecules that overwhelms the occurrence of oxygen, phosphorus, sulfur and the halogens. The relative influence of common atoms is seen in table 1. A negative mass defect is relatively rare (1/5), and so is a useful first cut in categorizing unknowns. We note that halogens (especially) are of importance in finding anthropogenic chemicals with toxic properties in the wide range of all chemicals.

DSSTox database structure

The current completed and validated version of DSSTox (10/2014) contains 34 669 compounds, their structures, formulas, and various identification numbers. Internally, we have further calculated integer mass, exact mass, and mass defect for each compound. For the ensuing evaluation, we consider only organic compounds (defined as containing at least one carbon atom); this reduces the total count to 32 424

![Figure 1](image1.jpg)  
Figure 1. Distribution of compound masses in DSSTox Database. Units in Da: GM = 245; GSD = 1.7; range (15–2248); 90% = 475; 95% = 586; 99% = 841. Frequency on y-axis is for 1 Da wide bins.

![Figure 2](image2.jpg)  
Figure 2. Empirical and modeled distribution of ‘true’ mass defect for all carbon containing entries in DSSTox database. The distribution is close to normal with \( \mu = 0.0912 \) and \( \sigma = 0.1206 \), and 95% of all values are between \(-0.156\) and \(0.347\); however, the actual range of mass defect is \(-1.23\) to \(2.04\) which tends to broaden the model (smooth line) beyond the empirical (bars) data.
# Table 3. Excerpt of DSSTox database as reconstructed for mass defect assessment (full database given in the supplementary material).

<table>
<thead>
<tr>
<th>DSSToxCID</th>
<th>CAS</th>
<th>Chemical name</th>
<th>Formula a</th>
<th>Mol. weight a</th>
<th>Monoisotopic mass (mass a)</th>
<th>Rounded mass (mass a)</th>
<th>Mass defect (rounded)</th>
<th>Integer (mass a)</th>
<th>Mass defect (integer)</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>33187</td>
<td>110999-46-7</td>
<td>1,2,3,7,8,9-Hexabromo-dibenzo-p-dioxin</td>
<td>C12H2Br6O2</td>
<td>657.5700</td>
<td>651.5155</td>
<td>652</td>
<td>−0.4845</td>
<td>652</td>
<td>−0.4845</td>
<td>Yes</td>
</tr>
<tr>
<td>30773</td>
<td>544-85-4</td>
<td>Dotriacontane</td>
<td>C32H66</td>
<td>450.8800</td>
<td>450.5165</td>
<td>451</td>
<td>−0.4835</td>
<td>450</td>
<td>0.5165</td>
<td>No</td>
</tr>
<tr>
<td>20521</td>
<td>56092-81-0</td>
<td>Ionomycin</td>
<td>C41H72O9</td>
<td>709.0180</td>
<td>708.5176</td>
<td>709</td>
<td>−0.4824</td>
<td>708</td>
<td>0.5176</td>
<td>No</td>
</tr>
<tr>
<td>9217</td>
<td>1569-01-3</td>
<td>1-Propoxy-2-propanol</td>
<td>C6H14O2</td>
<td>118.1760</td>
<td>118.0994</td>
<td>118</td>
<td>0.0994</td>
<td>118</td>
<td>0.0994</td>
<td>Yes</td>
</tr>
<tr>
<td>31893</td>
<td>55511-98-3</td>
<td>Buthidazole</td>
<td>C10H16N4O2S</td>
<td>256.3300</td>
<td>256.0994</td>
<td>256</td>
<td>0.0994</td>
<td>256</td>
<td>0.0994</td>
<td>Yes</td>
</tr>
<tr>
<td>22303</td>
<td>4726-14-1</td>
<td>Nitratin</td>
<td>C13H19N3O6S</td>
<td>345.3760</td>
<td>345.0995</td>
<td>345</td>
<td>0.0995</td>
<td>345</td>
<td>0.0995</td>
<td>Yes</td>
</tr>
<tr>
<td>2781</td>
<td>3577-01-3</td>
<td>Cephaloglycin</td>
<td>C18H19N3O6S</td>
<td>405.4310</td>
<td>405.0995</td>
<td>405</td>
<td>0.0995</td>
<td>405</td>
<td>0.0995</td>
<td>Yes</td>
</tr>
<tr>
<td>2760</td>
<td>60925-61-3</td>
<td>Ceferanide</td>
<td>C20H21N7O8S2</td>
<td>519.5630</td>
<td>519.0997</td>
<td>519</td>
<td>0.0995</td>
<td>519</td>
<td>0.0995</td>
<td>Yes</td>
</tr>
<tr>
<td>24868</td>
<td>2050-43-3</td>
<td>N-(2,4-Dimethylphenyl)-acetamide</td>
<td>C10H13NO</td>
<td>163.2200</td>
<td>163.0997</td>
<td>163</td>
<td>0.0997</td>
<td>163</td>
<td>0.0997</td>
<td>Yes</td>
</tr>
<tr>
<td>376</td>
<td>1163-19-5</td>
<td>Decabromodiphenyl oxide</td>
<td>C12Br10O</td>
<td>959.1710</td>
<td>949.1783</td>
<td>949</td>
<td>0.1783</td>
<td>949</td>
<td>−0.8217</td>
<td>No</td>
</tr>
<tr>
<td>34372</td>
<td>13654-09-6</td>
<td>1,1′-Biphenyl, 2,2′,3,3′,4,4′,5,5′-Decabromophenyl ester</td>
<td>C12Br10</td>
<td>943.1720</td>
<td>933.1834</td>
<td>933</td>
<td>0.1834</td>
<td>934</td>
<td>−0.8166</td>
<td>No</td>
</tr>
<tr>
<td>46967</td>
<td>71216-02-9</td>
<td>Carbonic acid, 1,2-ethanediylbis[(pentabromophenyl) ester</td>
<td>C16H4Br10O6</td>
<td>1091.2419</td>
<td>1081.1842</td>
<td>1081</td>
<td>0.1842</td>
<td>1082</td>
<td>−0.8158</td>
<td>No</td>
</tr>
</tbody>
</table>

a Values for ‘desalted’ mass of compound to reflect real world occurrence.
compounds each with a discrete entry using only the mono-isotopic mass. The full re-constructed database is given in the supplementary material (stacks.iop.org/JBR/10/012001/mmedia) as an excel file. Briefly, table 3 shows an example excerpt.

The excerpted compounds were arbitrarily chosen to show a range of mass defects based on the true integer mass to demonstrate how using the rounding technique for unknowns might go awry. In the complete database of all 32 424 carbon-containing compounds, the rounding method turns out to be quite robust. The details are discussed in the next section.

Accuracy of rounding method for unknowns

We have seen how the rounding method for approximating integer mass may give the wrong value depending on the true formula of the mass spectrometric feature. Granted, it requires an odd combination of many positive or negative mass defect atoms to cause a disruption, however, when grouping unknowns by structure, this may cause misidentification of a crucial feature. Upon investigating the total DSSTox Database of defined organic compounds (supplementary material), we found that overall, the rounding method assigned 1.31% erroneous mass defects, of which 1.22% were incorrectly labeled as negative, and 0.09% were incorrectly labeled as positive.

Under the assumption that the entries in the DSSTox database are representative of a broad spectrum of compounds that could be found in the environment or in biological samples, we now have reasonable (but not perfect) confidence for the mass defect grouping method for unknown features. We can further explore the classification error by sub-grouping these results by total molecular mass. Figure 3 shows the percent error (how many incorrect assignments of mass defect are expected), as a function of mono-isotopic mass. Here we see that the rounding method is completely valid for molecular mass up to about 450 Da, which represents about 90% of the entries in DSSTox. Beyond 450 Da, the remaining 10% of the entries have monotonically increasing error rates up to 100% at about 1900 Da.

Summary

Measuring the exact mass is a valuable tool for categorization, identification, and verification of organic chemicals as it adds another discrimination factor beyond standard time-elution and single-Dalton resolution. Herein we have explained the origins of mass differences among molecules with identical integer mass (total nucleon number), the evolution of the atomic mass unit (amu) and later the Dalton (Da), and finally how exact mass and the attendant mass defect are distributed, and how they can be interpreted for compound classification. The basis for all comparisons was the DSSTox database of about 33 000 compounds (from 2014); we assumed that this database is approximately representative of a broad variety of chemicals found in environmental and biological samples.

Specific conclusions

Based on this perspective, we can make a series of informed statements about the use of exact mass in the analytical laboratory:

- Atomic weight, as found in periodic tables, is not used for MS; it represents a weighted average of all isotopes.
Mono-isotopic atomic masses are used as the baseline for a molecule; they represent the masses of individual isotopes of specific atoms that add up to the molecular mono-isotopic mass.

- Integer mass of molecules is crucial; this can be directly calculated from the chemical formula using mono-isotopic atomic nucleon numbers.
- For unknown chemical formula, integer mass can be estimated by rounding the measured exact mass to the nearest integer.
- Rounding method for integer mass is only robust below ~450 Da.
- 90% of all DSSTox entries have integer mass <475 Da.
- 95% of mass defect values are between −0.156 and 0.347.
- Only about 22% of all DSSTox entries have negative mass defect.
- Negative mass defect is an indicator of excess halogens, oxygen, sulfur, and phosphorus atoms in a molecule.
- Positive mass defect is an indicator of excess hydrogen and nitrogen atoms in a molecule.

Caveats of using exact mass

We must recognize that isomeric structures, (compounds with the same atoms, but different arrangements) cannot be separated by exact mass. Also, it is always possible that two different chemical formulas could have exact masses so close together that the particular instrument cannot resolve the difference. Finally, exact mass and mass defect are not the complete answer to identifying unknowns; additional information from chromatographic retention time, isotopic mass abundances, and mass fragmentation patterns all contribute to unambiguously defining organic molecules. In short, added information is needed to discern which feature is actually the mono-isotopic molecule that can be used to identify a particular compound. We note that all high-resolution MS instruments have their own proprietary software that assists in making suggestions for chemical formulas and structures based on exact mass, estimated mass defect, and the other MS information mentioned above. For the present article, we restricted the discussion to the basics of mono-isotopic mass.

We note that the various proprietary computational approaches for identifying compounds using pattern recognition and rules-based tactics are effective, however, at some point a human must study the results and make the ultimate decision. As such, an a priori discriminator (or triage method) is important to streamline the ultimate interpretation. The exact mass and mass defect parameters described here are important because serve as such discriminators and are often used in larger high-throughput commercial laboratories where many of thousands of features need to be evaluated efficiently.

Ultimate advantage of using exact mass

In contrast to standard mass spectrometry with single Da resolution, the added information from high-resolution features narrows down the possible chemical formulas of a molecule, supports confirmation of targeted species, and helps identify unknowns. Once the integer mass and exact mass are established, the mass defect can be used to further classify compounds by the preponderance of atoms present; this is most useful for discerning halogenated molecules as a series or group in a complex sample. Finally, the use of rounding to estimate integer mass of unknowns can be considered robust for breath analysis as most exhaled gases are well below the 450 Da limit. In conclusion, exact mass adds a crucial third dimension for compound identification beyond chromatographic retention time and single Da fragmentation pattern.

Acknowledgments

The authors are grateful for the expert advice from Dr Mark Strynar, Dr Jon Sobus, and Dr Seth Newton of US EPA regarding non-targeted analysis and the DSSTox database. The United States Environmental Protection Agency through its Office of Research and Development has subjected this article to Agency administrative review and approved it for publication. Mention of trade names or commercial products does not constitute endorsement for use.

References

Baranska A, Smolinska A, Boots A W, Dallinga J W and van Schooten F J 2015 Dynamic collection and analysis of volatile organic compounds from the headspace of cell cultures J. Breath Res. 9 047102
Bean H D, Pleil J D and Hill J E 2015 Report: new analytical and statistical approaches for interpreting the relationships among environmental stressors and biomarkers Biomarkers 20 1–4
Bean H D, Zhu J, Sngle J C and Hill J E 2014 Identifying methicillin-resistant Staphylococcus aureus (MRSA) lung infections in mice via breath analysis using secondary electrospray ionization-mass spectrometry (SESI-MS) J. Breath Res. 8 041001