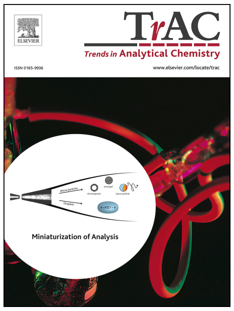
From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data

Nash, William; Dunn, Warwick

DOI:
10.1016/j.trac.2018.11.022


ELSEVIER
TrAC
Trends in Analytical Chemistry
ISSN 0165-9936
www.elsevier.com/locate/trac
Miniaturization of Analysis

From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data

William J. Nash, Warwick B. Dunn

PII: S0165-9936(18)30189-4

DOI: https://doi.org/10.1016/j.trac.2018.11.022

Reference: TRAC 15324

To appear in: *Trends in Analytical Chemistry*

Received Date: 3 May 2018

Revised Date: 15 November 2018

Accepted Date: 15 November 2018

From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data

William J. Nash[1] and Warwick B. Dunn[1,2,3†]

[1] School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[2] Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
[3] Institute of Metabolism and Systems Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Corresponding Author (†)
Email: w.dunn@bham.ac.uk
Telephone: +44 (0)121 4145458

Abstract

The relatively unbiased study of metabolites in biological systems is called untargeted metabolomics and the application of liquid chromatography-mass spectrometry platforms for data acquisition is now common across the world. When operating in its most unbiased form, this experimental strategy starts from assuming no knowledge of the metabolites to be detected and instead the data acquired is used to annotate or identify the detected metabolites on a study-by-study basis. The process of metabolite annotation is a bottleneck in untargeted metabolomics and to which significant progress has been made in the last ten years in understanding the limitations and developing new experimental and computational methods and tools to enhance our capabilities. In this review we will describe the current status of tools applied for metabolite annotation and discuss current areas where further work is required.

Keywords

Untargeted metabolomics; annotation; identification; electrospray; mass spectral libraries, gas phase fragmentation; metabolomics databases

Abbreviations

AIF: All Ion Fragmentation
DDA: Data Dependent Analysis
DIA: Data Independent Analysis
FWHM: Full Width Half Maximum
LC-MS: Liquid Chromatography-Mass Spectrometry
MS/MS: Tandem Fragmentation Mass Spectrometry
MS^n: Multistage Fragmentation Mass Spectrometry
m/z: mass-to-charge ratio
SWATH: Sequential Windowed Acquisition of All Theoretical Fragment Ion Mass Spectra

1. The complexities of metabolite annotation in untargeted metabolomics studies

The study of metabolites in biological samples is routinely defined as metabolomics and provides the capability to investigate metabolism on a global and relatively unbiased scale in comparison to traditional targeted studies focused on specific areas/pathways of metabolism and a small number of metabolites [1]. Changes in the concentrations of metabolites reflects dynamic and rapid changes in the phenotype of the system being studied; for example, mammalian muscle and blood lactate levels increase within minutes during an intense exercise event. This global approach provides opportunities to detect thousands of metabolites in hypothesis-generating (rather than hypothesis-testing) studies and to associate previously unknown metabolites with biologically important roles (e.g. metabolism, signalling, regulation and synthesis of larger biomolecules) in human health and disease, biotechnology, drug discovery and plant sciences. These discovery studies should not end here but should be validated both analytically and biologically applying targeted biological and metabolomic studies.

Significant developments related to instrumentation (e.g. increased mass resolving power; see [2] as an example), informatics (software and databases; see [3] for a recent review and Table 1 for a list) and analytical chemistry methods for sample collection and preparation (see [4] for a discussion on lipidomics) have increased the number of metabolites detected and annotated in untargeted metabolomic studies applying liquid chromatography-mass spectrometry platforms. Mass spectrometry is a relatively unbiased detector which has a high sensitivity and untargeted metabolomic studies apply a crude extraction of biological samples with no steps included to separate metabolites from non-biological chemicals. Today, thousands of ‘signals’ are detected in biological samples by liquid chromatography-mass spectrometry (LC-MS) platforms which relate to endogenous (e.g. amino acids) and exogenous (e.g. over-the-counter drug) metabolites as well as other chemicals whose source is not biological (e.g. contaminants in sample collection tubes or chemical solvents). These signals are reported with up to four different types of data – chromatographic retention time, full-scan mass-to-charge (*m/z*) ratio, MS/MS or MS^n (where n>2) mass spectrum and ion mobility drift time for full-scan or MS/MS data. Of importance is the knowledge that electrospray ionisation generates multiple signals for each metabolite [5] as will be discussed below in section 2. Therefore the processing of these data to remove non-biological signals and to integrate multiple signals in to a single metabolite are required to provide a cleaned dataset for further univariate and multivariate data analysis. A recent study [6] showed that in *Escherichia coli* intracellular extracts, up to 25,000 signals can be detected which relate to approximately 1000 metabolites and therefore demonstrated the disparity between the number of signals detected and the number of metabolites detected. It should be noted that the non-optimised use of raw data processing software (e.g. XCMS) can significantly inflate the number of signals detected in the authors experience and optimisation of software processing parameters is highly recommended (for example, see IPO for optimisation of XCMS parameters [7]).

A second complexity in untargeted metabolomics studies is that we do not know ALL of the metabolites which may be present in one or multiple biological samples and which could be detected. Figure 1 visualises the different factors impacting on the presence/absence and concentration of metabolites in human metabolomes. The concept of ‘dark matter’ in the metabolomes of different organisms has been introduced [8] and observations of the synthesis of unknown metabolites by promiscuous enzymes [9], the emergence of epi-metabolites [10] and modifications to damage-prone endogenous metabolites [11] have all been reported. There are still many metabolites not reported or searched for in metabolomics databases. Untargeted studies start from limited knowledge so to provide a relatively unbiased survey of metabolites and allow detection of previously unreported metabolites which have a previously unreported biological role. However, operating with these principles results in the requirement to annotate metabolites present in each biological study rather than have a predefined large list of metabolites whose annotation is already known and which lead to a rapid conversion of data to knowledge in all studies. The use of knowledge related to previously reported metabolites (for example, in metabolomic databases like the Human Metabolome Database, HMDB [12]) and Deep Metabolome Annotation (DMA) of each sample type applying advanced analytical approaches (as discussed in section 2) provides a long list of target metabolites that may be present. This allows rapid conversion of data to knowledge for these metabolites in semi-targeted assays, while still allowing unreported or unexpected metabolites not in these lists to be detected and annotated. If the list of metabolites to be detected can be accurately defined then the total search space is also accurately defined. As this search space increases in size to include metabolites that have a low probability to be present (but have to be included in the search) so does the difficulty in providing a single chemical structure as an annotation due to the presence of isomers and an increase in the probability of false annotations. For example, let us investigate the collection of endogenous and exogenous metabolites and peptides defined in the largest metabolite database for humans, the Human Metabolome Database (HMDB) [12]. The current version (v4.0) contains 114,000 metabolites, some of which have been detected in biological samples and some of which are predicted metabolites though not detected/reported in the scientific literature. The inclusion of predicted metabolites increases the complexity of the long list of targeted metabolites and can increase the confidence of any annotations because their inclusion provides the opportunity for previously unreported metabolites to be annotated and therefore a greater biological interpretation of the data to be deduced. Providing a full list of metabolites which can be detected allows more semi-targeted assays to be developed which target those known metabolites only, reduces the risk of false positive annotations, increases the probability of a single annotation and allows a quicker process of converting data to knowledge. A recent review article has discussed the need for Deep Metabolome Annotation of model species and other biological metabolomes [13] and the authors would suggest that this is needed for all metabolomes studied; for example, there are greater than one hundred different metabolomes in the human body when you consider biofluids, cell types and tissues. Applying a multi-platform approach similar to genome sequencing to experimentally and accurately define which metabolites are present (Deep Metabolome Annotation [13]) is urgently needed to enhance the derived knowledge and impact of metabolomics datasets.

A number of excellent reviews are available which discuss metabolite annotation and identification (for example, see [14-15]). In this review article we will describe the current methods and tools available for annotation of metabolites in untargeted metabolomics studies applying LC-MS platforms. A range of different data types, software and databases can be applied in the complete workflow as shown in Figure

2 which depicts the different steps that can (or can not) be applied in the annotation of metabolites. Importantly, approaches defined in proteomics do not necessarily translate to metabolomics and Böcker *et al* recently described this - ‘We can define that everything is better in proteomics because given the genome sequence we can infer all of the database of proteins and second given a peptide sequence we can simulate a MS/MS barcode mass spectrum’ [16]. This is not the case for metabolomics as we will see in the sections below. There are different levels of confidence in the accuracy and robustness of the annotation or identifications reported. The accuracy and robustness are based on which single data type or multiple data types have been applied and whether experimental data for biological samples have been compared to data acquired for authentic chemical standards. These confidence levels should be reported in all studies. The first set of reporting standards were designed by Sumner *et al.* as part of the Metabolomics Standards Initiative in 2007 with four different confidence levels [17]. More recently Schymanski *et al.* have reported a five-level confidence system [18].

2. Deriving the molecular formula from full-scan accurate mass data

As shown in Figure 2, the first process recommended by the authors is to reduce the number of possible metabolite structures to one or a small number. This process is performed primarily with full-scan accurate mass data, but as will be seen later, accurate mass MS/MS or MS^n data can also be applied. This process aims to limit the number of molecular formulae which can represent the measured *m/z* ratio in full scan data (and possibly the MS/MS data also). Let us focus on the full scan data first and its complexity.

In LC-MS all metabolites (and chemicals from non-biological sources) traverse an electrospray ionisation source which converts neutral molecules to a charged state allowing them to be manipulated by electrical and/or magnetic and/or RF energies dependent on their *m/z* ratio. The process of ionisation is complex involving heat, electrical voltages and a complex mixture of chemicals and metabolites; further information is available at [19]. Early work by Brown *et al.* investigated this complexity [5] and further work from the authors of this review has expanded this research (unpublished data). The early research showed that multiple signals (sometimes called metabolite features) are detected for a single metabolite and showed how expected (e.g. $[M+H]^+$, $[M+Na]^+$) and unexpected (e.g. $[M+HCOO]^-$) adducts, ion-source fragments (e.g. loss of ammonia from amino acids), isotopes (e.g. $^{32}S/^{34}S$), unexpected multiply charged ions (e.g. 3+ charge state) and unexpected loss of amino acids from conjugated metabolites (e.g. glycine from glycine-conjugated metabolites) can be created in the electrospray source and detected in LC-MS datasets. A recent publication has demonstrated the complexity observed for one class of metabolites, bile acids, and has shown that the signals were dependent on the specific bile acid, solvent flow rate and bile acid concentration [20]. A list of common adducts and in-source fragmentations are included in Table 2.

In the derivation of these discoveries, the important fact that different signals from the same metabolite have three specific relationships was applied as shown in Figure 3. These are (1) all signals for the same metabolite will have the same retention time as they enter and exit the electrospray ion source at the same time; (2) specific *m/z* differences between signals are commonly observed and not all possible *m/z* differences are experimentally observed and (3) the responses measured across multiple samples for two signals derived from the same metabolite are positively correlated as they are formed in an identical manner (and therefore the ratio for both will be identical across all samples) in the electrospray ion source across all samples. These relationships allow signals derived from the same metabolite to be grouped together and provide greater confidence to the annotation provided. For example, a *m/z* difference of 21.9819 between two signals with the same retention time indicates that the signal with lower mass is a protonated adduct and the signal with the higher mass is a sodiated adduct. Without this *m/z* difference it would be difficult to experimentally define whether the lower mass signal was a protonated adduct or a different type of adduct. The process applied has been reviewed recently [21] signals together and to annotate the 'ion type' to increase the level of confidence for any annotations. These include early and freely available software releases (PUTMEDID_LCMS [22], IDEOM [23] and CAMERA [24]) and more recent additions including MS-FLO [25], CEU Mass Mediator [26], RAMClust [6] and xMSannotator [27]. Metabolite databases including METLIN [28] and HMDB [12] as well as commercial software which also employ these capabilities. Interestingly, a comparison of these software by the authors shows that none use all of the possible ion types as depicted by other software; a robust assessment of all ion types experimentally observed across different manufacturer's instruments and sample types and a standardisation of these across all software is required. Other methods for integrating different signals for the same metabolite have been reported and focus on Bayesian methods (for example see [29]) or known metabolic networks (for example see [30]).

The mass resolution and accuracy of *m/z* measurements is dependent on the LC-MS platform, metabolite concentration, accuracy of mass calibration and whether internal or external mass calibration is applied. A degree of mass error is introduced during data acquisition and this has to be taken in to account when converting a full scan *m/z* signal to a neutral mass and molecular formula. A mass error can be calculated as follows:

Mass error (ppm) =

((experimentally measured *m/z* – theoretical *m/z*) / (theoretical *m/z*)) * 1,000,000

Typically, errors in *m/z* measurements for matching experimental to theoretical *m/z* values is < 5ppm which is within the installation specifications for most medium-to-high resolving power instruments. However, do consider the effects of space-charge effects and the influence of response on mass accuracy where lower intensity ions or ions with a saturated response may be measured with a poorer mass accuracy than for a simple solution of high concentration chemicals infused during installation. Also, two signals which are not mass resolved and are observed as a single *m/z* signal will have a measured *m/z* reported between the *m/z* of each signal and therefore an accurate *m/z* for each metabolite would not be detected. Also the error in an experimentally measured *m/z* difference applied when integrating different signals of the same metabolite (as shown above for protonated and sodiated adducts) has to be considered. For example, the *m/z* difference between ^{12}C and ^{13}C isotopic peaks are commonly applied to de-isotope data by removing ^{13}C isotopic peaks or to calculate the number of carbons present in a molecular formula (as described in the next paragraph). The *m/z* difference between ^{12}C and ^{13}C isotopic peaks is theoretically 1.0033 and if we assume an error range of +/-0.0005 then the mass error for a metabolite of measured *m/z* values for the ^{12}C and ^{13}C ions of 50.0033 and 51.0039 is 10ppm whereas for a metabolite of measured *m/z* values for the ^{12}C and ^{13}C ions of 100.0033 and 101.0039 then the mass error is 5ppm. However, many software and databases apply one single mass error value of 5ppm or less for all processes applied. Two mass errors should be defined, one for matching of experimental *m/z* to theoretical *m/z* data in databases and one for *m/z* differences for grouping of metabolite features.

The mass resolution (or resolving power) of the LC-MS platform can also be important in deriving information from these complex electrospray datasets. For example, the theoretical *m/z* for glucose ($[M+H]^+$) is 181.0707, for the $^{13}C_2$ ion is 183.0773 and for the ^{18}O is 183.0750. Although, the $[M+H]^+$ can be applied to derive the monoisotopic mass accurately, only high resolving power platforms can achieve accurate discrimination of the $^{13}C_2$ ion and the ^{18}O ion to aid in filtering of possible molecular formulas. Studies on hybrid Orbitrap instruments have assessed the influence of mass resolution and have shown that a mass resolution of 60,000 – 120,000 (FWHM at *m/z* 200) is required to maximise the number of signals detected by ensuring resolution of ions of very similar but not identical *m/z* values [31]. More recent work has shown the enhancements achievable in metabolite annotation at higher mass resolutions, specifically a high-field Orbitrap Fusion instrument operated at a mass resolution of 500,000 (FWHM at *m/z* 200) [2].

When applying the experimentally measured *m/z* value to derive a single or multiple molecular formula(s) there are two processes. The first converts the measured *m/z* value to a neutral mass and subsequently calculates the possible molecular formula(s) which match this neutral mass within a given mass error and then matches this molecular formula to metabolites present in metabolomic or chemical databases. The second option uses the measured *m/z* value and knowledge of ion type and searches directly for the metabolite in metabolomic or chemical databases without the step of converting to a molecular formula. One question often raised is whether we go from *m/z* to metabolite while ignoring the conversion to a molecular formula step? Our knowledge of all possible molecular formulae is greater than our knowledge of which metabolites we are expecting to detect (as discussed in section 1). Therefore, the conversion of raw data to a molecular formula can be the first step applied and which if no annotation to a specific metabolite(s) can be made still leaves the researcher with some information on the metabolite's identity. If the measured *m/z* was used directly to search metabolite databases then although matches to metabolites will be made there is the probability that some metabolites will not match to a metabolite in a database and no useful information (the molecular formula) is then available. In these cases then the conversion of *m/z* to molecular formula can be performed in a second subsequent process. Importantly, the application of DMA allows us to link a greater number of metabolites to known molecular formulae, which increases the number of annotated metabolites in databases and reduces the probability of no match to a metabolite when a measured *m/z* was used directly to search metabolite databases.

The molecular formula can also be derived with the use of MS/MS or MS^n data. Here, the possible molecular formula for the full-scan precursor signal and each of the product ions can be calculated. The precursor ion's molecular formula has to be constrained by the combination of the molecular formula for each of the product ions. For example, for a metabolite with an unknown precursor molecular formula, if one of the product ions possible molecular formula contains a sulphur atom then all possible precursor molecular formulas without a sulphur atom present can be deleted from the list of possible molecular formulas. Through this process it is possible to reduce the number of molecular formula for the precursor ion and therefore reduce the number of possible metabolites which match the list of molecular formulas (for an example see [32]). Another more traditional and common approach to filter the list of possible molecular formula is to use isotopic information. The most commonly used isotopic information is the $^{12}C/^{13}C$ relative isotope abundance (RIA) where each carbon in the molecular formula will contribute 1.1% to the ^{13}C isotopic peak intensity. Therefore, if you have a $^{12}C/^{13}C$ RIA of 11% the only molecular formulas of or close to containing 10 carbon atoms (11/1.1) are possible, a molecular formula with 18 carbons is much less probable. Figure 4 shows an example of this. Other elements which have two detectable isotopes can also be applied for filtering possible molecular formula and include $^{32}S/^{34}S$, $^{35}Cl/^{37}Cl$, $^{78}/^{80}Se$ and $^{79}/^{81}Br$. Fiehn *et al.* defined that even with a 1ppm mass accuracy a single molecular formula is not always achievable, and demonstrated that using a 3ppm mass error and combining with the use of isotopic filtering is more appropriate for filtering potential molecular formula than using a sub-ppm mass error on its own [33]. Recently a proposal to enhance van Krevelen diagrams from O:C and H:C to C:H:N:O:P stoichiometry has been reported to allow filtering of feasible and non-feasible molecular formula [34].

By applying the processes discussed in this section a researcher can start with a large search space of metabolites and quickly reduce the size of this search space to a single or small number of molecular formula. This single or multiple molecular formulae can be matched to metabolites present in metabolomic or chemical databases to derive a putative list of metabolite annotations ready for further data analysis to increase the confidence or reduce the list of putatively annotated metabolites. Of course, we must remember that a single empirical formula can represent multiple different isomers.

3. Deriving metabolite structures from gas-phase fragmentation mass spectra

In section 2 we discussed how a researcher can quickly and efficiently reduce the number of possible annotations for a single metabolite. These processes typically do not provide a single metabolite structure. The collection and use of mass spectral fragmentation data to increase the confidence of a single annotation or to reduce the list of possible annotations further should also be performed as a standard process. A recent review from Fenaille and colleagues discussed gas phase fragmentation and its current advantages and limitations in great depth [35].

The 'traditional' experimental approach to acquire MS/MS data is Data Dependent Analysis (DDA), where a top 'n' method is applied. Here, a pre-scan is collected and the top 'n' most intense precursor ions are then separately isolated and gas-phase fragmentation performed with the product ion mass spectrum collected (see [35] for further information on DDA experiments and also see Figure 5). This approach provides a high purity for the product ion MS/MS mass spectrum as the isolation window is small and ideally this window only contains ions for a single *m/z* value.

However, the limitation is that for the relatively short analysis times applied in untargeted metabolomics (typically less than 15 minutes) and because of the duty cycle of mass spectrometers, not all metabolites will have a MS/MS mass spectrum collected. If MS/MS data has to be acquired then further targeted analyses have to be performed focused on these metabolites; for example a nearline approach has been reported by Neumann et al. [36]. However, the intelligent use of DDA methods is currently limited though intelligent use of DDA methods is advised. For example, Mullard *et al.* showed that intelligent use of the precursor window range (applying smaller windows to increase the number of lower intensity metabolites with MS/MS data collected), type of gas-phase fragmentation applied (CID in a linear ion trap or HCD in a collision cell) and collision energy applied increases the number of metabolites for which informative MS/MS data can be acquired [37]. Yan and Yan more recently described a similar approach using gas-phase fractionation (similar to the use of smaller precursor window ranges as described above) and a staggered mass range [38]. Wang *et al* have reported enhanced MS/MS coverage using a target-directed DDA with a time staggered precursor ion list [39]. Here, a full-scan only run was performed and applied to develop a target list with associated retention time windows and the approach showed a greater number of metabolites with MS/MS data especially for lower abundance ions even in areas of high co-elution. The collision energy applied is also important and the study by Mullard *et al.* showed that fragmentation to generate an informative MS/MS mass spectrum is highly collision energy dependent and that no single collision energy should be applied [37]. If a single collision energy was applied then this could result in no fragmentation or too much fragmentation, both resulting in a lack of usable information. Today, stepped or multiple collision energies can provide a greater probability of obtaining an informative MS/MS mass spectrum. Another intelligent DDA approach is to use different DDA experiments for different biological samples in a study where collision energy and other parameters are different for each experiment (see [37]). This assumes that the qualitative composition of all samples is very similar and raises the question as to whether a metabolite has to be annotated once per study or once per study sample. The answer is dependent on the study and in studies where the qualitative composition of samples is very similar then metabolite annotation once per study can be appropriate. However, in studies where the qualitative composition of samples is very different then intelligent DDA applied to each biological sample or a pool of samples from each biological class (e.g. wildtype and mutant) should be performed. One other intelligent DDA method is to use intelligent inclusion and exclusion lists and multiple sample analyses where metabolites already fragmented are added to the exclusion list to allow less intense metabolites to be chosen for MS/MS fragmentation in subsequent samples. Koelmel *et al.* have applied a strategy where *m/z* peaks for which MS/MS data have been acquired are excluded from further analyses by being added to the exclusion list and provided 40-69 % more molecular identifications in a lipidomic positive ion mode study [40]. Intelligent inclusion lists which include only one signal for each metabolite will also maximise the number of metabolites for which MS/MS mass spectra are acquired when compared to multiple signals for each metabolite. This is important also because some adducts tend to produce less informative MS/MS mass spectra compared to other ion types. In conclusion, further research is required to experimentally determine how efficient an intelligent DDA method can be and to what percentage of metabolites informative MS/MS spectra can be acquired.

A more recent addition to the experimental toolbox is Data Independent Analysis (DIA), also called SWATH. This strategy operates by isolating wider precursor *m/z* regions than observed in DDA experiments and stepping these across the whole precursor *m/z* range so that all precursor ions are fragmented (see [35] for further information on DIA experiments and also see Figure 4). This approach has the benefit that any precursor ion above a specific signal will be fragmented and its MS/MS mass spectrum collected. However, the size of the isolation window will influence the number of *m/z* peaks isolated and therefore simultaneously fragmented; if the precursor *m/z* window contains more than one *m/z* peak then computational deconvolution is required to construct the pure MS/MS mass spectrum for each of the metabolites present. This is the case where one DIA window representing the whole precursor *m/z* range is applied, defined as All Ion Fragmentation (AIF). The deconvolution process is accurate when the MS/MS mass spectra of the metabolites are different and the peak shape and retention time (including peak apex) are different. However, if the peaks completely overlap and have the same peak shape and the MS/MS mass spectra are similar then inaccuracies can be observed. The number of DIA windows and their *m/z* width is dependent on the scan rate of the instrument, the higher the scan rate then the smaller the window size and the lower the probability for two or more ions to be present. Again the use of intelligent DIA experiments can enhance data information, for example, using variable DIA window sizes [41]. Another approach applied in proteomics and which has the potential to be transferred to metabolomics is to apply a DIA metabolite library which contains information on RT and MS/MS mass spectrum and then searches for each of these metabolites only. This allows confirmation of a metabolite's presence for known metabolites while still allowing unknown metabolites to be fragmented and MS/MS data collected for further interpretation (see [42,43] for examples).

The usefulness of the MS/MS mass spectrum collected for a metabolite is dependent on whether other ions of different *m/z* values were present in the isolation window and on their intensity in relation to the target precursor ion. A recent publication has taken an approach to assess the purity of a defined precursor *m/z* value in a defined isolation window from raw mass spectral data (msPurity). The results showed that the purity varied considerably within studies but (as expected) that data acquired using a DDA approach provided higher purities than data acquired using a DIA approach [44]. A low purity MS/MS spectrum has the possibility to not be matched to its correct metabolite in a mass spectral library with a suitably high match score when product ions derived from impurities in the isolation window are present in the product ion mass spectrum.

One question raised routinely in the metabolomics community is whether MS/MS data is sufficient for metabolite annotation. The majority of instruments available today only allow MS/MS data to be acquired unless in-source fragmentation is applied as a first stage of fragmentation. However, hybrid Orbitrap instruments allow MS^n data to be acquired where n>2. Here gas-phase fragmentation of product ions can be performed in multiple stages. For low complexity chemical structures, MS/MS data can suffice but for more complex structures and where two structurally similar isomers can differ only in the position of a single functional group then MS^n data provides increased confidence and accuracy (see [45,46] for examples of this application applying offline and online approaches). The collection of MS^n data online during LC analyses does have limitations in that when collecting high mass resolution MS^n mass spectra the number of unique metabolites where DDA data has been acquired is lower compared to when MS/MS DDA experiments are performed. However, in hybrid Orbitrap instruments low mass resolution MS^n DDA experiments can be applied much more quickly when using the linear ion trap for mass analysis and indeed a larger number of DDA experiments can be performed when applying this approach because of the fast cycle time and higher sensitivity of the linear ion trap compared to the Orbitrap mass analyser. The mass analyser applied for mass analysis, and the associated mass resolution, is a choice for the researcher and is a balance between speed, mass resolution and number of unique metabolites with informative MS^n data collected. Ideally, all MS and MS^n data would be collected at a high mass resolution to ensure separation of isobaric product ions, though this reduces the number of metabolites for which MS^n data can be acquired in a single chromatographic run. However, intelligent experiments applying inclusion and exclusion lists and multiple injections of the same sample can be applied to increase the number of metabolites with MS^n data acquired using a high mass resolution [40].

Where high quality MS/MS spectra are acquired then searching of the experimental MS/MS or MS^n data to data available in a mass spectral library is performed. There are many different mass spectral libraries available which focus either on metabolites only (e.g. mzCloud [47], METLIN [28] and MoNA [48]) or more broadly on chemicals which include metabolites (e.g. NIST18 [49]). These libraries are constructed by the analysis of pure authentic chemical standards and the inclusion of the MS/MS data in to the libraries. As discussed above the collision energy applied during experimental acquisition of MS/MS mass spectra influences the information content of the MS/MS mass spectrum and therefore collection of MS/MS mass spectra at different collision energies is recommended; many mass spectral libraries now do this including METLIN which uses three different collision energies [28] and mzCloud which uses up to twenty different collision energies [47]. Yanes *et al* recently reviewed the metabolites present across different mass spectral libraries and found that many MS/MS libraries include unique metabolites not included in any other MS/MS library and therefore a search of multiple libraries is recommended (METLIN, GNPS, NIST14 and MassBank provide the greatest number of unique metabolites, all over 40% of the total library were unique) [50]. Matching of a metabolite to a MS/MS mass spectra in a mass spectral library increases the confidence of annotation though matching with high match scores to multiple metabolites is commonly observed where different metabolites have similar or identical MS/MS mass spectra (for example, leucine and isoleucine). Therefore, caution should always be applied with mass spectral library searches as false positive, false negative or a lack of matches can be observed. The inclusion of complementary data (for example, retention time) is always recommended where possible to increase the confidence and robustness of the annotation or identification.

Importantly, metabolites can only be included in these experimentally-derived mass spectral libraries if they are available to be purchased or are synthesised and then analysed. Many metabolites are not available as authentic chemical standards and therefore can not be included in mass spectral libraries. This provides a significant quandary and two options are available. The first option is to construct in-silico mass spectral libraries which have been a large success in proteomics. However, the number and complexity of gas-phase fragmentations for proteins is much lower compared to the more structurally diverse range of metabolites and therefore this is an approach that has started to be applied in metabolomics but requires further developments and global application. The most widely applied in-silico MS/MS library currently is LipidBlast [51] and is applied to lipids whose structural similiarity makes it easier to apply the rules for one metabolite in a lipid class to all other metabolites in that class. Other libraries are available (e.g. [52]). The ability to accurately do this for the more structurally diverse water-soluble metabolites is currently limited but needs to be solved. The inclusion of quantum mechanical calculations can enhance the accuracy of in-silico MS/MS mass spectral construction and requires further developments. mzCloud applies these approaches within its library and an example of how this can be used is available at [53]. Another in-silico approach is to perform in-silico fragmentation on all metabolites remaining after filtering based on full-scan data (as discussed in section 1). Here fragmentation for all metabolites is performed in-silico and each in-silico MS/MS spectrum is then compared to the experimental MS/MS mass spectrum with a match score provided. Examples of freely available software which apply this strategy include MetFrag [54] and MS-FINDER [55]. The accuracy of this approach is dependent on the number of different fragmentation mechanisms allowed to be performed (see [56] for a good review). Mass spectral fragmentation is complex and therefore fragmentation libraries should be large to allow all mechanisms to have the potential to be included. The software with the most comprehensive list of fragmentation mechanisms is the commercially available MassFrontier [57], whose fragmentation library is comprehensive and is derived from the scientific literature. MassFrontier can not be operated directly in a batch mode, a newer software called HAMMER allows batch operations to be performed [58]. Finally, a large volume of MS/MS mass spectra for a diverse range of metabolites and chemicals are already available and can be applied to assist in the annotation of metabolites whose MS/MS mass spectra are not available. For example, the recent introduction of MS2LDA has driven this area forward offering an unsupervised method (inspired by text-mining) that extracts common patterns of mass fragments and neutral losses —Mass2Motifs— from collections of fragmentation spectra. Structurally characterized Mass2Motifs can be used to annotate molecules for which no reference spectra exist and expose biochemical relationships between molecules [59]. Treutler *et al* have shown how regulated metabolite families can be discovered using DIA LC-MS data and Hierarchical Clustering Analysis, the software is called MetFamily [60]. CASMI challenges have recently been used to investigate different in-silico tools for metabolite annotation [61,62].

4. Chromatographic retention time

Although the use of full scan and MS/MS and/or MS^n mass spectral data can provide a lot of information to be applied in the annotation of metabolites, other complementary data should be applied to provide greater confidence in reported annotations. Here we will discuss one of the data types.

Chromatographic retention times are based on a different property of the metabolite compared to mass. Here, the physicochemical properties of how a metabolite interacts with a stationary phase is the defining property being measured and these interactions can be optimised by changing solvents, stationary phase, column dimensions and temperature. As discussed above the combination of MS/MS mass spectra and retention times in metabolite libraries is possible where authentic chemical standards are available and are applied to experimentally derive retention time. However, unlike MS/MS mass spectra which are somewhat independent of the instrument where the data is acquired (i.e. the mass spectrometer can be operated reproducibly between different laboratories to produce the same MS/MS mass spectrum) the same is not true for the retention time. This is highly dependent on the analytical conditions applied including stationary phase, solvents, gradient elution, dead volumes; even the same type of column from different manufacturers can have different retention properties (for example, based on carbon coverage or changes in surface chemistry for reversed phase C_{18} stationary phases) means that reproducible retention times is somewhat limited because different laboratories apply different chromatographic parameters. Transferability between laboratories is much less achievable because only a small number of standardised assays are available and routinely applied across different laboratories for untargeted studies. The best examples of a standard assay are the p180 and p400 assay kits available from the company Biocrates which have been applied in different laboratories including in inter-laboratory comparisons (for an example see [63]). These assays are described as semi-targeted as the list of metabolites to be detected are derived before data collection and other metabolites present are not detected. These assays apply a liquid chromatographic assay for a subset of the metabolites and also includes direct infusion for assaying the other metabolites. Standard Reference Materials (SRMs) are available from the National Institute of Standards and Technology (NIST) including the most widely used SRM1950. These SRMs can be applied for method validation, inter-laboratory studies [64] and for development of quality control processes in metabolomics studies. Solutions in the future could follow two routes (1) standardised assays and retention indices or locking and (2) in-silico retention time prediction. The development of standardised untargeted LC-MS assays applied across multiple laboratories has not yet been achieved in the metabolomics community. Efforts are underway within an international consortium to apply standardised assays, called the International Phenome Centre Network, though this is in an early stage [65]. The difficult step is to persuade all groups to use the same LC-MS methods rather than use their own tried and tested assays in which they have the greatest confidence; this includes persuasion to use the same single supplier LC column. With standardised assays then a retention index system can be applied to compensate for small retention time drifts observed between laboratories. A second route is to use in-silico retention time prediction for metabolites where no authentic chemical standard is available (*cf* mass spectral libraries). Here data for known metabolites are applied to develop a prediction model for unknown metabolites, where the possible list of unknown metabolites can be derived from full scan or MS/MS data [66-68]. This strategy is again in early stages of its development and improvements in accuracy and precision are required. Many current models have a prediction accuracy of 30-60 seconds and in typical LC run times of 15 minutes or less many isomers of the same mass have retention times which fall within 10 seconds of each other; therefore increased accuracy is required.

Concluding remarks

We have come a long way in the last ten years in solving the issue of metabolite annotation, one of the major bottlenecks of untargeted metabolomics. Through assessment and characterisation of current methods and data collected, the metabolomics community have identified the complexities and limitations and have developed solutions to overcome these complexities and limitations. Through our growing knowledge of metabolites present in commonly studied metabolomes there has been the start of a move away from fully untargeted metabolomics studies to semi-targeted studies where the list of metabolites to be detected are known prior to data acquisition. However, we must remember that untargeted metabolomics is a game of confidence where all results reported can be assigned a level of confidence (e.g. a statistical p-value defines a level of confidence). This is the case for metabolite annotation in untargeted metabolomics studies where reporting standards related to confidence of an accurate and robust annotation have been presented.. One important conclusion from these reporting standards is that most metabolites reported are annotated and not identified. Identification defines that two or more complementary data types are compared to data collected for an authentic chemical standard applying the same analytical conditions. So retention time-MS/MS libraries constructed in-house allow identification. However, using full scan data only or MS/MS data only which are compared to online mass spectral libraries or metabolomics databases is not sufficient to provide an identification, these are annotations only. Care should always be taken when basing a biological conclusion on one annotated metabolite; validation of this discovery is needed and greater confidence in the importance of annotated metabolites should be based on multiple hits from the same class of metabolite where class can be based on metabolite structure of biological function.

Acknowledgements

The University of Birmingham and Thermo Fisher Scientific are Technology Alliance Partners.

Funding

This work was supported by BBSRC and Thermo Fisher Scientific (BB/L015579/1 and BB/N023013/1).

References

[1] W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre and J. L. Griffin. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. Chemical Society Reviews 40, no. 1 (2011): 387-426.

[2] P. Barbier Saint Hilaire, U. M. Hohenester, B. Colsch, J-C. Tabet, C. Junot, and F. Fenaille. Evaluation of the High-Field Orbitrap Fusion for Compound Annotation in Metabolomics. Analytical chemistry (2018).

[3] H. Tsugawa. Advances in computational metabolomics and databases deepen the understanding of metabolisms. Current opinion in biotechnology 54 (2018): 10-17.

[4] T. Cajka and O. Fiehn. Toward merging untargeted and targeted methods in mass spectrometry-based metabolomics and lipidomics. Analytical chemistry 88, no. 1 (2015): 524-545.

[5] M. Brown, W. B. Dunn, P. Dobson, Y. Patel, C. L. Winder, et al. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst 134, no. 7 (2009): 1322-1332.

[6] N. G. Mahieu and G. J. Patti. Systems-Level Annotation of a Metabolomics Data Set Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. Analytical chemistry 89, no. 19 (2017): 10397-10406.

[7] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, et al. IPO: a tool for automated optimization of XCMS parameters. BMC bioinformatics 16, no. 1 (2015): 118.

[8] R. R. da Silva, P. C. Dorrestein, and R. A. Quinn. Illuminating the dark matter in metabolomics. Proceedings of the National Academy of Sciences 112, no. 41 (2015): 12549-12550.

[9] J. G. Jeffryes, R. L. Colastani, M. Elbadawi-Sidhu, T. Kind, T. D. Niehaus, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of cheminformatics 7, no. 1 (2015): 44.

[10] M. R. Showalter, T. Cajka, and O. Fiehn. Epimetabolites: discovering metabolism beyond building and burning. Current opinion in chemical biology 36 (2017): 70-76.

[11] C. Lerma-Ortiz, J. G. Jeffryes, A. J. L. Cooper, T. D. Niehaus, A. M. K. Thamm, et al. ‘Nothing of chemistry disappears in biology’: the Top 30 damage-prone endogenous metabolites. Biochemical Society Transactions 44, no. 3 (2016): 961-971.

[12] D. S. Wishart, Y. D. Feunang, A. Marcu, A. C. Guo, K. Liang, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic acids research 46, no. D1 (2017): D608-D617.

[13] M. R. Viant, I. J. Kurland, M. R. Jones, and W. B. Dunn. How close are we to complete annotation of metabolomes?. Current opinion in chemical biology 36 (2017): 64-69.

[14] W. B. Dunn, A. Erban, R. J. M. Weber, D. J. Creek, M. Brown, et al. Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. Metabolomics 9, no. 1 (2013): 44-66.

[15] D. S. Wishart. Advances in metabolite identification. Bioanalysis 3, no. 15 (2011): 1769-1782.

[16] S. Böcker. Searching molecular structure databases using tandem MS data: are we there yet?. Current opinion in chemical biology 36 (2017): 1-6.

[17] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale R. Beger, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics 3, no. 3 (2007): 211-221.

[18] E. L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H. P. Singer and J. Hollender. Identifying small molecules via high resolution mass spectrometry: communicating confidence. Environ Sci Technol., 48, no. 4 (2014):2097-8.

[19] L. Konermann, E. Ahadi, A. D. Rodriguez, and S. Vahidi. Unraveling the mechanism of electrospray ionization. (2012): 2-9.

[20] P. Brophy, C. D. Broeckling, J. Murphy, and J. E. Prenni. Ion-neutral Clustering of Bile Acids in Electrospray Ionization Across UPLC Flow Regimes. Journal of The American Society for Mass Spectrometry 29, no. 4 (2018): 651-662.

[21] X. Domingo-Almenara, J. R. Montenegro-Burke, H. P. Benton, and G. Siuzdak. Annotation: a computational solution for streamlining metabolomics analysis. Analytical chemistry 90, no. 1 (2017): 480-489.

[22] M. Brown, D. C. Wedge, R. Goodacre, D. B. Kell, P. N. Baker, et al. Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. Bioinformatics 27, no. 8 (2011): 1108-1112.

[23] D. J. Creek, A. Jankevics, K. E. V. Burgess, R. Breitling, and M. P. Barrett. IDEOM: an Excel interface for analysis of LC–MS-based metabolomics data. Bioinformatics 28, no. 7 (2012): 1048-1049.

[24] C. Kuhl, R. Tautenhahn, C. Bottcher, T. R. Larson, and S. Neumann. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical chemistry 84, no. 1 (2011): 283-289.

[25] B. C. DeFelice, S. S. Mehta, S. Samra, T. Čajka, B. Wancewicz, J. F. Fahrmann, and O. Fiehn. Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography–Mass Spectroscopy (LC-MS) Data Processing. Analytical chemistry 89, no. 6 (2017): 3250-3255.

[26] A. G. de la Fuente, J. Godzien, M. F. López, F. J. Rupérez, C. Barbas, and A. Otero. Knowledge-based metabolite annotation tool: CEU Mass Mediator. Journal of pharmaceutical and biomedical analysis 154 (2018): 138-149.

[27] K. Uppal, D. I. Walker, and D. P. Jones. xMSannotator: an R package for network-based annotation of high-resolution metabolomics data. Analytical chemistry 89, no. 2 (2017): 1063-1067.

[28] C. Guijas, J. R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, et al. METLIN: A technology platform for identifying knowns and unknowns. Analytical chemistry 90, no. 5 (2018): 3156-3164.

[29] R. Daly, S. Rogers, J. Wandy, A. Jankevics, K. E. V. Burgess, and R. Breitling. MetAssign: probabilistic annotation of metabolites from LC–MS data using a Bayesian clustering approach. Bioinformatics 30, no. 19 (2014): 2764-2771.

[30] R. J. M. Weber and M. R. Viant. MI-Pack: Increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. Chemometrics and Intelligent Laboratory Systems 104, no. 1 (2010): 75-82.

[31] L. Najdekr, D. Friedecký, R. Tautenhahn, T. Pluskal, J. Wang, Y. Huang, and T. Adam. Influence of mass resolving power in orbital ion-trap mass spectrometry-based metabolomics. Analytical chemistry 88, no. 23 (2016): 11429-11435.

[32] M. Rojas-Chertó, P. T. Kasper, E. L. Willighagen, R. J. Vreeken, T. Hankemeier, and T. H. Reijmers. Elemental composition determination based on MSn. Bioinformatics 27, no. 17 (2011): 2376-2383.

[33] T. Kind and O. Fiehn. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC bioinformatics 7, no. 1 (2006): 234.

[34] A. Rivas-Ubach, Y. Liu, T. S. Bianchi, N. Tolič, C. Jansson, and L. Paša-Tolić. Moving beyond the van Krevelen diagram: A new stoichiometric approach for compound classification in organisms. Analytical chemistry (2018).

[35] F. Fenaille, P. Barbier Saint-Hilaire, K. Rousseau, and C. Junot. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand?. Journal of Chromatography A (2017).

[36] S. Neumann, A. Thum, and C. Böttcher. Nearline acquisition and processing of liquid chromatography-tandem mass spectrometry data. Metabolomics 9, no. 1 (2013): 84-91.

[37] G. Mullard, J. W. Allwood, R. Weber, M. Brown, P. Begley, et al. A new strategy for MS/MS data acquisition applying multiple data dependent experiments on Orbitrap mass spectrometers in non-targeted metabolomic applications. Metabolomics 11, no. 5 (2015): 1068-1080.

[38] Z. Yan and R. Yan. Improved data-dependent acquisition for untargeted metabolomics using gas-phase fractionation with staggered mass range. Analytical chemistry 87, no. 5 (2015): 2861-2868.

[39] Y. Wang, R. Feng, R. Wang, F. Yang, P. Li, and J-B. Wan. Enhanced MS/MS coverage for metabolite identification in LC-MS-based untargeted metabolomics by target-directed data dependent acquisition with time-staggered precursor ion list. Analytica chimica acta 992 (2017): 67-75.

[40] J. P. Koelmel, N. M. Kroeger, E. L. Gill, C. Z. Ulmer, J. A. Bowden, et al. Expanding lipidome coverage using LC-MS/MS data-dependent acquisition with automated exclusion list generation. Journal of The American Society for Mass Spectrometry 28, no. 5 (2017): 908-917.

[41] Y. Zhang, A. Bilbao, T. Bruderer, J. Luban, C. Strambio-De-Castillia, et al. The use of variable Q1 isolation windows improves selectivity in LC–SWATH–MS acquisition. Journal of proteome research 14, no. 10 (2015): 4359-4371.

[42] T. Bruderer, E. Varesio, A. O. Hidasi, E. Duchoslav, L. Burton, R. Bonner, and G. Hopfgartner. Metabolomic spectral libraries for data-independent SWATH liquid chromatography mass spectrometry acquisition. Analytical and bioanalytical chemistry 410, no. 7 (2018): 1873-1884.

[43] Y. Chen, Z. Zhou, W. Yang, N. Bi, J. Xu, et al. Development of a Data-Independent Targeted Metabolomics Method for Relative Quantification Using Liquid Chromatography Coupled with Tandem Mass Spectrometry. Analytical chemistry 89, no. 13 (2017): 6954-6962.

[44] T. N. Lawson, R. J. M. Weber, M. R. Jones, A. J. Chetwynd, G. Rodríguez-Blanco et al. msPurity: automated evaluation of precursor ion purity for mass spectrometry-based fragmentation in metabolomics. Analytical chemistry 89, no. 4 (2017): 2432-2439.

[45] J. J. J. van der Hooft, J. Vervoort, R. J. Bino, and R. C. H. de Vos. Spectral trees as a robust annotation tool in LC–MS based metabolomics. Metabolomics 8, no. 4 (2012): 691-703.

[46] J. Václavík, K. L. M. Coene, I. Vrobel, L. Najdekr, D. Friedecký et al. Structural elucidation of novel biomarkers of known metabolic disorders based on multistage fragmentation mass spectra. Journal of Inherited Metabolic Disease (2017): 1-8.

[47] https://www.mzcloud.org/; last accessed May 3rd 2018

[48] http://mona.fiehnlab.ucdavis.edu/; last accessed May 3rd 2018

[49] https://www.nist.gov/srd/nist-standard-reference-database-1a-v17; last accessed May 3rd 2018

[50] M. Vinaixa, E. L. Schymanski, S. Neumann, M. Navarro, R. M. Salek, and O. Yanes. Mass spectral databases for LC/MS-and GC/MS-based metabolomics: state of the field and future prospects. TrAC Trends in Analytical Chemistry 78 (2016): 23-35.

[51] T. Kind, K-H. Liu, D. Y. Lee, B. DeFelice, J. K. Meissen, and O. Fiehn. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nature methods 10, no. 8 (2013): 755.

[52] F. Allen, R. Greiner, and D. Wishart. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics 11, no. 1 (2015): 98-110.

[53] B.G. Janesko, L. Li, and R. Mensing. Quantum Chemical Fragment Precursor Tests: Accelerating de novo annotation of tandem mass spectra. Analytica chimica acta 995 (2017): 52-64.

[54] C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. Journal of cheminformatics 8, no. 1 (2016): 3.

[55] H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Analytical chemistry 88, no. 16 (2016): 7946-7958.

[56] D. P. Demarque, A. E. M. Crotti, R. Vessecchi, J. L. C. Lopes, and N. P. Lopes. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. Natural product reports 33, no. 3 (2016): 432-455.

[57] http://www.highchem.com/index.php/component/content/article?id=81; last accessed May 3rd 2018

[58] J. Zhou, R. J. M. Weber, J. W. Allwood, R. Mistrik, Z. Zhu, et al. HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. Bioinformatics 30, no. 4 (2013): 581-583.

[59] J. J. J. van Der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess, and S. Rogers. Topic modeling for untargeted substructure exploration in metabolomics. Proceedings of the National Academy of Sciences 113, no. 48 (2016): 13738-13743.

[60] H. Treutler, H. Tsugawa, A. Porzel, K. Gorzolka, A. Tissier, S. Neumann, and G. U. Balcke. Discovering regulated metabolite families in untargeted metabolomics studies. Analytical chemistry 88, no. 16 (2016): 8082-8090.

[61] F. Allen, R. Greiner, and D. Wishart. CFM-ID Applied to CASMI 2014. Current Metabolomics 5, no. 1 (2017): 35-39.

[62] I. Blaženović, T. Kind, H. Torbašinović, S. Obrenović, S. S. Mehta, et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. Journal of cheminformatics 9, no. 1 (2017): 32.

[63] A. P. Siskos, P. Jain, W. Römisch-Margl, M. Bennett, D. Achaintre, et al. Interlaboratory reproducibility of a targeted metabolomics platform for analysis of human serum and plasma. Analytical chemistry 89, no. 1 (2016): 656-665.

[64] J.A. Bowden, A. Heckert, C. Z. Ulmer, C. M. Jones, J. P. Koelmel, et al. Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950-Metabolites in Frozen Human Plasma. J. Lipid Res., 58, no. 12 (2017): 2275-2288.

[65] https://phenomenetwork.org/; last accessed May 3rd 2018

[66] J. Stanstrup, S. Neumann and U. Vrhovšek. PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. Analytical chemistry 87, no. 18 (2015): 9421-9428.

[67] F. Falchi, S. M. Bertozzi, G. Ottonello, G. F. Ruda, G. Colombano, et al. Kernel-based, partial least squares quantitative structure-retention relationship model for UPLC retention time prediction: A useful tool for metabolite identification. Analytical chemistry 88, no. 19 (2016): 9510-9517.

[68] D. J. Creek, A. Jankevics, R. Breitling, D. G. Watson, M. P. Barrett, and K. E. V. Burgess. Toward global metabolomics analysis with hydrophilic interaction liquid chromatography–mass spectrometry: improved metabolite identification by retention time prediction. Analytical chemistry 83, no. 22 (2011): 8703-8710.

Figures and Tables

Figure 1. The range of factors which influence biofluid and tissue metabolomes in the human population including the intake of metabolites and other chemicals from the environment or other microbial genomes, physical characteristics including age as well as life and work choices including levels of exercise.

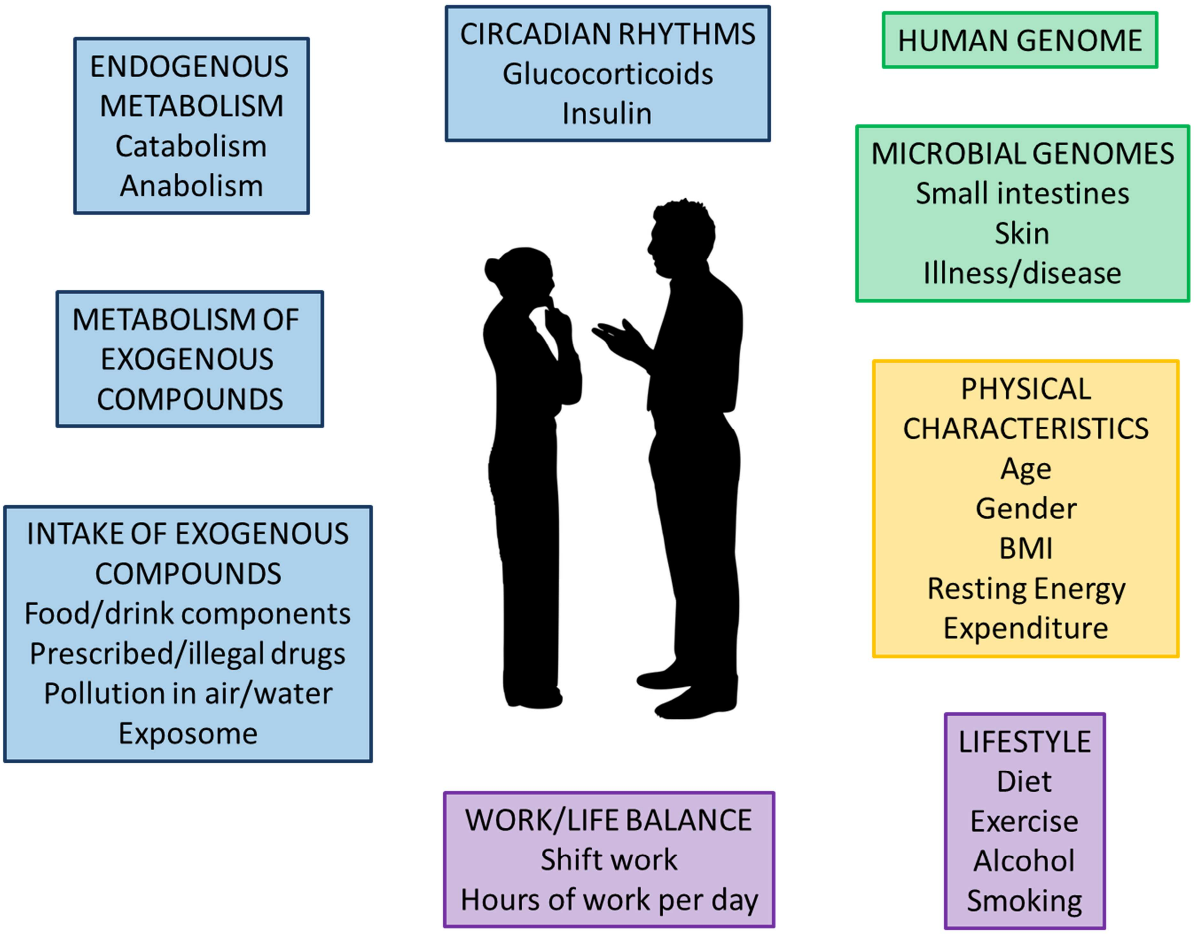


Figure 2. The process of metabolite annotation or identification in untargeted metabolomics studies applying LC-MS. The use of full scan data (blue) and MS/MS or MS^n data (purple) are applied routinely for the annotation of metabolites. Increasingly the use of in-silico approaches (in-silico mass spectral libraries or in-silico prediction of properties; purple and green) are being observed. Most metabolites are annotated (and should be reported as annotated) unless two complementary properties are matched to the same properties for a chemical standard analysed applying identical analytical conditions for biological sample and chemical standard. The reporting of confidence in the annotation or identification should be performed and reporting standards are available [17,18].

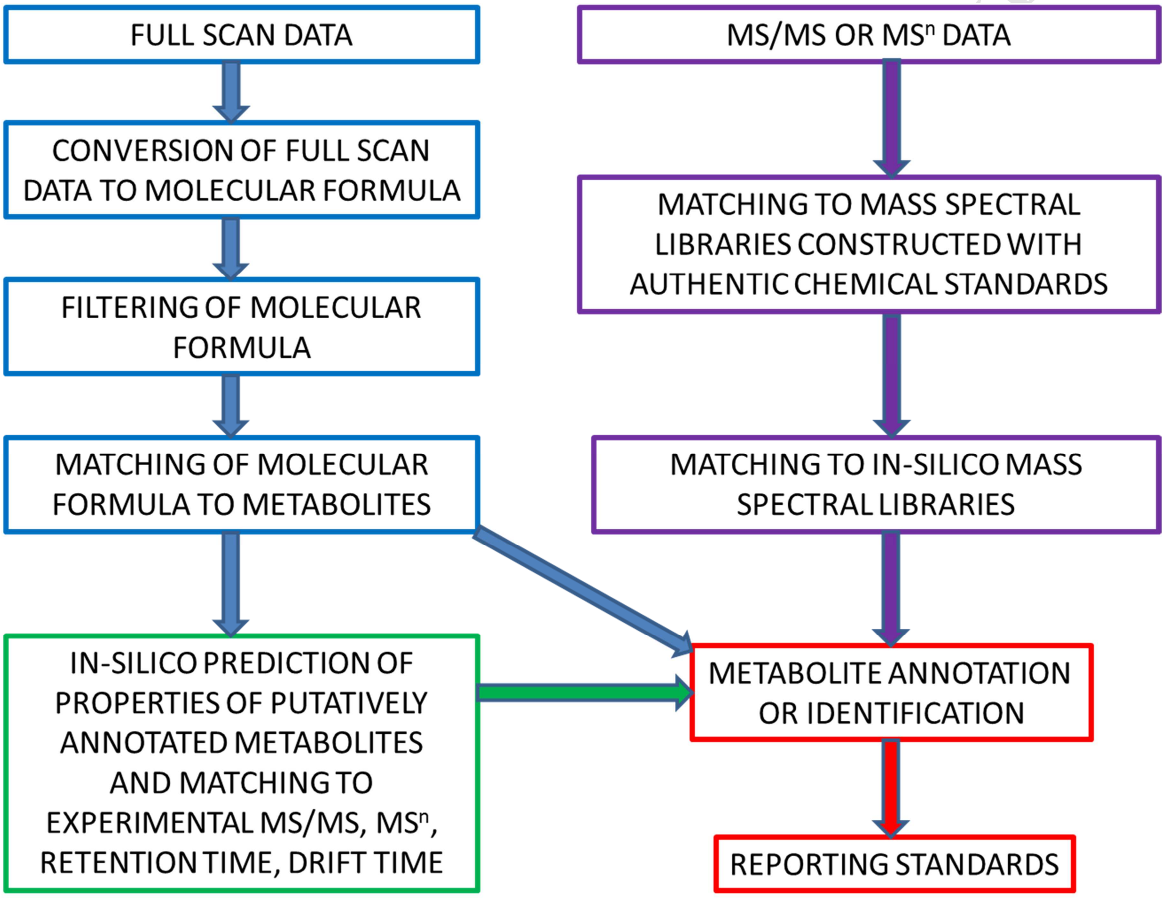

FULL SCAN DATA
CONVERSION OF FULL SCAN DATA TO MOLECULAR FORMULA
FILTERING OF MOLECULAR FORMULA
MATCHING OF MOLECULAR FORMULA TO METABOLITES
IN-SILICO PREDICTION OF PROPERTIES OF PUTATIVELY ANNOTATED METABOLITES AND MATCHING TO EXPERIMENTAL MS/MS, MSⁿ, RETENTION TIME, DRIFT TIME
MS/MS OR MSⁿ DATA
MATCHING TO MASS SPECTRAL LIBRARIES CONSTRUCTED WITH AUTHENTIC CHEMICAL STANDARDS
MATCHING TO IN-SILICO MASS SPECTRAL LIBRARIES
METABOLITE ANNOTATION OR IDENTIFICATION
REPORTING STANDARDS

Figure 3. The use of full scan accurate mass data to derive information and reduce complexity in metabolite annotation. Different signals for the same metabolite have the same retention time and peak shape (A), responses for pairs of signals are positively correlated (B) and specific *m/z* differences are observed and these *m/z* differences are not random (C).

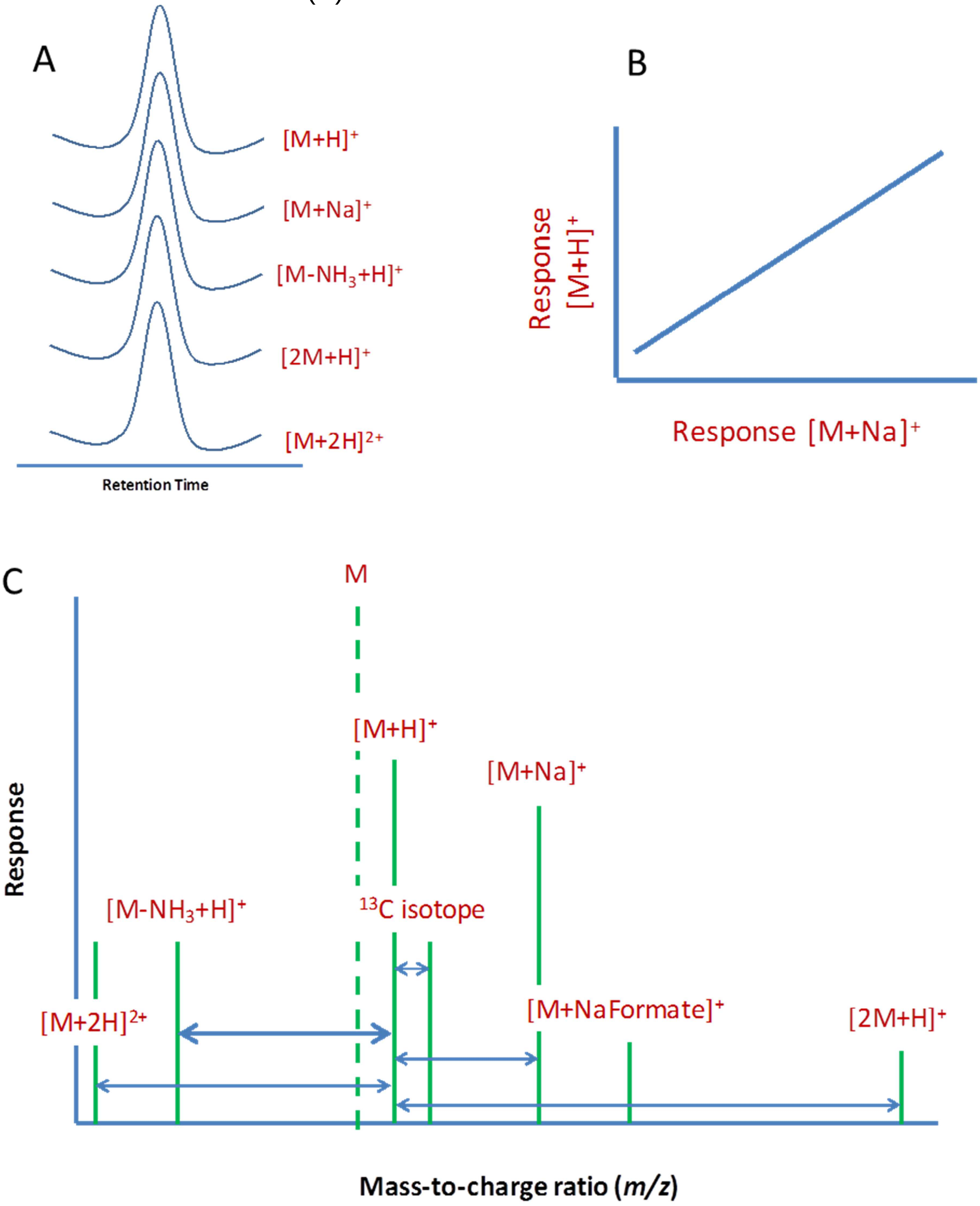


Figure 4. Calculation of the Relative Isotope Abundance and its use for filtering lists of potential molecular formula. The peak area for the ^{12}C and ^{13}C peaks are normalized to the peak area of the ^{12}C peak (100%). In this example, the ^{13}C normalized peak area is 33%. The relative isotope abundance is calculated by dividing the ^{13}C normalized peak area by 1.1, in this example to produce a RIA of 30. The metabolite therefore must have a molecular formula containing approximately 30 carbon atoms; here all molecular formula with 30 +/- 10% carbon atoms are potential molecular formula while molecular formulae outside this range are removed.

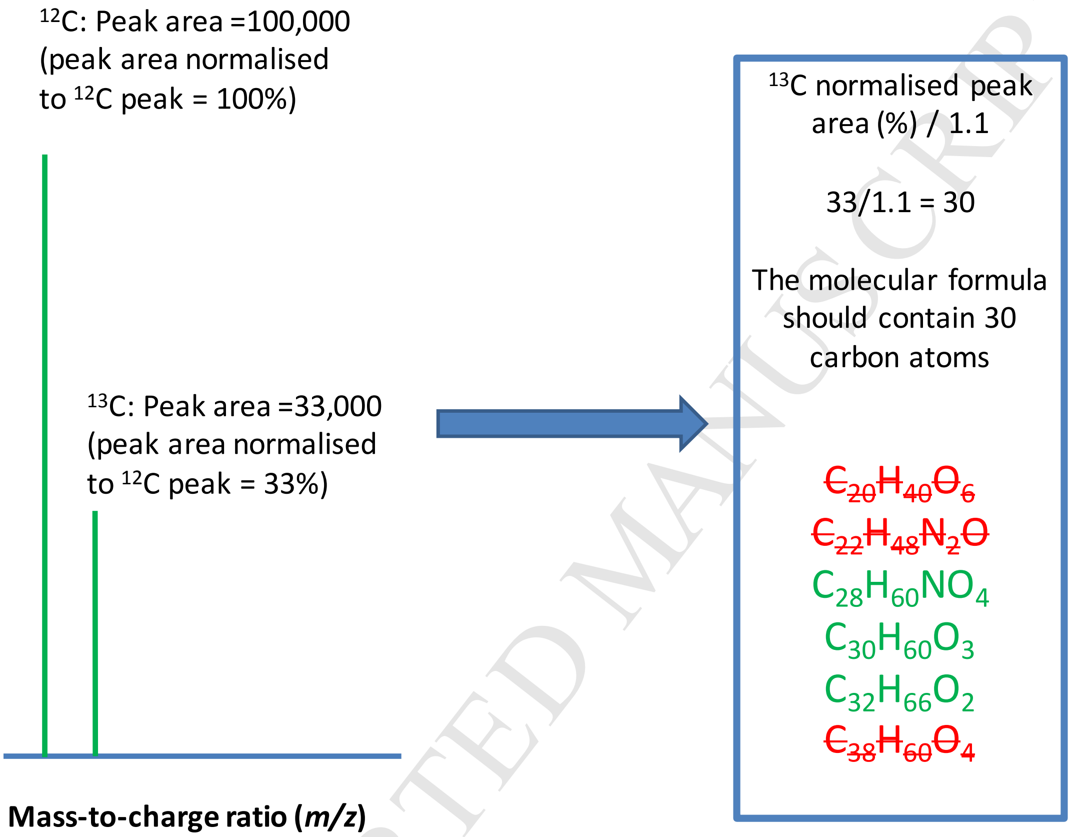

^{12}C: Peak area =100,000 (peak area normalised to ^{12}C peak = 100%)
^{13}C: Peak area =33,000 (peak area normalised to ^{12}C peak = 33%)
Mass-to-charge ratio (*m/z*)
^{13}C normalised peak area (%) / 1.1
33/1.1 = 30
The molecular formula should contain 30 carbon atoms
~~$C_{20}H_{40}O_6$~~
~~$C_{22}H_{48}N_2O$~~
$C_{28}H_{60}NO_4$
$C_{30}H_{60}O_3$
$C_{32}H_{66}O_2$
~~$C_{38}H_{60}O_4$~~

Figure 5. The processes applied to perform DDA and DIA experiments differ in the width of the isolation window (DDA=narrow, DIA=wider), the coverage of the precursor *m/z* range (DDA=lower coverage, DIA=complete coverage) and the purity of the signal in the isolation window (DDA=higher purity, DIA=lower purity).

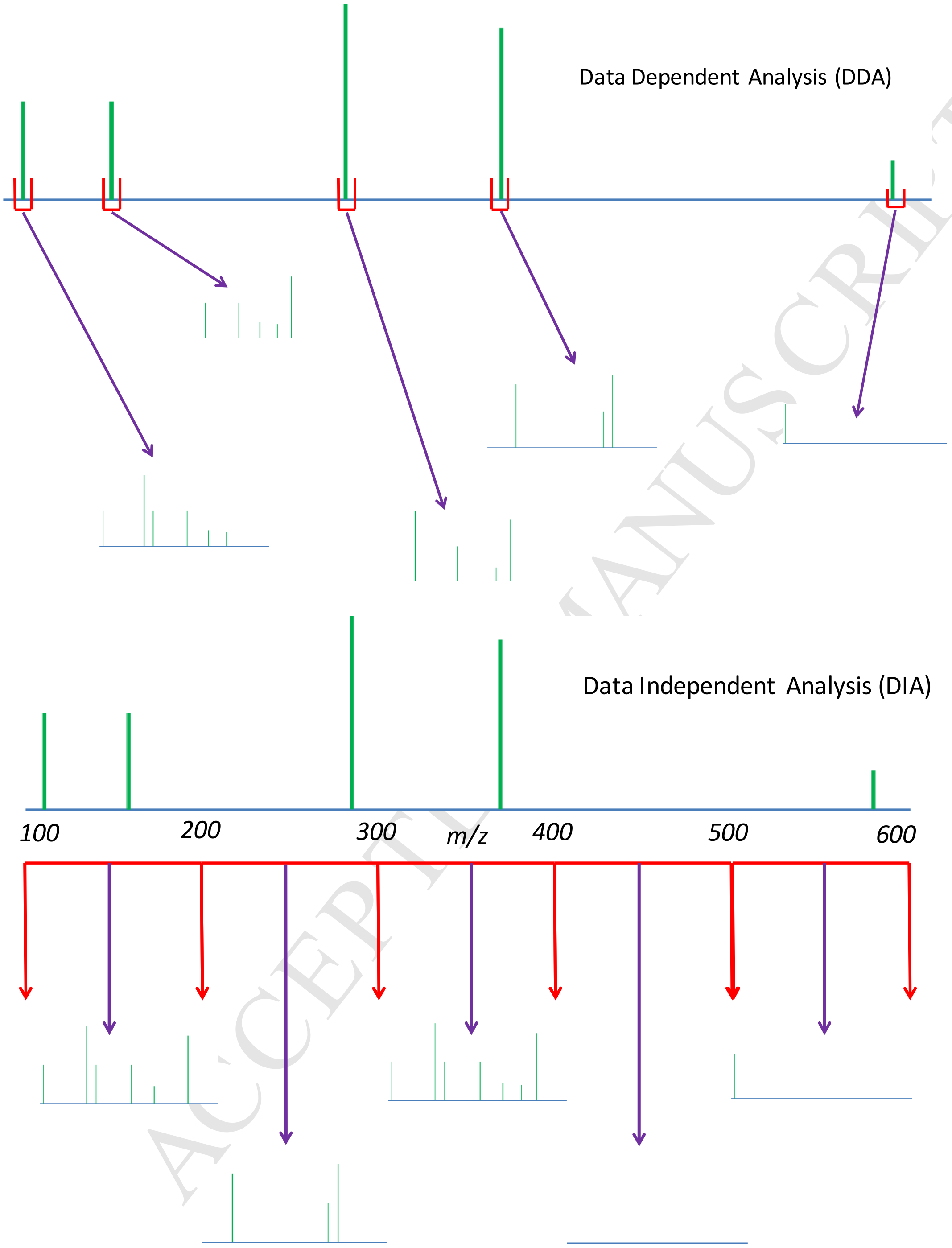

Data Dependent Analysis (DDA)
Data Independent Analysis (DIA)
100
200
300
m/z
400
500
600

Table 1. A list of open access software, databases and libraries for metabolite annotation and identification.

Software/Database	Annotation from full scan data	Annotation from from MS/MS or MS^n data	In-Silico Tool	Web Address
Golm Metabolome Database (GMD)	x			http://gmd.mpimp-golm.mpg.de/
Toxin and Toxin Target Database (T3DB)	x			http://www.t3db.ca/
FooDB	x			http://foodb.ca/
DrugBank	x			https://www.drugbank.ca/
Human Metabolome Database (HMDB)	x			http://www.hmdb.ca/
KEGG	x			http://www.genome.jp/kegg/
PubChem	x			https://pubchem.ncbi.nlm.nih.gov/
ChEBI	x			https://www.ebi.ac.uk/chebi/
BioCyc	x			https://biocyc.org/
HumanCyc	x			https://humancyc.org/
LipidMAPS	x			http://www.lipidmaps.org/
ChemSpider	x			http://www.chemspider.com/
MINE	x			http://minedatabase.mcs.anl.gov/
Recon2	x			http://www.ebi.ac.uk/biomodels-main/MODEL1109130000
PUTMEDID_LCMS	x			http://www.mcisb.org/resources/putmedid.html
IDEOM	x			http://mzmatch.sourceforge.net/ideom.php
CAMERA	x			https://bioconductor.org/packages/release/bioc/html/CAMERA.html
MS-FLO	x			http://msflo.fiehnlab.ucdavis.edu/
CEU Mass Mediator	x			http://ceumass.eps.uspceu.es/
xMSannotator	x			https://sourceforge.net/projects/xmsannotator/
MZedDB	x			http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html
Rdisop	x			https://bioconductor.org/packages/release/bioc/html/Rdisop.html
SIRIUS	x			https://github.com/boecker-lab/sirius
MI-Pack	x			https://github.com/Viant-Metabolomics/MI-Pack
ProbMetab	x			http://labpib.fmrp.usp.br/methods/probmetab/
MetAssign-mzMatch	x			http://mzmatch.sourceforge.net/
RAMClust	x	x		https://rdrr.io/github/cbroeckl/RAMClustR/
MyCompoundID	x	x		http://www.mycompoundid.org/mycompoundid_IsoMS/
METLIN	x	x		https://metlin.scripps.edu/
MassBank		x		http://www.massbank.jp/
MS-DIAL		x		http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/
mzCloud		x		https://www.mzcloud.org/
NIST		x		https://www.nist.gov/srd/nist-standard-reference-database-1a-v17
LipidBlast			x	http://fiehnlab.ucdavis.edu/projects/LipidBlast
MetFrag			x	http://c-ruttkies.github.io/MetFrag/
MS-FINDER			x	http://prime.psc.riken.jp/Metabolomics_Software/MS-FINDER/
HAMMER			x	http://www.biosciences-labs.bham.ac.uk/viant/hammer/
MS2LDA			x	http://ms2lda.org/
MetFamily	x	x	x	https://msbi.ipb-halle.de/MetFamily/
MetFusion			x	http://mgerlich.github.io/MetFusion/
MIDAS			x	https://github.com/chongle/midas-metabolomics
CFM-ID			x	http://cfmid.wishartlab.com/
FT-BLAST			x	https://bio.informatik.uni-jena.de/research/
MAGMa			x	https://github.com/NLeSC/MAGMa
CSI:FingerID			x	https://www.csi-fingerid.uni-jena.de/
MOLGEN-MS/MS			x	http://www.molgen.de/

Table 2. List of commonly detected adducts and in-source neutral losses observed in untargeted LC-MS datasets applying electrospray ionisation.

Adduct type	**In-source fragments**
$[M+H]^+$	$[M+H-H_2O]^+$ (water loss)
$[M+Na]^+$	$[M+H-NH_3]^+$ (ammonia loss)
$[M+K]^+$	$[M+H-CO]^+$ (carbon monoxide loss)
$[M+NH_4]^+$	$[M+H-CO_2]^+$ (carbon dioxide loss)
$[M-H]^-$	$[M+H-H_2S]^+$ (hydrogen sulfide loss)
	$[M+H-CH_2O_2]^+$ (formate loss)
	$[M+H-C_6H_8O_6]^+$ (glucuronide loss)
	$[M+H-H_3PO_4]^+$ (phosphate loss)
	$[M+H-H_2SO_4]^+$ (sulphate loss)

From mass to metabolite in human untargeted metabolomics: recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data

William J. Nash[1] and Warwick B. Dunn[1,2,3†]

[1] School of Biosciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[2] Phenome Centre Birmingham, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[3] Institute of Metabolism and Systems Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Corresponding Author (†)

Email: w.dunn@bham.ac.uk

Telephone: +44 (0)121 4145458

Highlights

- Untargeted metabolomics provides an unbiased study of human metabolomes
- These studies require the annotation of metabolites using data acquired, no prior list of metabolites is applied
- Reduction of the search space using full-scan data is applied first with filters incorporating isotopic information as an example
- Gas phase fragmentation can be applied to provide information on chemical structure used to differentiate between metabolites including isomers
- Development of new analytical and computational tools and resources has driven metabolite annotation forward significantly in the last ten years