HHS Public Access

Author manuscript

Nat Prod Rep. Author manuscript; available in PMC 2017 August 27.

Published in final edited form as:

Nat Prod Rep. 2016 August 27; 33(8): 942–950. doi:10.1039/c6np00024j.

Modern mass spectrometry for synthetic biology and structure-based discovery of natural products

Matthew T. Henke[a] and **Neil L. Kelleher**[a,b,*]

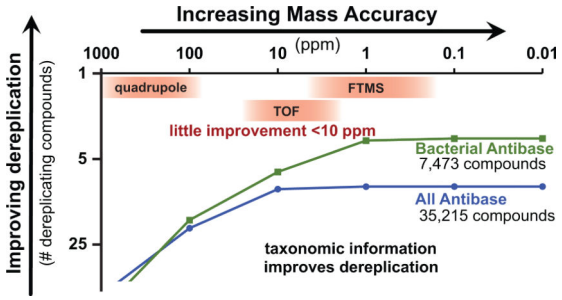
[a] Department of Molecular Biosciences, Northwestern University, Evanston IL 60208

[b] Department of Chemistry and the Feinberg School of Medicine Northwestern University, Evanston IL 60208

Abstract

In this highlight, we describe the current landscape for dereplication and discovery of natural products based on the measurement of the intact mass by LC-MS. Often it is assumed that because better mass accuracy (provided by higher resolution mass spectrometers) is necessary for absolute chemical formula determination (≤1 part-per-million), that it is also necessary for dereplication of natural products. However, the average ability to dereplicate tapers off at ~10 ppm, with modest improvement gained from better mass accuracy when querying focused databases of natural products. We also highlight some recent examples of how these platforms are applied to synthetic biology, and recent methods for dereplication and correlation of substructures using tandem MS data. We also offer this highlight to serve as a brief primer for those entering the field of mass spectrometry-based natural products discovery.

Graphical Abstract



This highlight serves as a primer for those curious about the abilities of mass spectrometry for natural products discovery and engineering.

1 Introduction

1.1 The importance of dereplication in discovery pipelines

Natural products from the microbial world are a valuable source of medically useful compounds.[1] An appreciation for the untapped potential of the microbial world has

* n-kelleher@northwestern.edu.

developed recently because of the increased availability of microbial genomes. Efficient discovery of these molecules is dependent on technologies capable of expediting the exploration and prioritization of this chemical space.

For decades, natural product chemists have relied on a workflow based on the bioactivity of natural products (Fig. 1, top). In this approach, extracts from microbial or plant sources are screened for activity (e.g., cytotoxicity against a panel of cancer cell lines or infectious microbes). If an extract displays activity, it is fractionated and the resulting fractions are then rescreened for that same activity. This process is repeated until the compound responsible for the activity is pure enough for NMR. Often at this stage, relatively pure compound is subjected to high-resolution mass spectrometry (HRMS) to obtain a chemical formula and NMR spectral data are acquired.

This pipeline has yielded a wealth of natural product scaffolds, which became or inspired several pharmaceutical success stories. After the golden age of natural product discovery (1960s through 1980s),[2] rediscovering the same natural products became a major challenge, and many pharmaceutical companies elected to forego natural products research in favour of screening synthetic libraries.[3]

However, incorporation of spectrometric data early in the discovery pipeline allows structural information to be used to identify, and filter out any known compounds that are present in the extract (i.e., dereplication). Unlike the traditional bioactivity-based pipeline, this means of natural products discovery is considered to be ‘structure-based’ (Fig. 1, bottom). While a structure-based approach may be more likely than a bioactivity-based approach to yield new chemical scaffolds (and therefore with a potential activity against new protein targets), it is also possible to identify compounds, for which no bioactivity can be found. Practically, the bioactivity-based approach will incorporate dereplication as early in the pipeline after finding a bioactive fraction; and, a structure-based approach will often incorporate bioactivity early on once finding extracts with many new compounds.

In addition to discovering natural products from the microbes that natively produce them, the structure-based approach has been used in synthetic biology efforts to link cryptic gene clusters to a natural product through genome mining efforts[4] and to prioritize strains for natural product isolation.[5] Genome mining starts with an orphan gene cluster of interest and seeks to identify associated natural products. This can be done by manipulation of the native strain[6] or through heterologous expression of entire gene clusters in either targeted[7] or high-throughput fashions.[8, 9] Engineering efforts that seek to make ‘unnatural products’ also rely on the structure-based approach to screen and characterize the products made by engineering pre-existing gene clusters.[10]

1.2 Methods for the detection of natural products

Instead of using bioactivity as the driving basis for a primary screen, the use of spectrometric or spectroscopic data to dereplicate known compounds has recently emerged as a viable option to screen for new natural products. The most commonly implemented means of detection for early stage dereplication are 1) ultraviolet-visible spectroscopy (UV/Vis)[11] provides information on chromophores present in a compound; 2) nuclear magnetic

resonance spectroscopy (NMR)[12] provides information on chemical environment and connectivity within a molecule through NMR-active isotopes (e.g., ^{1}H, ^{13}C, ^{15}N); and, 3) mass spectrometry (MS) provides the mass of compounds and, if the instrument is capable of fragmentation, the masses of subsequent fragment ions. Often these techniques can be placed in-line with each other (UV-MS, UV-NMR, etc.), thereby maximizing the complementarity of the information they provide.[13]

Of the detection methods, NMR and MS provide far more structural detail than UV/Vis. When compared to NMR, MS can be orders of magnitude more sensitive – often only nanograms of material (10^{-9} g) are needed for MS compared to the milligram requirements (10^{-3} g) for routine NMR; though, more recent NMR setups are capable of collecting data on just micrograms of material.[14] A list of benefits and drawbacks of these methods is summarized in Table 1. With these factors in mind, we view MS and tandem MS as a method efficient enough to serve as the front line of a structure-based pipeline for the high-throughput dereplication and discovery of new natural products.

While natural products discovery has benefited from the use of high-throughput mass spectrometry,[7, 15-19] many still rely on UV/Vis for detection.[20, 21] This stands in contrast with the primary metabolomics community, where robust mass spectrometry workflows have been developed to routinely identify and quantify hundreds of metabolites in central metabolism.[22-24] Our goal here is not to review the fundamentals of mass spectrometry. For that we refer to other reviews that highlight the more recent developments of various mass spectrometers for natural products discovery pipelines.[24-28] This highlight will focus on the ability of intact mass measurements (MS^1) and tandem MS (MS^2, MS/MS or MS2) data for dereplicating previously known compounds.

2 Before mass spectrometry

Before a complex mixture can be introduced into a mass spectrometer, there are several preparation steps that should be considered: source of biological material, method for extraction, chromatographic separation, and ionization of compounds. Each of these choices introduces biases that ultimately determine which compounds will be observed and dereplicated.

2.1 Biological sources – natural and synthetic

The first source of bias comes from the biological material itself: a microbial isolate that natively produces the natural product, or a heterologous system where natural product biosynthetic gene clusters have been introduced or even engineered. For native strains, culturing the microbe on defined media may greatly reduce background signals and significantly increase signals from target compounds, yet, without proper environmental cues, the organism may not produce natural products in detectable levels. To circumvent the often cryptic nature of natural products, the one-strain, many-compounds approach (OSMAC) has been articulated and implemented with success.[29-31] With OSMAC, a single strain is cultured across many different conditions with the hope that some combination of conditions will stimulate natural product biosynthesis.

Apart from culture conditions, more refined and targeted manipulations can be done to native strains to coax silent clusters into expression. HDAC inhibition has been used in various fungal species to keep repressive heterochromatin from forming and allowing for transcription of the biosynthetic gene clusters found in these regions of the genome.[16, 20, 32] Others have driven expression of cryptic gene clusters by upregulating the transcription factors found in biosynthetic gene clusters through targeted promoter replacement. When done for all non-reducing-polyketide synthase (NR-PKS) genes in *Aspergillus nidulans*, Ahuja et al. unveiled seven new compounds.[9]

Leaving the native host, which may not even be amenable to culturing, heterologous expression of entire natural product gene clusters provides conclusive association of a natural product with its gene cluster when successful. Recently, Petersen et al. combined ^{13}C-labeling with metabolomics and recombinantly expressed a single NR-PKS gene from *Aspergillus aculeatus* in *A. nidulans* to find 6-methylsalicylic acid and aculin A and B.[33]

Recombinant expression of an entire genome's worth of biosynthetic gene clusters has recently been achieved. Through the use of fungal artificial chromosomes, Bok et al. recombinantly expressed each of the 56 biosynthetic gene clusters from *Aspergillus terreus* in *A. nidulans* and identified the gene cluster for astechrome.[8] This technology has great potential when characterizing and exploring the biosynthetic gene clusters from unculturable or pathogenic fungi. Another platform that has disruptive potential was recently shown by Kang et al., where they developed a general CRISPR/Cas9 tool in yeast that combines high-throughput heterologous expression of microbial biosynthetic gene clusters with promoter replacement.[34]

The tools and techniques of structure-based approaches in natural products discovery are easily transferred to engineering projects. For instance, by relying on a structure-based approach, Newman et al. were able to rearrange and reprogram NR-PKS systems to provide deeper understanding behind the mechanisms of domains interaction that ultimately yield natural products.[35] Similarly, swapping domains between two fungal NR-PKS systems provided insights for chain length control by ketosynthase domain.[36] Such detailed understanding of the control points and mechanisms of these systems is necessary if the eventual goal of ‘designer unnatural products’ is to be achieved.

Beyond verifying the success of synthetic biology approaches in generating targeted compounds, structure-based approaches have been applied to help engineer ‘green’ alternatives to for the production of industrially used materials. Hagen et al. modified a PKS construct from borrelidin biosynthesis to make adipic acid.[10] And Asai et al. expressed a partial NR-PKS gene from *Chaetomium indicum* in *Aspergillus oryzae* to obtain a reactive intermediate that was then used as a starting point for diverse synthesis, producing a variety of ‘pseudonatural’ products.[21]

Once biological material is in hand, it is often extracted to remove background metabolites that interfere with detection of the natural products. The most common first pass extraction is with an inexpensive, non-miscible organic solvent, often ethyl acetate, which removes

most highly polar compounds, such as sugars and salts – and unfortunately, also highly charged natural products. The compounds removed by organic extraction are also the most frequent culprits that dirty and cause poor performance with MS instrumentation. Therefore, organic extraction can greatly reduce sample background and increase overall compatibility with downstream instrumentation, at the risk of removing some natural products. Recently a microscale platform has been developed that allows for the growth and extraction of microbes with the potential for truly high-throughput screening of culture conditions, microbial species and even co-cultures.[37]

2.2 Reducing sample complexity with chromatography

With samples cleaned, they are typically separated by chromatography as they are introduced to the instrument (the 'online' modality). Separation aids in detection of low abundance metabolites that may otherwise be suppressed by more abundant compounds. Several separation methods have been coupled with MS, but gas chromatography (GC) and reversed-phase liquid chromatography (RPLC) are by far the most used for natural products discovery.

GC is unparalleled in the separation of nonpolar, volatile compounds such as waxes, terpenoids, and fatty acids (that have been converted to their methyl ester derivatives), and has been used for dereplication of plant active metabolites.[38] GC is most often paired with an electron ionization source (EI) by passing compounds eluting from the GC column through a stream of electrons to generate molecular ions ($M^{+\cdot}$). EI imparts enough energy such that compounds are also fragmented (broken apart) upon ionization and the resulting mass spectrum is often unique for each molecule. Such spectra are deemed pseudo-MS^2 because the intact ion is not first isolated by the instrument and then fragmented, as is the case with MS capable of first isolating a specific ion (in essence purifying it) and then fragmenting it. Because GC-EI-MS spectra for hundreds of thousands of organic compounds are highly reproducible and massive repositories of ~10^6 of these spectra exist, natural products amendable to GC can often be compared to these databases for unambiguous dereplication. However, GC has limitations both in size (<500 Da) and polarity (hydrophobic compounds) of what can be analysed.

On the other hand, RPLC is the most utilized method for separation of natural products and can be coupled with a variety of on- and off-line detection method (see Table 1).[39-41] RPLC is capable of separating compounds over a broad range of masses and hydrophobicity. However, very hydrophobic compounds tend to adhere too strongly, and highly hydrophilic compounds often pass directly through the reversed-phase column without being retained (these may be better suited to hydrophilic interaction liquid chromatography (HILIC), which is not discussed here).

2.3 Ionization of molecules

Once separated by RPLC, solution phase molecules must be converted to gas phase ions by an ion source in order to be analysed by the mass spectrometer. There are several types of ionization sources,[42] but the most widely used are the atmospheric pressure ionization (API) methods. Compared to EI, API methods are significantly gentler and result in little to no

fragmentation. Among API there are several means to generate ions: atmospheric pressure chemical ionization (APCI) and atmospheric pressure photoionization (APPI) nebulize the solvent and compounds, and then transfer a proton from the mobile phase to the compounds in solution by either electrical or photon activation.[42] In addition, electrospray ionization (ESI), perhaps the most widely use ionization source, directly nebulizes molecules from solution conditions where they are already charged (e.g., acidic conditions to protonate basic compounds).

All API methods generate the protonated molecule $(M+H)^+$ in positive mode or the deprotonated molecule $(M-H)^-$ in negative mode. Highly acidic molecules are less likely to be observed in positive mode, while highly basic compounds are less likely to be observed in negative mode; therefore, no single MS polarity can cover all of natural product space. Some experimental set ups allow for the mass spectrometer to switch between modes (positive and negative) either in sequential runs or alternating each scan during a LC-MS run. However, if instrument time is a limiting factor, then a choice must be made. To answer this, Nielsen et al. tested a collection of 719 natural product standards and found that >90% were detectable in positive ESI compared to ~80% detectable in negative ESI.[43]

3 Types of mass spectrometers

Resolving Power and Mass Accuracy

For those just entering the field of mass spectrometry, the variety and number of abbreviations and hyphenations can be overwhelming. Simplistically there are three major types of instruments: 1) quadruple-based instruments (quad or Q); 2) time-of-flight instruments (TOF); and, 3) Fourier-transform instruments (FT). These instruments and some key information about them are shown in Table 2. The most prevalent distinction between these types of instruments is their ability to resolve molecules by their mass-to-charge (*m/z*) ratio – also known as resolving power - and to accurately measure mass values.

Often chemists state that a molecular formula was determined for an isolated natural product from a high-resolution MS (HRMS) measurement of the intact mass. However, this is not strictly true, as accurate mass is not a direct result of having high-resolving power instruments. Resolution refers to the width of the *m/z* peaks, which is calculated as the width of the peak at 50% abundance – full-width, half max (FWHM). If an *m/z* peak is narrower, its absolute center can more reliably be determined than a much broader peak. Theoretically a broad peak could be centered at the correct *m/z* position, but the confidence in that measurement is clouded by the fact that it cannot be determined if a mixture of compounds actually constitutes the observed peak, and the observed *m/z* value is a result of the averaging effect of the overlapping peaks. Therefore, increased resolving power is a significant component in the ability to more accurately measure *m/z*. Even still, an instrument with high-resolving power must be calibrated with known compounds in order to provide high mass accuracy measurements. A general trend for mass spectrometers (and perhaps most importantly for scientists considering which instrument type to purchase) is that the higher the resolving power, the more the instrument will cost.

Explaining Mass Accuracy

As mass spectrometers are capable of making increasingly higher mass accuracy measurements, they reduce the deviation in observed mass measurements from theoretical mass values. For instance, measuring a molecule with a theoretical mass of 500 Da will likely result in a measurement of 500 Da ± 0.25 Da (0.05% error) on a quadrupole-based instrument, while on an FT-based instrument, this measurement would have an expected variation of ± 0.0005 Da (0.0001% error) from the theoretical mass. Because the percent error values are so low for these measurements, mass spectrometrists often represent mass error as parts-per-million accuracy (ppm, where ppm = [theoretical mass – observed mass]/ theoretical mass•10^6). For comparison, percent is just another way of saying parts-per-hundred.

Targeted Detection Modes

Quadruple-based mass spectrometers include single quadrupole instruments, triple quadrupole instruments (triple quads or QqQ), and ion traps. These have the lowest resolving power of all, but have high duty cycles (spectra collected/time), can achieve low limits of detection, and are relatively inexpensive. While they cannot obtain the level of resolving power that is necessary for high accuracy mass measurements, they have very attractive performance metrics when set in series as a triple quadrupole instrument (QqQ). Here, ‘triple quads’ can be used to perform the Single (or Multiple) Reaction Monitoring assays (SRM/MRM).

In an SRM (or MRM) mode, one can achieve exquisite specificity and reliability in compound detection by relying on the observation of unique fragment ions that result from particular intact masses (called transitions). These assays can detect and quantify dozens to hundreds of natural products in complex samples.[44] However, in order to develop a quantitative SRM/MRM assay, standards of each compound to be dereplicated should have validated and unique transitions. A dereplication pipeline that relies on SRM/MRM and the use of standards is not feasible unless standards for known subsets of the ~100,000 known natural products can be obtained and a unique transition for SRM (or set of transitions in MRM) can be established for each compound for which a targeted assay is desired.

High Resolution MS

There are two generations of MS that provide high enough mass accuracy data to determine molecular formula. TOFs have higher resolving power than quadrupole instruments, with values often quoted to be ~10,000 and can provide 1 to 5 ppm mass accuracy. But mass spectrometers that collect data in the frequency domain and then convert those signals to *m/z* scale by means of Fourier Transform (FT) provide the highest resolution data of all, often >100,000.[45] For decades, this resolution regime was dominated by Ion Cyclotron Resonance MS (FT-ICRs), which require expensive superconducting magnets and attentive maintenance. More recently, the Orbitrap-based instruments have provided a compromise between resolution and instrument robustness, and are often able to provide accurate mass ± 0.001 Da.

4 Surveying dereplication based on MS^1 alone

Accurately measuring masses of components in an extract is only half of dereplication; one must also have a database of known natural products and their theoretical masses. For a discussion of natural product databases and the distinctions among them, we recommend the following reviews.[46, 47] In particular, *Antibase* (~35,000) and *Dictionary of Natural Products* (~120,000) are dedicated to natural products, and METLIN (http://metlin.scripps.edu) and GNPS (http://gnps.ucsd.edu) are outstanding mass spectral databases.

With a database chosen, and a list of experimentally determined intact masses of the compounds in an extract, a simple comparison of lists is performed. If an intact mass measurement is within the mass accuracy tolerance (which depends on instrument used, see Table 2), then it dereplicates to that entry. However, it is possible for a single intact mass to dereplicate to many entries in the database, and this can happen for two reasons.

Firstly, the compound entries, by chance, are similar in mass despite having different chemical formulas, for example, okaramine G ($C_{32}H_{34}N_4O_3$, 522.2631 Da) and calbistrin E ($C_{31}H_{38}O_7$, 522.2618 Da) differ in mass by 0.0013 Da (2 ppm). This ambiguity can often be overcome by increasing the mass accuracy of the measurement. The difference here cannot be distinguished with a mass accuracy of 10 ppm, but can be at 1 ppm mass accuracy.

Another example of two compounds with similar masses is found in (E)-6-Chloro-10,11-dehydrocurvularin ($C_{16}H_{17}O_5Cl$, 324.0765) and asparenomycin C ($C_{14}H_{16}N_2O_5S$, 324.0780). These differ in mass by 0.0015 Da, and could be distinguished by higher mass accuracy measurements, or simply by interrogation of isotopic distribution patterns even with lower mass accuracy: the chlorine imparts a unique +2 Da isotope peak at approximately 33% abundance of the monoisotopic peak.

The second reason for ambiguous dereplication is the result of structural isomers in the database, that is, multiple entries that have the same chemical formula. Resolution of structural isomers by measuring the intact mass of a compound is impossible regardless of the mass accuracy of the instrument.[48] An additional source of data (e.g., NMR or MS^2) must be used to fully dereplicate these cases (see below).

While ambiguity does exist in dereplication of natural products based on intact mass alone, we would like to briefly contextualize the magnitude of this issue, particularly as it relates to the mass accuracy of the instrument being used. To show this, we analysed a natural products database (*Antibase* 2011)[49] as a proxy for the size distribution of natural products that one is likely to encounter when dereplicating compounds encountered in a natural products pipeline.

For each metabolite in the database, a mass window centered at each accurate mass entry was calculated and the number of metabolites within the database that fell within those mass windows were counted. By narrowing the mass window – effectively measuring the metabolites with increasing mass accuracy (1000, 100, 10, 1, 0.1, 0.01 ppm) – ambiguity in dereplication based on intact mass decreases with increasing mass accuracy as expected (Fig. 2).

By increasing mass accuracy from 1000 ppm to 100 ppm, on average dereplication ambiguity decreases by 71.6% (65.3 entries/observed mass at 1000 ppm to 18.5 entries/observed mass at 100 ppm). These mass accuracy values straddle the typical value expected of a quadrupole-based instrument. An additional 52.1% decrease in dereplication ambiguity was observed when increasing mass accuracy from 100 ppm to 10 ppm (18.5 entries/observed mass to 8.9 entries/observed mass), the latter representing the typical mass accuracy of TOF instruments. Surprisingly, there was only a 4.7% decrease in dereplication ambiguity on average when increasing mass accuracy from 10 ppm to 1 ppm (which is typical for FT-based instruments) (8.9 entries/observed mass to 8.5 entries/observed mass). From this analysis, as it relates to natural products dereplication, little benefit is gained when increasing mass accuracy from 10 ppm to 1 ppm. This observation is in fact experimentally supported,[50] and initially may appear counterintuitive, especially when considering the necessity for high mass accuracy (≤1 ppm) in determination of chemical formulae *de novo*.

Instead of considering averages across the database, the number of possible unique dereplications based on intact mass alone (that is how many known compounds can be identified by intact mass measurement alone) effectively plateaus at 10,099 at 1 ppm mass accuracy (Fig. 4). The theoretical maximum is 10,185 compounds capable of being dereplicated unambiguously based on their intact mass alone. This means that approximately 70% of database (>25,000 entries) will result in ambiguous dereplication based on intact mass alone, regardless of mass accuracy.

While 70% ambiguity is high, ~5,200 of the ambiguous entries are ambiguous between just 2 structural isomers, and another ~3,300 are ambiguous among 3 structural isomers. Collectively these bi- and tri-ambiguous entries represent nearly 25% of the database. This level of ambiguity can reasonably be manually interrogated in a result set, for instance by comparing an experimental MS^2 spectrum to real or theoretical MS^2 spectra for the ambiguous hits.

Dereplication at the level of intact measurement (not relying on MS^2 data) can be further improved with isotopic labelling. This can be done broadly with ^{13}C labelling allowing for carbon counting (measured as a mass shift of 1 Da for each carbon).[51] Or in a more targeted manner, by using heavy isotope analogues of potential amino acids thought to be incorporated in the natural product, as in the case of nidulanin A.[52] The use of isotopic labelling requires the ability to culture on media containing metabolic precursors with stable isotopes where the natural product expression is still observed. If instead, MS^2 data can be collected, more detailed information can be collected on many more compounds and their substructures present in an extract.

5 Enhancing dereplication with MS^2 data

In order to collect MS^2 data (also called MS2, MS/MS or tandem MS), the instrument must be capable of isolating ions of a compound, adding energy to fragment it (often by colliding it with an inert gas), and then measuring the mass of the fragment ions (Fig. 4). In general, the ability to perform MS^2 increases the cost of the instrument.

Based on the structure of the compound, it will fragment generally at the weakest bonds. For instance, the peptide bond is particularly labile and often fragments, as do prenyl groups. The measurement of fragment ions produces a pattern called a fragmentation spectrum (or MS^2 spectrum, see panel at the far right of Fig. 4). These patterns, which are representative of substructures, are often unique to individual compounds, enough so that they can be used to overcome the limitations of dereplication without HRMS.[44] MS^2 data have also been used to automatically narrow down potential chemical formulae.[53-55] In fact, this method can be paired with matching isotope patterns for *de novo* chemical formula determination, which is particularly helpful when dealing with an unknown compound and beginning a full structure determination.[56,57]

However, for the purposes of dereplication, if a fragmentation spectrum of a standard of the natural product of interest can be obtained, this can be used to confidently confirm or reject identity of the experimental compound when it dereplicates to multiple database entries. But, if a fragmentation spectrum cannot be found in the literature, or cannot be generated, then the experimental fragmentation spectrum can be compared to the structures of the multiple database entries to which it dereplicates. At this point, the natural products chemist must either rely on intuition and experience to predict what fragment ions a structure is likely to generate, or use software (e.g., Mass Frontier) that can predict theoretical fragmentation spectra for comparison purposes. MS^2 data can be used to correlate sub-structures and therefore bypass the limitations of MS^1-only dereplication. The interpretation of MS^2 spectra when it comes to manual dereplication can be quite time-consuming. To overcome this, many efforts have been launched to automate the interpretation of fragmentation data,[58] and through assembly of fragmentation trees, such as by CSI:FingerID.[59]

MS^2 spectra can also be compared *en masse* in an automated fashion to build MS/MS networks to cluster related compounds.[60] When doing this, overlaid with dereplication based on intact mass, compounds can be dereplicated or deemed less interesting targets if they cluster with known compounds. The power of molecular networking has recently been extended further with the incorporation of *in silico* prediction of MS2 data for natural products databases and automated comparison with experimental data.[61] Going beyond simple dereplication (and microbial sources), Heiling et al. compared the MS^2 spectra of known 17-hydroxygeranyllinalool diterpene glycosides to unknown metabolites and were able to propose an additional 105 novel structural analogues from 35 species of plants.[62]

Instead of analysing data in a new way, Hoffmann et al. recently sought to have the mass spectrometer itself acquire data in a new way. Typically, mass spectrometers fragment the most abundant ions at a given time; however, by pre-screening microbial extracts and media blanks, a list of *m/z* values (metabolites) that are only seen in the microbial extracts was generated. The mass spectrometer then used this targeted list to seed MS^2 experiments, which not only focused the resulting data to include only metabolites from the microbial extract, it also allowed the instrument to analyse ions at far lower levels than more traditional acquisition can achieve, providing a deeper dive into the microbial metabolome.[63]

Indeed, analysis of metabolomics data in innovative ways has led to the discovery of bioactive metabolites. Kurita et al. developed a high-throughput means of integrating untargeted metabolomics and bioactivity data to generate networks of bioactive metabolites. Not only did this platform allow for the identification of new natural products, the quinocinnolinomycins, but based on comparison of bioactivities with known metabolites it allowed the researchers to determine that the new natural products likely target the endoplasmic recticulum.[64]

As stated earlier natural products chemists rely on multiple sources of structural data when dereplicating compounds in an extract.[65, 66] Particularly illustrative of this was work by El-Elimat et al., which demonstrated the creation of a database by collecting MS^1 and MS^2 in both positive and negative modes, and UV data of 172 fungal natural products. They then screened 106 fungal cultures for bioactivity. Positive extracts were then screened and dereplicated, which allowed the authors to eliminate 50% of the bioactive extracts because they were producing already known compounds.[67]

Even ignoring the incorporation of multiple datasets, the measurement of intact mass alone provides a great means to dereplication when combined with taxonomic information. For instance, when dereplicating an extract from an actinobacterial species, if the mass 324.0634 Da ($C_{18}H_{12}O_6$) is observed it is much more likely to be fluostatin C from Streptomyces (bacteria) than it is to be sterigmatocystin from *Aspergillus* species (fungi). Therefore, we recommend a dereplication pipeline that includes multiple sources of data, relying most strongly on intact accurate mass measurements, MS/MS networking, and taxonomic information (Fig. 5).

6 Conclusions

Dereplication of known natural products is a necessary step in the efficient discovery of novel compounds. Mass spectral data is the most powerful means to achieve robust and high-throughput dereplication. While some level of ambiguity exists if measurement of intact mass alone is the basis for dereplication, this problem can often be alleviated by incorporation of MS^2 data and taxonomic information. To further strengthen the ability for MS in natural products dereplication, we would also like to second the call for construction of an open-source natural products database that combines all known spectral data (MS, MS^2, UV/Vis and NMR) with taxonomic data and biosynthetic gene cluster data,[68] as such a resource would prove invaluable for the efficient harvest of microbial natural products.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank OS Skinner for many helpful discussions that guided the analysis of the natural products database, and that shaped both figures and text, and RA McClure and AW Goering for conversations that shaped figure 2. Northwestern University and the National Institutes of Health supported this work under grant numbers GM 067725 and AT 009143.

References

1. Newman DJ, Cragg GM. Journal of natural products. 2016; 79:629–661. [PubMed: 26852623]
2. Watve M, Tickoo R, Jog M, Bhole B. Archives of microbiology. 2001; 176:386–390. [PubMed: 11702082]
3. Baltz RH. Journal of industrial microbiology & biotechnology. 2006; 33:507–513. [PubMed: 16418869]
4. Bachmann BO, Van Lanen SG, Baltz RH. Journal of industrial microbiology & biotechnology. 2014; 41:175–184. [PubMed: 24342967]
5. Hou Y, Braun DR, Michel CR, Klassen JL, Adnani N, Wyche TP, Bugni TS. Analytical chemistry. 2012; 84:4277–4283. [PubMed: 22519562]
6. Lukassen MB, Saei W, Sondergaard TE, Tamminen A, Kumar A, Kempken F, Wiebe MG, Sorensen JL. Marine drugs. 2015; 13:4331–4343. [PubMed: 26184239]
7. Chiang YM, Oakley CE, Ahuja M, Entwistle R, Schultz A, Chang SL, Sung CT, Wang CC, Oakley BR. Journal of the American Chemical Society. 2013; 135:7720–7731. [PubMed: 23621425]
8. Bok JW, Ye R, Clevenger KD, Mead D, Wagner M, Krerowicz A, Albright JC, Goering AW, Thomas PM, Kelleher NL, Keller NP, Wu CC. BMC genomics. 2015; 16:343. [PubMed: 25925221]
9. Ahuja M, Chiang YM, Chang SL, Praseuth MB, Entwistle R, Sanchez JF, Lo HC, Yeh HH, Oakley BR, Wang CC. Journal of the American Chemical Society. 2012; 134:8212–8221. [PubMed: 22510154]
10. Hagen A, Poust S, Rond T, Fortman JL, Katz L, Petzold CJ, Keasling JD. ACS synthetic biology. 2016; 5:21–27. [PubMed: 26501439]
11. Hansen ME, Smedsgaard J, Larsen TO. Analytical chemistry. 2005; 77:6805–6817. [PubMed: 16255577]
12. Bradshaw J, Butina D, Dunn AJ, Green RH, Hajek M, Jones MM, Lindon JC, Sidebottom PJ. Journal of natural products. 2001; 64:1541–1544. [PubMed: 11754607]
13. Waridel P, Wolfender JL, Lachavanne JB, Hostettmann K. Phytochemistry. 2004; 65:2401–2410. [PubMed: 15381014]
14. Wang X, Duggan BM, Molinski TF. Magn Reson Chem. 2016 DOI: 10.1002/mrc.4415.
15. Petersen LM, Hoeck C, Frisvad JC, Gotfredsen CH, Larsen TO. Molecules. 2014; 19:10898–10921. [PubMed: 25068785]
16. Albright JC, Henke MT, Soukup AA, McClure RA, Thomson RJ, Keller NP, Kelleher NL. ACS chemical biology. 2015; 10:1535–1541. [PubMed: 25815712]
17. Doroghazi JR, Albright JC, Goering AW, Ju KS, Haines RR, Tchalukov KA, Labeda DP, Kelleher NL, Metcalf WW. Nature chemical biology. 2014; 10:963–968. [PubMed: 25262415]
18. Mohimani H, Liu WT, Kersten RD, Moore BS, Dorrestein PC, Pevzner PA. Journal of natural products. 2014 DOI: 10.1021/np500370c.
19. Sidebottom AM, Johnson AR, Karty JA, Trader DJ, Carlson EE. ACS chemical biology. 2013; 8:2009–2016. [PubMed: 23777274]
20. Asai T, Morita S, Taniguchi T, Monde K, Oshima Y. Organic & biomolecular chemistry. 2016; 14:646–651. [PubMed: 26549741]
21. Asai T, Tsukada K, Ise S, Shirata N, Hashimoto M, Fujii I, Gomi K, Nakagawara K, Kodama EN, Oshima Y. Nature chemistry. 2015; 7:737–743.
22. Benton HP, Ivanisevic J, Mahieu NG, Kurczy ME, Johnson CH, Franco L, Rinehart D, Valentine E, Gowda H, Ubhi BK, Tautenhahn R, Gieschen A, Fields MW, Patti GJ, Siuzdak G. Analytical chemistry. 2015; 87:884–891. [PubMed: 25496351]
23. Benton HP, Wong DM, Trauger SA, Siuzdak G. Analytical chemistry. 2008; 80:6382–6389. [PubMed: 18627180]
24. Gao P, Xu G. Analytical and bioanalytical chemistry. 2015; 407:669–680. [PubMed: 25216964]
25. Carter GT. Natural product reports. 2014; 31:711–717. [PubMed: 24468674]
26. Bouslimani A, Sanchez LM, Garg N, Dorrestein PC. Natural product reports. 2014; 31:718–729. [PubMed: 24801551]

27. Jarmusch AK, Cooks RG. Natural product reports. 2014; 31:730–738. [PubMed: 24700087]
28. Gaudencio SP, Pereira F. Natural product reports. 2015; 32:779–810. [PubMed: 25850681]
29. Bode HB, Bethe B, Höfs R, Zeeck A. Chembiochem : a European journal of chemical biology. 2002; 3:619–627. [PubMed: 12324995]
30. Hewage RT, Aree T, Mahidol C, Ruchirawat S, Kittakoop P. Phytochemistry. 2014; 108:87–94. [PubMed: 25310919]
31. Yuan C, Guo YH, Wang HY, Ma XJ, Jiang T, Zhao JL, Zou ZM, Ding G. Scientific reports. 2016; 6:19350. [PubMed: 26839041]
32. Mao XM, Xu W, Li D, Yin WB, Chooi YH, Li YQ, Tang Y, Hu Y. Angewandte Chemie. 2015; 54:7592–7596. [PubMed: 26013262]
33. Petersen LM, Holm DK, Gotfredsen CH, Mortensen UH, Larsen TO. Chembiochem : a European journal of chemical biology. 2015; 16:2200–2204. [PubMed: 26374386]
34. Kang HS, Charlop-Powers Z, Brady SF. ACS synthetic biology. 2016 DOI: 10.1021/acssynbio.6b00080.
35. Newman AG, Vagstad AL, Storm PA, Townsend CA. Journal of the American Chemical Society. 2014; 136:7348–7362. [PubMed: 24815013]
36. Liu T, Sanchez JF, Chiang YM, Oakley BR, Wang CC. Organic letters. 2014; 16:1676–1679. [PubMed: 24593241]
37. Barkal LJ, Theberge AB, Guo CJ, Spraker J, Rappert L, Berthier J, Brakke KA, Wang CC, Beebe DJ, Keller NP, Berthier E. Nature communications. 2016; 7:10610.
38. Gu J-Q, Wang Y, Franzblau SG, Montenegro G, Timmermann BN. Phytochemical Analysis. 2006; 17:102–106. [PubMed: 16634286]
39. Strege MA. Journal of Chromatography B. 1999; 725:67–78.
40. Potterat O, Wagner K, Haag H. Journal of Chromatography A. 2000; 872:85–90. [PubMed: 10749489]
41. Eugster PJ, Boccard J, Debrus B, Breant L, Wolfender JL, Martel S, Carrupt PA. Phytochemistry. 2014; 108:196–207. [PubMed: 25457501]
42. Bhardwaj C, Hanley L. Natural product reports. 2014; 31:756–767. [PubMed: 24473154]
43. Nielsen KF, Mansson M, Rank C, Frisvad JC, Larsen TO. Journal of natural products. 2011; 74:2338–2348. [PubMed: 22026385]
44. Nothias-Scaglia LF, Dumontet V, Neyts J, Roussi F, Costa J, Leyssen P, Litaudon M, Paolini J. Fitoterapia. 2015; 105:202–209. [PubMed: 26151856]
45. Marshall AG, Hendrickson CL. Annual review of analytical chemistry. 2008; 1:579–599.
46. Füllbeck M, Michalsky E, Dunkel M, Preissner R. Natural product reports. 2006; 23:347–356. [PubMed: 16741583]
47. Johnson SR, Lange BM. Frontiers in bioengineering and biotechnology. 2015; 3:22. [PubMed: 25789275]
48. Kind T, Fiehn O. BMC bioinformatics. 2006; 7:234. [PubMed: 16646969]
49. Laatsch H. Antibase. 2011; 2011
50. Glauser G, Veyrat N, Rochat B, Wolfender JL, Turlings TC. Journal of chromatography. A. 2013; 1292:151–159. [PubMed: 23274073]
51. Clendinen CS, Stupp GS, Ajredini R, Lee-McMullen B, Beecher C, Edison AS. Front Plant Sci. 2015; 6:611. [PubMed: 26379677]
52. Klitgaard A, Nielsen JB, Frandsen RJ, Andersen MR, Nielsen KF. Analytical chemistry. 2015 DOI: 10.1021/acs.analchem.5b01934.
53. McDonald LA, Barbieri LR, Carter GT, Kruppa G, Feng X, Lotvin JA, Siegel MM. Analytical chemistry. 2003; 75:2730–2739. [PubMed: 12948143]
54. Konishi Y, Kiyota T, Draghici C, Gao J-M, Yeboah F, Acoca S, Jarussophon S, Purisima E. Analytical chemistry. 2007; 79:1187–1197. [PubMed: 17263353]
55. Rojas-Chertó M, Kasper PT, Willighagen EL, Vreeken RJ, Hankemeier T, Reijmers TH. Bioinformatics. 2011; 27:2376–2383. [PubMed: 21757467]
56. Vizcaino MI, Crawford JM. Nature chemistry. 2015; 7:411–417.

57. Goering AW, McClure RA, Doroghazi JR, Albright JC, Haverland NA, Zhang Y, Ju KS, Thomson RJ, Metcalf WW, Kelleher NL. ACS Cent Sci. 2016; 2:99–108. [PubMed: 27163034]
58. Hufsky F, Scheubert K, Bocker S. Natural product reports. 2014; 31:807–817. [PubMed: 24752343]
59. Duhrkop K, Shen H, Meusel M, Rousu J, Bocker S. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112:12580–12585. [PubMed: 26392543]
60. Nguyen DD, Wu CH, Moree WJ, Lamsa A, Medema MH, Zhao X, Gavilan RG, Aparicio M, Atencio L, Jackson C, Ballesteros J, Sanchez J, Watrous JD, Phelan VV, van de Wiel C, Kersten RD, Mehnaz S, De Mot R, Shank EA, Charusanti P, Nagarajan H, Duggan BM, Moore BS, Bandeira N, Palsson BO, Pogliano K, Gutierrez M, Dorrestein PC. Proceedings of the National Academy of Sciences of the United States of America. 2013; 110:E2611–2620. [PubMed: 23798442]
61. Allard PM, Peresse T, Bisson J, Gindro K, Marcourt L, Pham VC, Roussi F, Litaudon M, Wolfender JL. Analytical chemistry. 2016; 88:3317–3323. [PubMed: 26882108]
62. Heiling S, Khanal S, Barsch A, Zurek G, Baldwin IT, Gaquerel E. Plant J. 2016; 85:561–577. [PubMed: 26749139]
63. Hoffmann T, Krug D, Huttel S, Muller R. Analytical chemistry. 2014; 86:10780–10788. [PubMed: 25280058]
64. Kurita KL, Glassey E, Linington RG. Proceedings of the National Academy of Sciences of the United States of America. 2015; 112:11999–12004. [PubMed: 26371303]
65. Nielsen K, Smedsgaard J. Journal of Chromatography A. 2003; 1002:111–136. [PubMed: 12885084]
66. Sica VP, Raja HA, El-Elimat T, Kertesz V, Van Berkel GJ, Pearce CJ, Oberlies NH. Journal of natural products. 2015 DOI: 10.1021/acs.jnatprod.5b00268.
67. El-Elimat T, Figueroa M, Ehrmann BM, Cech NB, Pearce CJ, Oberlies NH. Journal of natural products. 2013; 76:1709–1716. [PubMed: 23947912]
68. Nielsen KF, Larsen TO. Frontiers in microbiology. 2015; 6:71. [PubMed: 25741325]

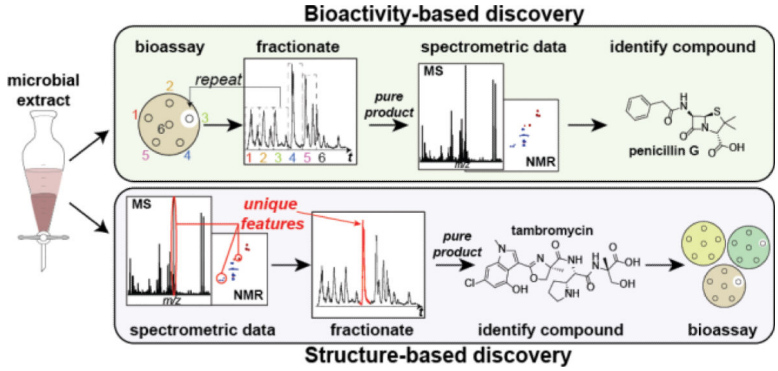


Fig. 1.

A comparison of two different workflows for natural products discovery. Structure-based discovery (bottom) emphasizes the use of spectrometric data early in the pipeline to quickly dereplicate known compounds and scaffolds, allowing researchers to save resources and focus on unknown compounds.

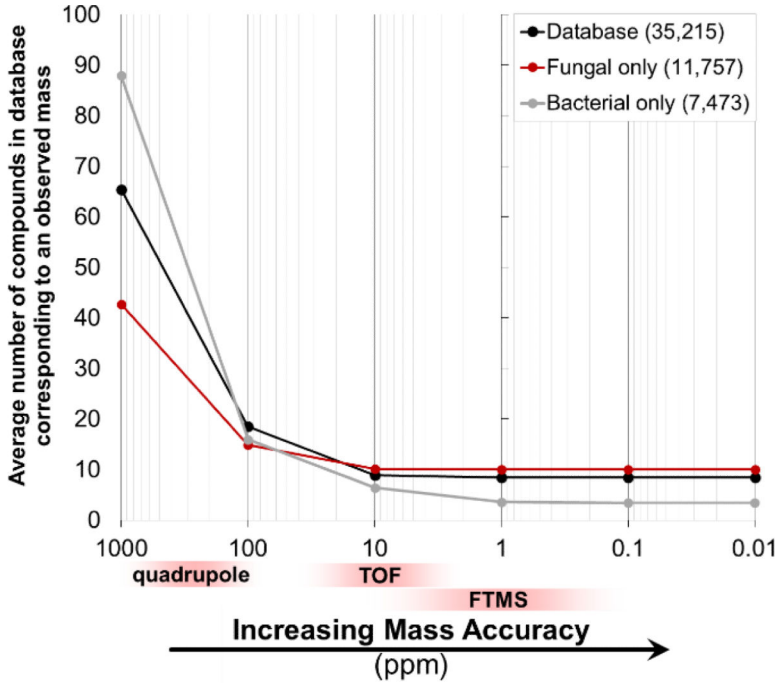


Fig. 2.

With a lower mass error (i.e., better mass accuracy), the observed mass of a compound will dereplicate to a fewer number of candidate compounds in a database of 35,215 natural products (Antibase was used here). On average, there is only a 4.7% gain in dereplication ability by increasing mass accuracy from 10 to 1 ppm.

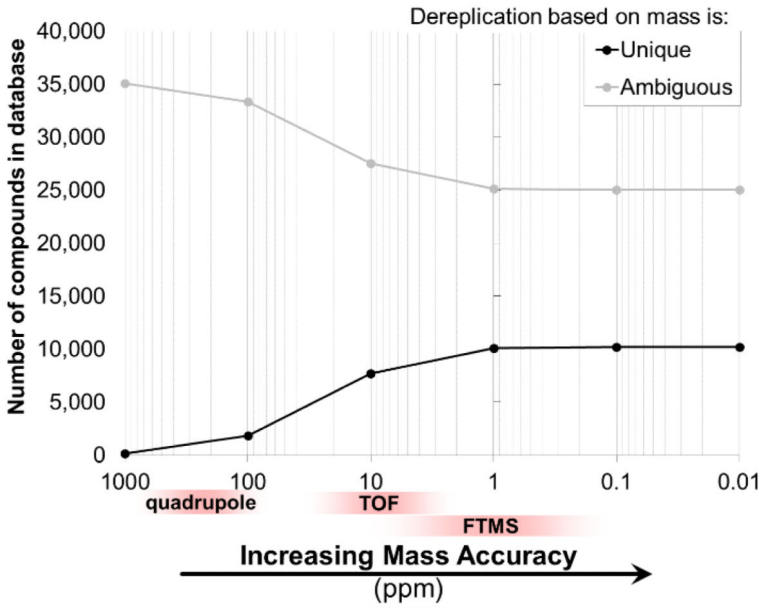


Fig. 3.

As mass accuracy increases, the number of compounds in the database that can be unambiguously dereplicated based on measurement of the intact mass increases drastically by going from 100 to 10 ppm, aga in to a lesser extent from 10 ppm to 1 ppm, and barely from 1 to 0.1 ppm.

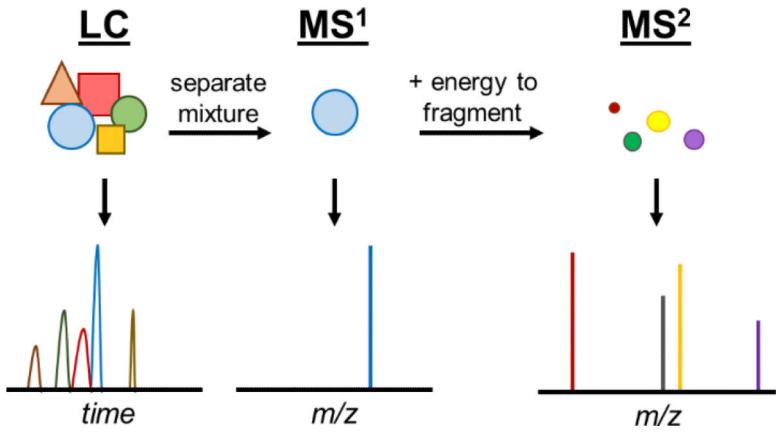


Fig. 4.

By selectively fragmenting a natural product, MS^2 can provide unique structural details that can assist dereplication.

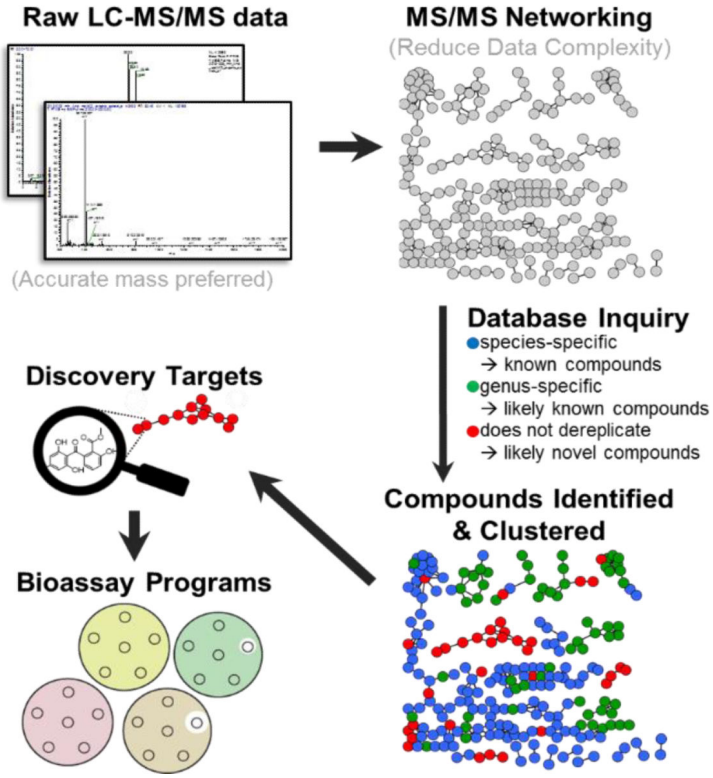


Fig 5.

A dereplication pipeline for natural products discovery. This pipeline takes advantage of the incredible amount of structural information that can be gleaned from LC-MS, especially when accurate mass (1-5 ppm) and MS^2 data are used together.

Table 1

Pros and cons for the detection methods most used for natural products dereplication

Detection method	Pros	Cons
UV/Vis	• Inexpensive • Non-destructive • Easily placed in-line with MS	• Provides little structural information • Compounds must be UV-active
MS	• Very sensitive (~ng) • Provides molecular formula (MS^1) • Can provide substructures (MS^n)	• Ionization is highly dependent on compound and conditions • Can be very expensive • No good spectral databases for MS^2 • Destructive
NMR	• Highest level of structural detail • Applicable to all compounds • Non-destructive	• Very expensive • Low sensitivity

Table 2

Types of mass spectrometers and various pieces of information that are important to know about them

Mass spectrometer	Resolving Power	Mass Accuracy	Typical Mass Error[b] (Da)	Cost
quadrupole/ion trap	1000	~500 ppm	0.25	$
TOF	10,000 (50,000)	1-5 ppm[a]	0.0005-0.0025	$$ ($$$)
FT-Orbitrap	100,000	~1 ppm	0.0005	$$$
FT-ICR	100,000-1,000,000	<1 ppm	<0.0005	$$$$

[a] these levels are attainable with an internal standard

[b] expected deviation in mass measurement of a 500 Da compound