Advanced Multidimensional Separations in Mass Spectrometry: Navigating the Big Data Deluge

Jody C. May and John A. McLean

Department of Chemistry, Center for Innovative Technology, Vanderbilt Institute for Chemical Biology, Vanderbilt Institute for Integrative Biosystems Research and Education, Vanderbilt University, Nashville, Tennessee 37235; email: john.a.mclean@vanderbilt.edu

Annu. Rev. Anal. Chem. 2016. 9:387–409

First published online as a Review in Advance on March 30, 2016

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

This article's doi: 10.1146/annurev-anchem-071015-041734

Keywords

systems biology, integrated omics sciences, structural elucidation, peak capacity, mass defect analysis, metabolites, metabolomics

Abstract

Hybrid analytical instrumentation constructed around mass spectrometry (MS) is becoming the preferred technique for addressing many grand challenges in science and medicine. From the omics sciences to drug discovery and synthetic biology, multidimensional separations based on MS provide the high peak capacity and high measurement throughput necessary to obtain large-scale measurements used to infer systems-level information. In this article, we describe multidimensional MS configurations as technologies that are big data drivers and review some new and emerging strategies for mining information from large-scale datasets. We discuss the information content that can be obtained from individual dimensions, as well as the unique information that can be derived by comparing different levels of data. Finally, we summarize some emerging data visualization strategies that seek to make highly dimensional datasets both accessible and comprehensible.

INTRODUCTION

All grand challenges in which mass spectrometry (MS) plays a role are characterized by the big data paradigm (**Table 1**). Proteomics seeks to detect and measure all proteins found in an organism (1), which based on several recent drafts numbers between 16,000 and 19,000 for basic human proteins (2–4) but could be as high as several million once protein variants and modifications are taken into account (5). The inclusion of spatially resolved protein information from imaging studies will increase this number even further (6). The human metabolome is represented by many diverse classes of small-molecule metabolites, of which over 40,000 have been annotated with support from MS techniques (7–9), but estimates place the possible number of human metabolites as high as 180,000 for lipids alone (10). Another grand challenge, systems biology, seeks to form connections between all the various classes of biomolecules in both space and time toward the comprehensive diagnosis of disease states (11), and MS is at the forefront of integrated omics approaches that will help realize this vision (12–14). Drug discovery initiatives aim to find the proverbial needles in haystacks in a molecular landscape of over 10^{60} possible chemical structures (15, 16), which is a haystack containing a novemdecillion straws of hay, or about 40 orders of magnitude greater than the number of grains of sand on Earth. To address this formidable challenge, researchers turn to high-throughput screening using MS-based assays that are capable of screening up to 100,000 compounds a day from combinatorial small-molecule libraries (17, 18). Genomics, which is driven by massively paralleled DNA sequencers (19), is currently feeling the burden of big data, with 2 to 40 million terabytes of genomic data projected to be generated in the next 10 years, representing 100 million to 2 billion complete human genomes sequenced. According to a recent report, this volume of genomics data will surpass that of YouTube, Twitter, and the future Square Kilometre Array by 2025 (20).

The above examples underscore only one aspect of big data: big numbers. But big data challenges are much more complex than dealing with large-scale datasets. The three Vs are often invoked to identify a big data challenge: (*a*) a large *volume* of data, (*b*) being generated at high *velocity*, (*c*) and characterized by a *variety* of different subsets of data (21). Lusher et al. (22, p. 861) define a big data challenge more broadly in terms of "whether the [researchers are] able to extract the relevant information from their rapidly growing data resources". Thus, the definition of big data is relative to the field in which the data are generated. A relevant example is the massive data generated in the field of astronomy, datasets on the order of petabytes (1,000 terabytes), which

Table 1 Grand challenges addressed by MS-based research

Grand challenge	Description	Scope of data volume
Systems biology	Map the interconnectivity of all biomolecules in space and time	>100,000 discrete biomolecules, over 10^9 possible binary connections
Omics sciences	Genomics	>20,000 human genes
	Proteomics	Approximately 20,000 base human proteins; 10^6 possible protein variants
	Metabolomics Lipidomics Glycomics	>40,000 annotated human metabolites; >200,000 possible metabolites
Drug discovery	Find chemicals with desirable pharmacological properties	$>10^{60}$ possible drug targets; 10^{11} virtual drug-like chemical structures mapped
Synthetic biology	Engineer and chemically characterize surrogate biosystems for translational research	Elements of all of the above plus xenometabolites and temporal sampling on the order of seconds

are efficiently handled due to decoupling data storage with data analysis, bridged through cloud computing (23). Another example is CERN's Large Hadron Collider data analysis network, which is built upon a massively distributed system architecture (24). MS as a field is beginning to embrace the concept of a decentralized computing infrastructure, but it is not there yet.

In this review, we trace the challenges of big data in analytical chemistry from its origins in the enumeration of chemical isomers to the (arguably) current locus in the field of multidimensional analysis based on MS. The need for multidimensional analysis is rationalized based on the problem of enumerating small molecules (25), and beyond high dimensionality and enhanced peak capacities, the correlation between discrete dimensions of data is presented to illustrate the rich information content afforded by scaling to higher dimensions. Finally, we provide an overview of some recent and creative visualization strategies that provide comprehensible access to higher-order dimensional information in a low-dimensional format.

Cloud computing: a resource-sharing concept of deploying computationally intensive work across a distributed network of computers

THE FOUNDATION OF BIG DATA IN CHEMICAL RESEARCH

Generating Chemical Knowledge

In many of his works, the futurist Buckminster Fuller argued that human knowledge was increasing at an unprecedented rate in human history, and this provided unique opportunities in the areas of science, engineering, and design. Fuller (26) supported his argument by plotting a timeline of the discovery of the chemical elements (**Figure 1*a***) and noted that knowledge was essentially increasing exponentially over time. In light of new data, the discovery of chemical elements has occurred at a relatively fixed rate for the past 50 years, reflecting the difficulty in creating stable nuclei of the superheavy elements. Knowledge stems from myriad sources and is more aptly illustrated from discoveries made in the absence of limitations. An updated observation of knowledge doubling in the chemical sciences can be seen in **Figure 1*b***, which plots over time the indexing of new chemical substances in the Chemical Abstracts Service (CAS) Registry. As chemical space is vast, the number of unique chemical registry numbers is increasing near exponentially since the CAS Registry system was first introduced in 1965, reaching 100 million substances at the year of this publication (27), and 200 million substances are expected to be indexed within the next five years. The open-access repository, PubChem (28), is growing even faster, with over 150 million chemical substances indexed since its introduction in 2004 (29). Over 60 million of these have been validated as unique chemical compounds (30). These and other efforts (31–34) to catalog all chemical compounds discovered represent our known chemical universe (35), which according to one estimate represents less than 1 in 10^{50} possible molecular structures for small organic compounds alone (36). To put this number in context, 1 in 10^{50} is a greater disparity of scale than the height of a person compared to the diameter of the observable universe, or 1 in 10^{26} m (37).

Although enumeration of all possible molecular structures is impossible, recent progress toward enumerating compounds of 17 atoms or less containing C, H, N, O, S, and halogens has resulted in 166.4 billion virtual drug-like compounds of approximately 350 Da or less (38, 39). Such new compound discovery and annotation efforts are big data challenges that represent the chemical sciences in the purest sense, and form the basis for the nascent fields of chemography (40) and cheminformatics (41, 42). **Figure 1*c*** contains a histogram of the number of chemical abstracts indexed by the CAplus system, which is accessed through SciFinder. As of January 2016, over 45 million chemical abstracts have been indexed (43), which based on current trends is expected to double within the next decade. This number of publications is on scale with the estimated 50 million total number of peer-reviewed articles across all disciplines as of 2009 (44), of which a little over half (approximately 30 million) were represented in CAplus at that time. Extrapolating this observation

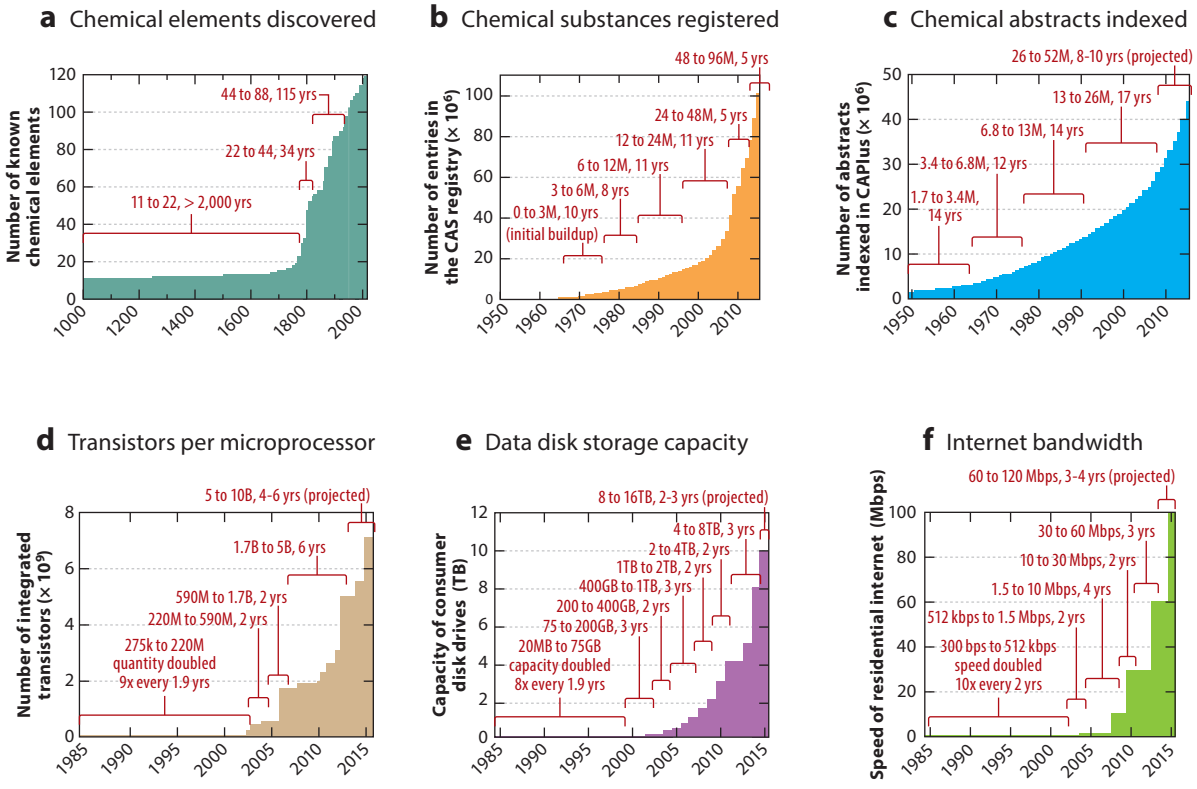

Figure 1

Histograms illustrating the increasing rate of knowledge and innovation relevant to the analytical sciences. These histograms describe growth numbers for (*a*) chemical elements discovered; (*b*) chemical substances registered in the Chemical Abstracts Service (CAS) Registry system; (*c*) chemical abstracts indexed in CAplus; (*d*) number of transistors in each generation of microprocessor, also known as Moore's law; (*e*) data storage capacity for each generation of consumer disk drive; and (*f*) accessible internet bandwidth for residential customers.

to the present suggests that at least 75 million journal articles are currently in existence. In a broader sense, this trend illustrates dissemination of greater quantities of chemical data into chemical information, which follows an exponential growth rate. If the canonical goal of the analytical sciences is to separate, identify, and quantify chemical substances from a variety of sources, then there is an incredibly vast amount of chemical space left to explore.

Translating Chemical Information

Moore's law: coined by Intel founder Gordon Moore, who initially observed that microprocessor component density doubled approximately every two years

The ability to acquire and analyze large amounts of data is driven by advances in the computer sciences. Panels *d* through *f* of **Figure 1** illustrate the so-called digital laws, which describe (*a*) exponential scaling of computer processor speeds (Moore's law; **Figure 1*d***) (45), (*b*) increasing capacity of consumer-grade data storage (**Figure 1*e***) (46), and (*c*) available bandwidth of residential broadband internet (**Figure 1*f***) (47). Collectively, these digital laws represent innovation and, on a broader level, humanity's current capacity to process, store, and disseminate information. From the doubling brackets annotated on each graph, it is apparent that computer processor speed (related to the number of transistors) now doubles approximately every four to six years,

whereas data storage capacity is doubling every two to three years. These observations infer two key points for large-scale analytical data mining: (*a*) Data volume will increase faster than the informatics tools necessary to interpret the data, and (*b*) broader data accessibility will provide new opportunities to handle large analytical datasets. The first is already being realized in many areas of MS-based research, and the second point is beginning to emerge from two notable concepts: crowdsourcing and cloud computing.

Crowdsourcing: a division of labor concept whereby many individuals are tasked at solving a complicated problem

LC-MS: liquid chromatography coupled to mass spectrometry

Crowdsourcing Mass Spectrometry Data

Crowdsourcing describes a division of labor concept in which the combined efforts of a large group of individuals are applied toward solving a complex problem. Typically, the contributing individuals are not experts in the problem they are tasked to solve, but rather citizen scientists who utilize their natural human capabilities to address scientific challenges. Crowdsourcing in science as a concept is not new. In 1714, the British government issued the Longitude Prize to anyone who could determine, by relatively simple means, the longitude of a ship at sea, which fostered advances in cartography and celestial navigation while creating the new science of marine chronometry (48). Recently, the problem of three-dimensional protein folding has been addressed by means of a novel online video game, Foldit, whereby players attempt to fold proteins, many of which have unsolved crystal structures. Foldit is an interactive puzzle game that is built upon Rosetta folding algorithms (49, 50), allowing online players to move subdomains, as well as shake and wiggle a protein structure to minimize its energy (51). Foldit players have so far been able to solve the three-dimensional structure of a retroviral protease (52, 53), as well as improve the biological activity of a computationally designed enzyme (54). In MS, there have been a few, although sparse, notable efforts to crowdsource large-scale efforts. Several MS database initiatives source from user-submitted data, including MassBank, HMDB (Human Metabolome Database), and mzCloud. Bradley et al. (55) described a gameplay approach similar to Foldit whereby players of the web-based Spectral Game are asked to match mass spectra to their corresponding molecular structures. Utilizing citizen scientists, Du et al. (56) described a study whereby property owners collected soil samples that were screened by liquid chromatography–mass spectrometry (LC-MS) for potential natural products. These efforts yielded a novel fungal metabolite, maximiscin, which exhibited antitumor activity in a mouse model (56).

An open call for more crowdsourced data analysis resources in MS-based proteomics has recently been made (57), which, based on other similar efforts in structural and network biology (58, 59), will foster more innovation and standardization across the field. Pragmatically speaking, it makes sense to query large groups of individuals for data analysis, as the capacity of the human brain surpasses that of conceptualized mathematical algorithms for discerning complex patterns across datasets, e.g., images (60). We envision the above strategies developed for structural biology and the nascent efforts now under way to crowdsource data analysis in MS will be critical to inferring important patterns or information from massive datasets, including those from discrete fields (e.g., transcriptomics, proteomics), but ultimately to linking datasets spanning the breadth of systems biology. Whereas the promise of data-driven discovery has been classically framed in the context of autonomous algorithms combing through large amounts of data, the human element in such efforts should not be understated (61).

Cloud Computing in Mass Spectrometry Research

Cloud computing, or specifically the concept of conducting data-intensive computational work across a distributed network of computers (62), is rapidly being adopted in many MS-based workflows (63). One example is XCMS Online, which offers cloud-based processing of LC-MS

metabolomics datasets, from feature extraction to normalization of data dimensions and statistical analysis of the results, and is capable of handling terabytes of data (64, 65). Another example, OpenMSI, outsources the US Department of Energy's National Energy Research Scientific Computing Center (NERSC) supercomputing facility toward processing highly dimensional imaging MS datasets, which in a single experiment can exceed 50 GB in size (66, 67). One of the earliest big data initiatives in MS, proteomics, has seen several recent offerings of open-access software tools for processing tandem MS (MS/MS) data (23, 68–70), and notable among these is the Trans-Proteomic Pipeline (71), which supports outsourcing of the software to the Amazon Web Services cloud computing infrastructure (72). These and other efforts to decentralize the computational resources required to handle MS-based data will help facilitate the development of educational and research programs where large investments in computer infrastructure are no longer necessary, and will help offset the need for institutions to invest heavily in instrumentation altogether. Open-source repositories of MS data help make this possible (73, 74). In the near future, it is conceivable that entire research programs will conduct MS research on digitally streamed data without direct access to a mass spectrometer.

MS/MS: tandem mass spectrometry

Ion mobility coupled to mass spectrometry (IM-MS): in a manner that utilizes the configuration as a distinct analytical technique

MULTIDIMENSIONAL METHODS BASED ON MASS SPECTROMETRY

The Genesis of Big Data in Mass Spectrometry

MS is well suited to address big data challenges, as the throughput and information density of the technique are both extraordinarily high. **Figure 2** highlights the peak capacity and the peak production rate of some MS-based analytical techniques that are used, and have the potential to be used, in big data initiatives. Peak capacities are commonly reported for condensed-phase separations such as liquid and gas chromatography (75, 76), but are less common in ion mobility (IM) and MS research, and so specific considerations are taken to generate these metrics in this present work. For example, MS calculations for peak capacity are based on methods developed by Frahm et al. (77) that account for instrument resolving powers and isotope redundancy and utilize typical rather than optimal parameters for each method (78). For IM, peak capacities are obtained from measured values of different techniques where available (79–81) or otherwise calculated from reported resolving powers (82, 83). For IM coupled to MS (IM-MS), there is a correlation between size and mass, and so IM-MS data in **Figure 2** are scaled by a factor of 0.25 (25% unique space occupancy) to reflect this reduced orthogonality, which is based on experiments conducted in the authors' laboratory. The power and potential of multidimensional analytical separations become evident when comparing both the peak capacities and peak production rates of individual analytical separation dimensions with those of multidimensional techniques. A cursory look at IM separations at the top of the scale reveals that the peak capacities are quite low, less than 100 (84), but IM can produce data at rates exceeding 100 peaks per second (85).

The next data-dense separation technique is liquid chromatography, for which peak capacities exceed 100 for one-dimensional LC and several thousand if two-dimensional LC is utilized (86, 87). The latter is on scale with two-dimensional electrophoresis (88, 89), as well as traditional MS techniques such as triple quadrupoles and ion trap instrumentation. High-resolution MS techniques exhibit peak capacities approaching 100,000 or greater and are capable of very high peak production rates ranging from 100,000 peaks per second for Orbitrap MS [Fourier transform MS (FTMS)] to over 100 million peaks per second for time-of-flight (TOF) MS. Fourier transform ion cyclotron resonance (FTICR) is capable of peak capacities approaching 1 million. Peak capacities beyond approximately 1 million are available through MS/MS experiments and/or by coupling multiple dimensions of separation to MS. For example, IM coupled to TOF MS (IM–TOFMS)

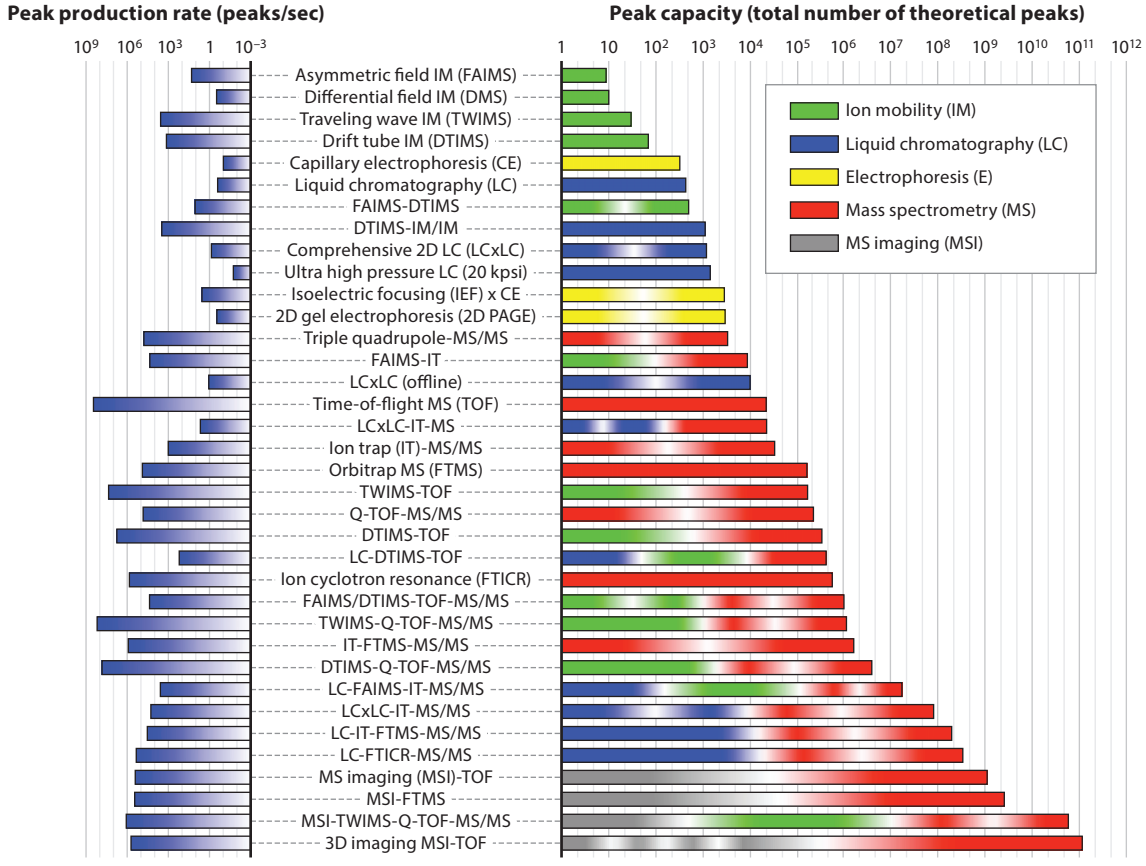

Figure 2

Peak capacities (*right*) and peak capacity production rates (*left*) for hybrid multidimensional mass spectrometry and related techniques. Specific techniques and combinations are selected based on available information in the literature.

offers a similar peak capacity as FTICR but can generate approximately one order of magnitude more peaks, at 10 million peaks per second (85). The addition of LC to MS and IM-MS combined with MS/MS achieves between 1 million and 100 million peaks depending on the specific configuration, with production rates of 1,000 spectra or more per second (79, 90).

Mass spectrometry imaging (MSI) adds spatial information to the analysis, increasing peak capacity to over 1 billion peaks, with a production rate of approximately 100,000 peaks per second (91). A combination of MSI experiments and IM-MS and MS/MS analyses generates over 10 billion peaks at a rate of approximately 1 million peaks per second (92). Finally, three-dimensional MSI, which obtains layered spatial information, is capable of very high data density, here illustrated by a recent example of over half a million MSI pixels coupled to TOF, which generates over 100 billion peaks at a rate of more than 100,000 peaks per second (93). If IM experiments and/or MS/MS were included, then a theoretical peak capacity in excess of 1 trillion would be possible with three-dimensional MSI. This small sampling of possible analytical configurations does not take into consideration the inclusion of experimental time-points or cohorts, which are a component of comprehensive biological studies. Conservatively, multidimensional MS data generation is on

MSI: mass spectrometry imaging

the order of thousands of peaks per second, with capabilities for resolving tens of thousands of molecular signals in a single experimental sequence.

The Importance of Multidimensional Mass Spectrometry

The fundamental analytical approach to system complexity is to reduce a complex problem into manageable subsets of data. In analytical chemistry, this involves elucidating the chemical structure of an analyte from an initially unknown sample or, conversely, characterizing a complex sample by descriptors of its molecular constituents. For MS, a conventional approach to molecular identification is to use the mass measurement to reduce several million possible molecular formulas to a few thousand or less toward initial characterization (94). Other orthogonal pieces of information are then used in a complementary fashion to assign a structural identity to the unknown analyte. This challenge is often understated, and so it is instructive to revisit the challenge of structural elucidation by MS using a relevant example.

Figure 3 illustrates an example of molecular identification using an MS-centric approach. For molecules less than 2,000 Da, there are over 8 billion possible molecular formulas that can be assigned to an unknown analyte (95). Typically, an initial sample fractionation step such as

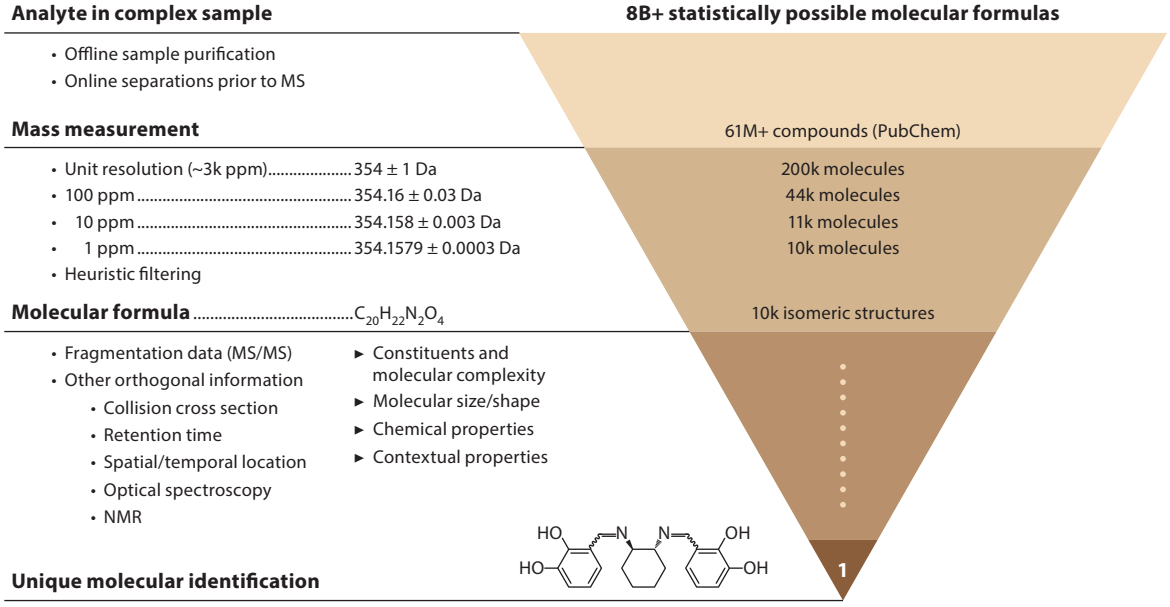

Figure 3

An example of the molecular characterization workflow using multidimensional mass spectrometry. From top to bottom: Over 8 billion possible molecular formulas between 0 and 2,000 Da exist. Obtaining the mass measurement allows database searching, which is illustrated by the over 61 million compounds indexed in PubChem as of January 2016. Subsequent levels of mass accuracy reduce the number of possible molecular formulas from over 200,000 (unit resolution) to approximately 10,000 at 1 ppm mass accuracy for the example mass of 354 Da. Using higher mass accuracy and/or a heuristic filtering approach obtains a unique molecular formula, which still represents several thousand isomeric compounds. Obtaining a unique molecular identity requires additional measurement dimensions, such as MS/MS, LC, IM, and perhaps measurements from other analytical techniques (e.g., optical spectroscopy, NMR). Abbreviations: IM, ion mobility; LC, liquid chromatography; MS/MS, tandem mass spectrometry; NMR, nuclear magnetic resonance; ppm, parts per million.

chromatography is utilized to reduce complexity but also to avoid ion suppression effects, which are inherent in all MS ion sources (96). The latter motivation is important because it is often stated that post-ionization separations such as IM can offset the need for condensed-phase separations (electrophoresis or chromatography), but such strategies cannot address limitations of the ion source itself. Once a mass measurement is obtained, the number of possible structures can be reduced significantly. In this example, a mass of 354 Da at unit resolution represents over 200,000 possible chemical structures indexed in PubChem, which contains over 60 million verified compounds as of 2016. At 100 ppm (354.16 ± 0.03 Da) the number of structures reduces to approximately 44,000, and at 10 ppm (354.158 ± 0.003 Da) there are just over 11,000 indexed structures. With a combination of high mass accuracy (approximately 1 ppm) and invoking filtering rules based on isotope pattern matching and probable structures (95), a molecular formula can be assigned to the analyte with high confidence. Marshall and colleagues (97) have demonstrated that a mass accuracy on the order of 0.1 mDa (0.2 ppm at 500 Da) is necessary to unambiguously assign a molecular formula based on mass measurement alone, and this level of high mass accuracy has been demonstrated for FTICR (98) and Orbitrap MS (99). Whereas these examples demonstrate that it is tractable to assign a unique molecular formula based on the mass measurement alone, the molecular formula is not a specific descriptor of the analyte. For example, **Figure 3** demonstrates that there are still over 10,000 possible isomeric structures in PubChem for this particular chemical formula ($C_{20}H_{22}N_2O_4$). To translate molecular formula information to a unique chemical structure, other orthogonal pieces of information, such as MS/MS data, must be utilized.

Parts per million (ppm): the difference in the measured mass to the exact mass multiplied by 10^6

Mass defect: the difference in the nominal mass to the exact mass due to the nuclear binding energy

To place this example in the context of big data, **Figure 4** illustrates the support and limitations that large-scale databases can provide for MS-based characterizations. The current scope of the PubChem Compound Database for molecules between 0 and 1,000 Da is projected in the histogram in **Figure 4*a***. Two previous surveys of the database are also shown (95, 100), which provide a sense of the volume of data that is currently being generated. Smaller molecules are being added at a faster rate, which has shifted the distribution of compounds represented since the 2007 survey. The survey conducted in this study reveals a bimodal distribution at low mass, which we interpret as reflecting the greater effort being made in comprehensive annotation of the smaller molecules. Over time, we anticipate that the distribution will normalize and shift back to higher mass, where more possible isomeric structures reside in chemical space. The dotted line in **Figure 4*a*** depicts the theoretically possible unique molecular formulas based on valence rules (95). At low mass, the number of possible isomeric structures for each molecular formula decreases significantly, but regardless, a very large number of isomeric structures less than approximately 400 Da are represented in the PubChem database. For example, valence and chemical stability rules suggest that the number of possible molecular formulas at approximately 250 Da for an organic molecule (C, H, N, and O) is on the order of 5,000, but empirically over 200,000 validated chemical structures are in this mass range. Even at this tractable mass range, the challenge of assigning a unique structure to an analyte is still quite formidable.

The inset in **Figure 4*a*** shows the distribution of molecules within a 10 Da mass window, and at this level of zoom the so-called forbidden zones resulting from quantized mass spacing are clearly visible (101). Thus, although MS has a very high peak capacity, much of the possible signal clusters within narrow regions of mass space to an extent that is specific to the atomic composition originating from the mass defect [properly, the mass excess (77)]. In this 10 Da window, there are over half a million verified chemical structures.

Figure 4*b* illustrates three panels of increasing levels of zoom in which the histogram resolutions are scaled to different mass accuracies. A 1 Da mass window at 100 ppm focuses on a single cluster of molecular structures within the database, containing over 200,000 structures. Higher

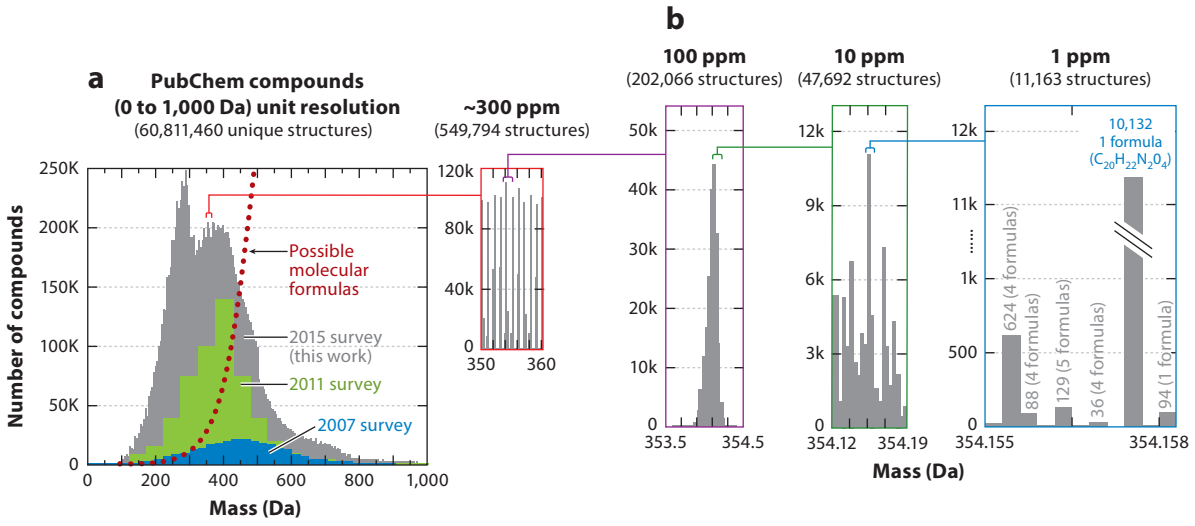

Figure 4

An illustration of the amount of information density present at different levels of mass measurement accuracy, using the validated entries in the PubChem Compound Database. (*a*) The distribution of molecules between 0 and 1,000 Da in the PubChem Compound Database, as surveyed in this review and in two previous surveys from 2007 and 2011. As new compounds are discovered and archived, the distribution has shifted to lower mass, with most entries currently centered between 100 and 600 Da. The dotted line illustrates theoretical molecular formulas determined from chemical stability rules, indicating that most of these entries are isomers. The inset zooms in on a 10 Da window, in which over half a million compounds are represented. (*b*) At increasing levels of mass accuracy, the number of possible molecular formulas can be reduced to a few thousand, but in one extreme case shown at 1 ppm, one formula is represented by over 10,000 isomers in the database. Mass spectrometry can significantly reduce complexity, but it cannot fully address molecular characterization without other dimensions of information.

mass bin resolution (10 ppm) brings this number of structures down to approximately 48,000, and at 1 ppm mass accuracy, the number of structures can be reduced to approximately 10,000, here represented by five possible molecular formulas or less for each 1 ppm resolution bin. For most molecular formulas in this range, only tens to hundreds of structures are represented, but highlighted is the dramatic case described above in which one molecular formula ($C_{20}H_{22}N_2O_4$) represents over 10,000 isomeric structures indexed in the PubChem Compound Database.

Integrating additional separation dimensions with MS is thus necessary to address sample complexity. Although many combinations of condensed- and gas-phase separations have been demonstrated to work well with MS, there are inherent technological limitations imposed for specific combinations. For example, seamless coupling of LC to MS requires a continuous liquid stream ion source operated at ambient pressure, such as electrospray or atmospheric pressure chemical ionization (102). On the other hand, MSI is conventionally coupled to MS by means of a pulsed ion beam or laser source to provide high-speed, discrete ionization of spatial locations on the sample (103, 104). Developments in MSI-MS using liquid sampling probes, which reduce sample pretreatment at a cost of throughput and spatial resolution, are now beginning to emerge (105, 106). IM-MS requires transferring ions across disparate pressure regions and analyses in an efficient manner, and this challenge has been addressed through electrodynamic ion optics and the nesting of analytical timescales (83, 107, 108). Despite some limitations, numerous combinations of separations have been adapted to MS, each of which provides a unique level of information when combined.

COMPARISON OF ORTHOGONAL SEPARATION DIMENSIONS

Mass Defect Analysis

The accurate mass measurement provides a highly specific measurement of an intrinsic property of the analyte, the exact mass, but as noted above, the mass measurement alone does not provide specific information beyond the chemical composition. Several MS studies have exploited the additional information that can be gained from derivative comparisons (i.e., change in mass) using the single dimension of MS analysis, namely through the correlation of the small mass shift from nominal mass owing to the mass defect. Mass defect refers to the change in the nominal mass due to the binding energy of nucleons, which manifests as the decimal mass measurement in high-resolution MS data (109). Because this mass shift reflects the chemical composition of the analyte, an orthogonal comparison between the nominal mass and the mass defect groups the measurements into chemical class families, which provides a convenient means of locating related chemical species in a complex MS spectrum. MS measurements are based on the International Union of Pure and Applied Chemistry (IUPAC) mass scale, which normalizes the measurement to carbon-12 ($^{12}C = 12$ Da), but mass defect analysis commonly utilizes a rescaled mass axis to help identify small mass differences relative to a reference mass that is more representative of the molecule of interest.

Two mass defect scales that have been utilized are the Kendrick scale and the averagine scale. The Kendrick scale normalizes the mass axis to CH_2, which is 14 Da (110), and has found widespread utility in analyzing the chemical constituents in petroleum (111, 112), as well as in MS analysis of lipids (113). **Figure 5** depicts a Kendrick mass defect plot, which aligns families of compounds into easily discernible groups, as is shown for the separation of constituents contained in crude oil (111). The averagine scale is normalized to the averagine subunit, which is a theoretical amino acid based on the statistically weighted occurrence of amino acids in the Protein Information Resource (PIR) protein database as of 1995 ($C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$) (114, 115). The averagine mass scale has been used in MS proteomics to aid in peptide identification (116), although axis rescaling is not necessary, and peptide mass defect analysis has been carried out using the conventional IUPAC scale (117). Of note is an effort to theoretically enumerate all possible tryptic peptides of 3.5 kDa or less to support mass defect–based identifications (118, 119), among other initiatives. Mass defect analysis is commonly used in drug metabolism studies, where it is desirable to search out exogenous metabolites that possess chemical compositions related to the drug species of interest (120–122). Mass defect analysis is also particularly effective at identifying surfactant and halogenated compound contaminants in complex samples (123, 124).

Ion Mobility and Mobility-Mass Correlations

IM has emerged in recent years as a mature analytical technique capable of rapid separations that seamlessly integrate into a variety of MS instrument platforms (83). The benefit of the separation is through increasing the peak capacity of the analysis as well as providing an additional level of discrimination by filtering out desirable signals from complex matrices. IM also provides complementary information to MS in the form of structural information by means of the gas-phase collision cross section (CCS), which is now a relatively routine parameter determined from a variety of IM experiments. In contrast to that from MS, however, the structural information from IM is a less-specific descriptor of the analyte than of the mass, as the CCS is an orientationally average shape parameter that reduces the three-dimensional structure to a two-dimensional area (125). CCS is also an extrinsic property that depends on the specific parameters of the experiment,

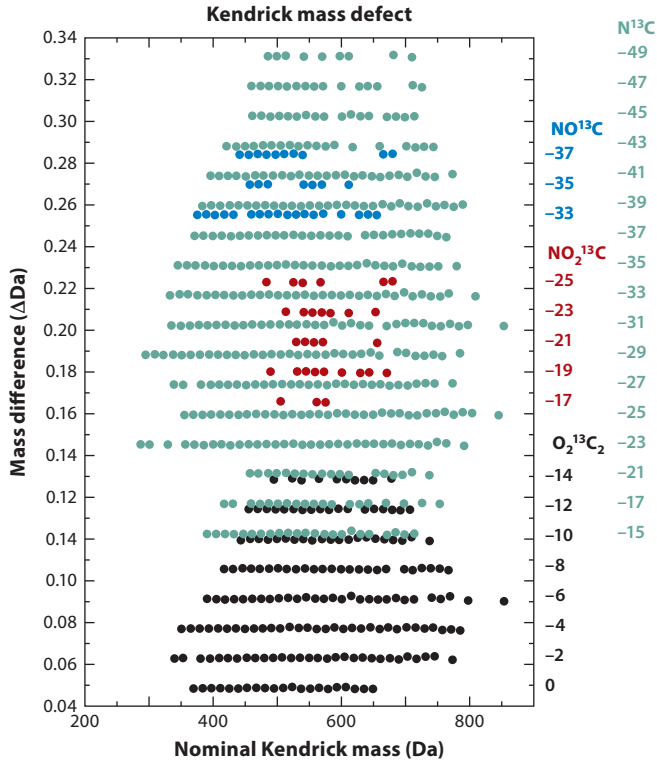

Figure 5

A Kendrick mass defect plot of crude oil. The plot projects the nominal (integer) Kendrick mass on the x-axis and the difference between nominal and exact Kendrick mass on the y-axis. In this way, chemically similar compounds align horizontally, and chemical class families group into distinct regions on the plot. Here, odd-mass species (e.g., carbon-13) are projected to validate the assignments made for the primary even-mass species in the original work. The numbers below each class header at right correspond to the number of degrees of unsaturation within each chemical family. Reprinted with permission from Reference 111. Copyright 2001 American Chemical Society.

and so CCS values are reported with statistics (standard deviation and number of observations) to gauge the relative specificity of the measurement. Despite these limitations, CCS databases have facilitated molecular identification by providing an additional parameter for characterization (126, 127). Aided by computational methods (128), the CCS can also be used to infer more detailed structural information regarding the analyte (129–131).

Combined Ion Mobility–Mass Spectrometry and Mass Defect Analysis

Although IM alone provides some level of specificity in the measurement, additional information is gained when it is combined with MS. At a fundamental level, IM-MS separates analytes by size and mass, which collectively describe the relative gas-phase densities of different structural populations (132, 133). These mobility-mass correlations are useful for discerning related structural families toward chemical-class-specific filtering (134, 135) and characterization schemes (136). As noted above, class information can also be inferred from analysis using the mass defect originating from the MS measurement. A recent study conducted in the authors' laboratory for lipids is depicted

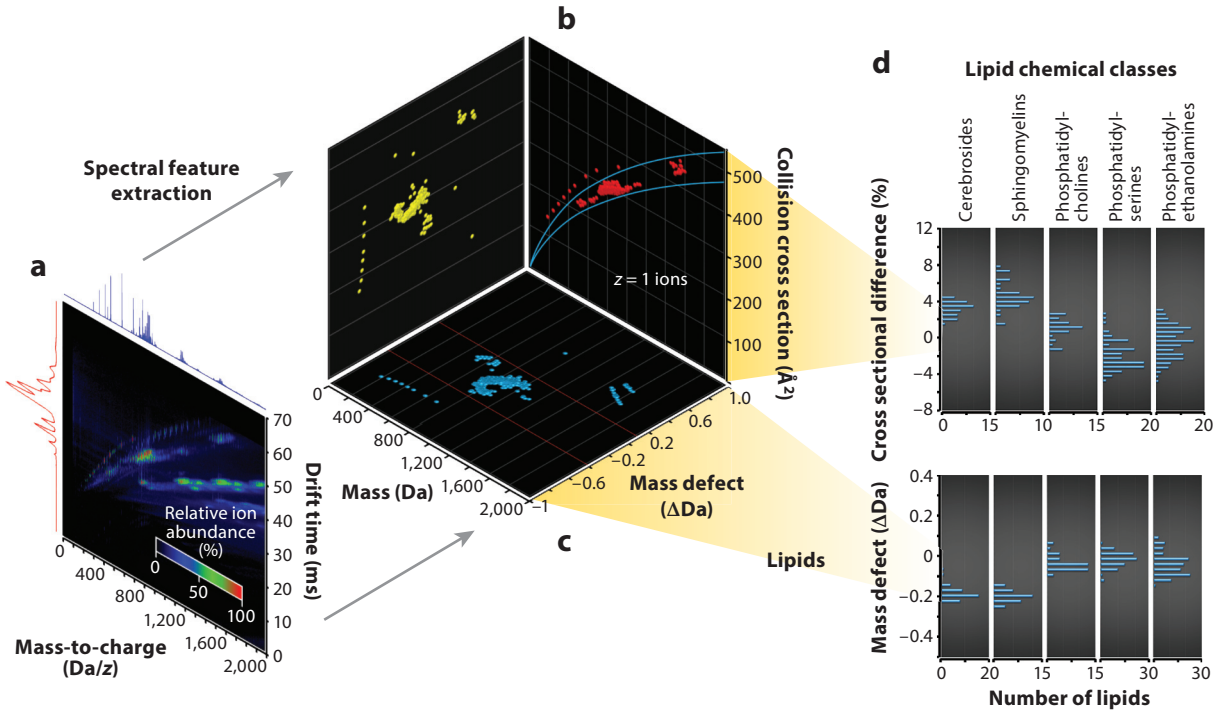

Figure 6

A multidimensional analysis of five lipid classes (two sphingolipids and three glycerophospholipids) using data obtained from an IM-MS experiment. (*a*) The raw IM-MS spectrum is projected as a heat map, with *m/z* on the x-axis, IM drift time on the y-axis, and signal intensity on the color mapping scale. (*b*) Feature extraction of singly charged ions is performed, resulting in a mass versus collision cross section conformational space plot. (*c*) The accurate mass measurement (approximately 5 ppm in this work) can also be subjected to a mass defect scaling, resulting in a mass versus mass defect plot. (*d*) Both the IM-MS conformational space map and the mass defect plot reveal groupings of data based on their respective lipid class, with sphingolipids separated from phospholipids. Abbreviations: IM, ion mobility; MS, mass spectrometry; *m/z*, mass-to-charge ratio; ppm, parts per million.

in **Figure 6**, which compares mass defect analysis to IM-MS correlations for five classes of lipids (137). Raw IM-MS data are visualized in a heat map projection (**Figure 6*a***), which projects the mass-to-charge ratio (*m/z*) (x-axis) versus the IM drift time (y-axis) while retaining the third dimension of signal intensity as a color map. In this particular example, a tricolor gradient is used to differentiate the islands of high signal abundance for polar lipids and a series of alkyl ammonium salts introduced as internal calibrants that form distinct class-specific trends in the raw data (136). Spectral features are then extracted, in this case for singly charged analytes, and drift times are converted to CCS (**Figure 6*b***). Using the accurate mass measurement, the data can also be projected as mass versus mass defect (**Figure 6*c***), which reveals chemical relationships in the extracted ion signals. Here, the inorganic ammonium salts are easily differentiated from the lipids, and two lipid subclasses, sphingolipids and glycerophospholipids, can be differentiated from each other using either the IM-MS projection (**Figure 6*b***) or the MS–mass defect projection (**Figure 6*c***). This differentiation of lipid subclasses is observed in a distribution analysis of both dimensions of separation (**Figure 6*d***), illustrating that similar chemical class information can be obtained from both IM-MS and high-accuracy MS. Thus, IM-MS provides an additional level of information that goes beyond the sum of its parts.

NOVEL STRATEGIES FOR VISUALIZING BIG DATA

As the rate and volume of data increase, visualizing important information derived from the data becomes increasingly challenging. Because of the fundamental limitations of human perception, large-scale data cannot be adequately comprehended unless they are reduced to lower-dimensional projections. Successful data visualization strategies thus simplify the level of data complexity and also incorporate familiar and intuitive visual cues that help infer connections between higher dimensions or otherwise provide access to the unseen higher dimensions. We briefly review a few recent and noteworthy means of visualizing large datasets.

Cloud Plots

A contemporary metabolomics experiment incorporating MS analysis typically detects on the order of several thousand metabolites, representing a diverse array of chemical classes (138–140). Because of limited time and information, the majority of these detected metabolites are characterized only by a few nonspecific descriptors originating from the multidimensional experiment, such as retention time, signal intensity, and molecular mass (141). As such, it is highly desirable to find underlying relationships between discrete metabolites that can be directly correlated to the experimental measurements. Recently, Patti et al. (142) describe a novel visualization format known as a cloud plot, where detected metabolites are projected as bubbles onto a plot of retention time versus *m/z*.

A cloud plot of metabolite data from a sepsis study is depicted in **Figure 7**, where the vertical *m/z* scaling represents a positive or negative fold-change for each metabolite, and the size of each bubble represents the fold-change of the corresponding metabolite. In this way, the cloud plot projects five dimensions of information simultaneously: retention time (x-axis), *m/z* (y-axis), directional fold-change of signal intensity (positive or negative y-axis projection), magnitude of

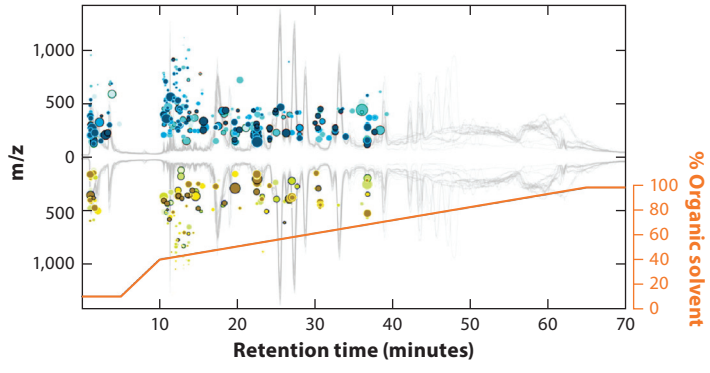

Figure 7

A cloud plot of LC-MS metabolomics data obtained from a mouse sepsis model. In a cloud plot, the LC retention time is projected on the x-axis, the positive/negative fold-change *m/z* is plotted on the y-axis, and individual spectral features are plotted as bubbles, with size indicating the magnitude of the fold-change and color shading indicating the statistical significance. Outlined bubbles indicate positive database matches for that particular feature. Also shown are the superimposed LC chromatogram and LC gradient (*orange*), which provide hydrophobic and hydrophilic information regarding the analytes. In this way, the cloud plot projects a large number of measurement dimensions onto an intuitive graphic that provides ready access to each level of information. Reprinted with permission from Reference 142. Copyright 2013 American Chemical Society. Abbreviations: LC, liquid chromatography; MS, mass spectrometry; *m/z*, mass-to-charge ratio.

fold-change (bubble size), and the corresponding *p*-value from a statistical *t*-test (bubble color shading). The LC chromatogram and gradient method are also superimposed on the plot, providing additional information regarding the separation method and, by association, the relative chemical hydrophobicity of each detected metabolite. In this particular example, 29,920 extracted spectral features (unique retention time and *m/z*) are distilled down to the 487 most significant metabolites (p-value $\leq 1.0 \times 10^{-4}$, fold-change ≥ 3), which are then projected onto the cloud plot, providing a means of directly accessing only the most relevant signals originating from the experiment. The cloud plot is an interactive graphic implemented in XCMS Online such that clicking on each individual metabolite bubble reveals detailed information such as the putative metabolite assignment from a METLIN metabolite database query (143). Although developed specifically for metabolomics data, the underlying concept of cloud plots as an intuitive and interactive visualization tool has broad relevance in all multidimensional MS initiatives.

p-value: statistical probability that another result will be found that is similar to or more significant than the current result

Self-organizing map (SOM): a map organized by an artificial neural network to cluster data based on learned similarities in the primary dimensions

Self-Organizing Maps

The majority of MS-based metabolite data have been interpreted through the use of comparative, multivariate statistics that performs binary comparisons of dimensionally transformed data to find underlying relationships between detected signals. Of these, principal component analysis (PCA) is the most widely utilized in metabolomics. PCA enables the visualization of clusters of related signals in the data but does not provide direct, quantitative information regarding analyte similarity, as the data have been mathematically transformed prior to conducting comparisons (144). One complementary approach to identifying relationships between detected signals in large datasets is to cluster similar signals using a self-organizing map (SOM). SOMs have found use in untargeted MS-based metabolomics research, where biological relationships are inferred from the clustered metabolites (145, 146). Although an SOM originating from a single metabolomics experiment can provide insight into the metabolites expressed in one experimental context, the information content is greatly increased when differential analysis is utilized. SOM differential analysis allows numerous comparisons between SOMs originating from discrete samples and/or time-points of the experiment. Another benefit of the SOM approach is the ability to decrypt specific nodes on the map to extract primary feature information, such as retention time and *m/z* (147, 148).

A recent example of using SOMs to systems-level mapping of molecules for an organ-on-chip human liver bioreactor exposed to acetaminophen (APAP) is shown in **Figure 8** (149). This particular synthetic organ consists of a network of hollow fibers around which are seeded with cultured cells harvested from a human cadaver liver. Cell culture media is perfused through the hollow fibers, and time-points from the waste stream are analyzed by IM-MS, which provides a comprehensive analysis of the metabolites secreted from the cells. In this example, two sets of SOMs are created, representing positive (**Figure 8*a***) and negative (**Figure 8*b***) fold-change of detected features. Following exposure to APAP, several regions of the map change in intensity, representing a fold-change for a group of metabolite features. Groups of nodes can be decrypted at any time to obtain primary information that can be utilized for molecular identification (**Figure 8*c***). In this case, the upregulation of APAP-bound glucuronic acid is observed in addition to the dysregulation of bile acid, the latter of which is a classic marker for liver stress.

These few examples underscore the importance of effective data visualization strategies for MS-based data intensive experiments. Whereas big data holds the promise of new discoveries, the reality of achieving these discoveries is contingent upon our ability to effectively navigate and find connections between multiple dimensions of information.

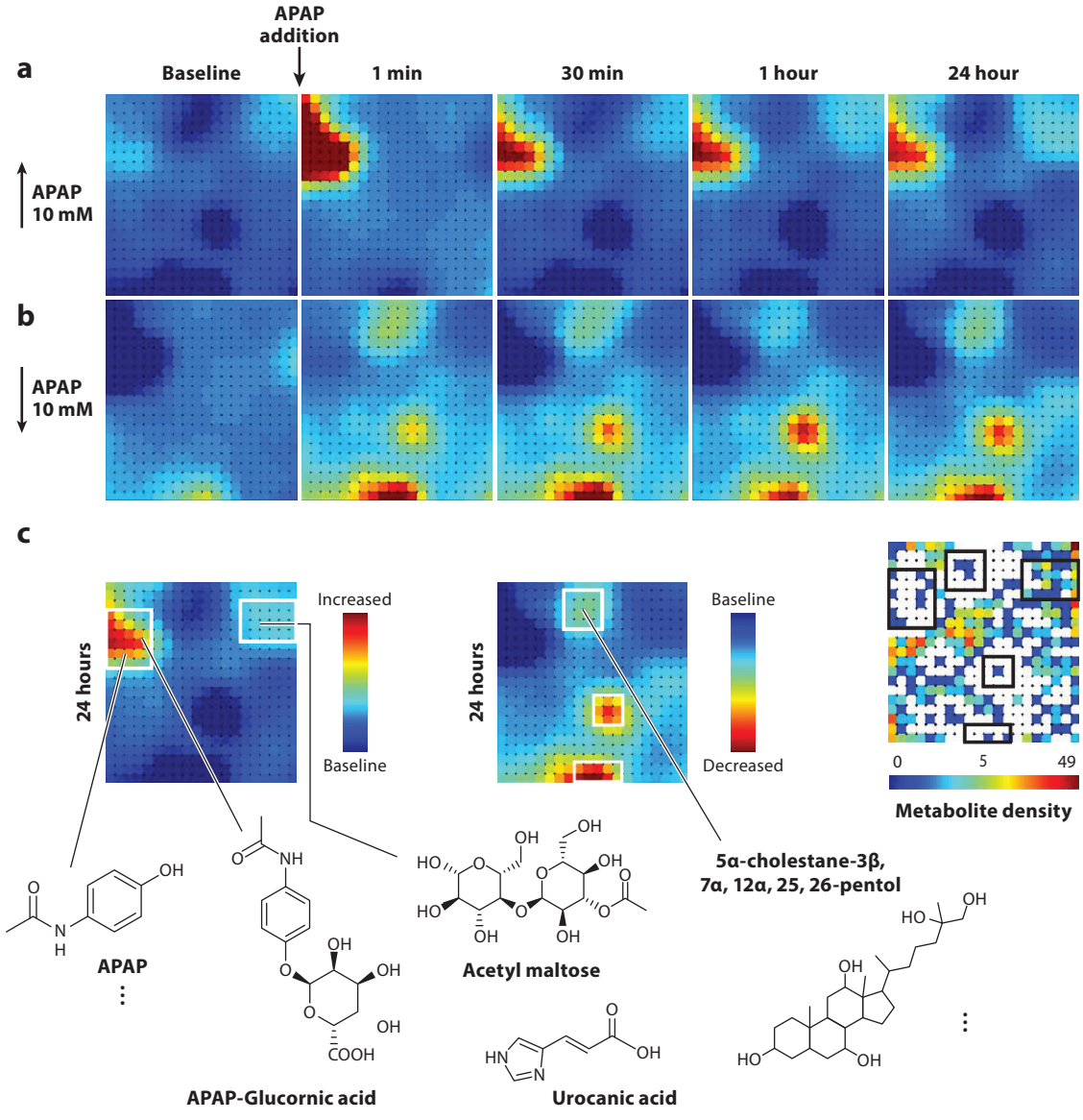

Figure 8

Time-course monitoring of a lab-on-a-chip liver bioreactor's exposure to APAP using SOMs, from data obtained by LC-IM-MS. Each SOM organizes detected features based on similarities in signal intensity, and sets of SOMs are differentially analyzed to determine signals that are (*a*) increased in abundance and (*b*) decreased in abundance as a function of time. (*c*) A single SOM from any point in time can be decrypted to extract the primary measurement information across all dimensions, which can be used for molecular identification. In this case, several metabolites related to liver stress are upregulated and downregulated in response to APAP exposure, and these signatures for metabolic stress persist even 24 hours following exposure. Reprinted with permission from Reference 149. Copyright 2016 American Association for Clinical Chemistry. Abbreviations: APAP, acetaminophen; IM, ion mobility; LC, liquid chromatography; MS, mass spectrometry; SOM, self-organizing map.

CONCLUSION

MS is a rapidly evolving field that now encompasses myriad allied analytical techniques and disciplines. The current state of the art in hybrid MS instrumentation generates data that are dense with information and can be obtained much faster than they can be interpreted. This is the big data challenge, and it can be considered the breadth and scope of analytical space, which includes the information from each separation dimension, the number of total separation dimensions, and the derived information from comparing dimensions to one another, all placed within the context of a spatial and temporal location. In this context, one may argue that, similar to chemical space, analytical space is both immense and vastly unexplored, providing unique opportunities for future innovation and discovery. We can speculate many things about the future of MS-based analytical sciences, but one prediction we are certain about is given the continuing integration of high-resolution instrumentation into hybrid architectures, the field of MS will retain a place at the forefront of big data.

SUMMARY POINTS

1. Multidimensional MS generates large volumes of data at a high rate, representing different measurement dimensions, which makes it a big data driver.
2. On the basis of exponential growth trends in the chemical and computer sciences, the data generated by MS-based analytical techniques are projected to increase significantly in the next few years.
3. The increase in multidimensional MS data generation will be driven by technological advances, as higher-resolving-power instruments are developed and integrated into hybrid analytical architectures, as is the current trend in the field.
4. Looking to other areas of science where big data is generated, such as astronomy and particle physics, it is recommended that the field of MS embrace distributed computing and open-source initiatives.
5. Multidimensional MS data are highly information dense, with information generated from the primary data dimensions as well as gained from conducting binary or higher-order comparisons between data dimensions.
6. Creative means of visualizing highly dimensional datasets in an intuitive and comprehensible manner are necessary to deal with multidimensional MS datasets and address the limitations of human perception.
7. The size of analytical space representing the scope of information that can be gained from multidimensional analysis techniques is vast and largely unexplored.

DISCLOSURE STATEMENT

The authors are unaware of any potential bias that may affect the objectivity of the review, but do acknowledge collaborative arrangements with Agilent Technologies (Santa Clara, CA) and Waters Corporation (Milford, MA). The Vanderbilt University Center for Innovative Technology is designated as an Agilent Thought Leader Laboratory and a Waters Center of Innovation.

ACKNOWLEDGMENTS

The authors are grateful to the following sources of financial support for this work: the National Science Foundation (no. CHE-1229341), the National Center for Advancing Translational Sciences of the National Institutes of Health (no. 5UH3TR000491-04 and 3UH3TR000491-04S1), and the National Institute of General Medical Sciences of the National Institutes of Health (no. R01GM92218). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors also acknowledge intramural support for this work provided by the Vanderbilt University Center for Innovative Technology, the Vanderbilt Institute for Chemical Biology, the Vanderbilt Institute for Integrative Biosystems Research and Education, and the Vanderbilt University College of Arts and Sciences.

LITERATURE CITED

1. Aebersold R, Mann M. 2003. Mass spectrometry-based proteomics. *Nature* 422:198–207
2. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, et al. 2014. Mass-spectrometry-based draft of the human proteome. *Nature* 509:582–87
3. Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. 2014. A draft map of the human proteome. *Nature* 509:575–81
4. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, et al. 2015. Tissue-based map of the human proteome. *Science* 347:1260419
5. Uhlén M, Ponten F. 2005. Antibody-based proteomics for human tissue profiling. *Mol. Cell. Proteomics* 4:384–93
6. Marx V. 2015. Mapping proteins with spatial proteomics. *Nat. Methods* 12:815–19
7. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, et al. 2007. HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35:D521–26
8. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al. 2013. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res.* 41:D801–7
9. Junot C, Fenaille F, Colsch B, Bécher F. 2014. High resolution mass spectrometry based techniques at the crossroads of metabolic pathways. *Mass Spectrom. Rev.* 33:471–500
10. Yetukuri L, Ekroos K, Vidal-Puig A, Oresic M. 2008. Informatics and computational strategies for the study of lipids. *Mol. BioSyst.* 4:121–27
11. Hood L, Heath JR, Phelps ME, Lin B. 2004. Systems biology and new technologies enable predictive and preventative medicine. *Science* 306:640–43
12. Nicholson JK, Wilson ID. 2003. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2:668–76
13. Feng X, Liu X, Luo Q, Liu B-F. 2008. Mass spectrometry in systems biology: an overview. *Mass Spectrom. Rev.* 27:635–60
14. Graessel A, Hauck SM, von Toerne C, Kloppmann E, Goldberg T, et al. 2015. A combined omics approach to generate the surface atlas of human naive $CD4^+$ T cells during early T-cell receptor activation. *Mol. Cell. Proteomics* 14:2085–102
15. Bohacek RS, McMartin C, Guida WC. 1996. The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16:3–50
16. Peironcely JE, Reijmers T, Coulier L, Bender A, Hankemeier T. 2011. Understanding and classifying metabolite space and metabolite-likeness. *PLOS ONE* 6:e28966
17. Gurard-Levin ZA, Scholle MD, Eisenberg AH, Mrksich M. 2011. High-throughput screening of small molecule libraries using SAMDI mass spectrometry. *ACS Comb. Sci.* 13:347–50
18. de Rond T, Danielewicz M, Northen T. 2015. High throughput screening of enzyme activity with mass spectrometry imaging. *Curr. Opin. Biotechnol.* 31:1–9
19. Mardis ER. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133–41

20. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, et al. 2015. Big data: astronomical or genomical? *PLOS Biol.* 13:e1002195
21. Laney D. 2001. *3D data management: controlling data volume, velocity, and variety*. File 949. Application Delivery Strategies. Stamford, CT: META Group
22. Lusher SJ, McGuire R, van Schaik RC, Nicholson CD, de Vlieg J. 2014. Data-driven medicinal chemistry in the era of big data. *Drug Discov. Today* 19:859–68
23. Askenazi M, Webber JT, Marto JA. 2011. mzServer: web-based programmatic access for mass spectrometry data analysis. *Mol. Cell. Proteomics* 10:M110.003988
24. Bird I. 2011. Computing for the Large Hadron Collider. *Annu. Rev. Nuclear Part. Sci.* 61:99–118
25. Reymond J-L, Ruddigkeit L, Blum L, van Deursen R. 2012. The enumeration of chemical space. *Wiley Interdiscip. Rev. Comp. Mol. Sci.* 2:717–33
26. Fuller RB, Marks RW. 1973. *The Dymaxion World of Buckminster Fuller*. Garden City, NY: Anchor Press/Doubleday
27. Wang L. 2015. Chemical Abstract Service marks multiple milestones. *Chemical and Engineering News*. July 1. American Chemical Society
28. Bolton EE, Wang Y, Thiessen PA, Bryant SH. 2008. Chapter 12 - PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* 4:217–41
29. PubChem Substance Database. **https://pubchem.ncbi.nlm.nih.gov.** Accessed September 28, 2015
30. PubChem Compound Database. **https://pubchem.ncbi.nlm.nih.gov.** Accessed September 28, 2015
31. Pence HE, Williams A. 2010. ChemSpider: an online chemical information resource. *J. Chem. Educ.* 87:1123–24
32. NIST Chemistry WebBook. 2015. NIST Standard Reference Database Number 69, eds. PJ Linstrom, WG Mallard. **http://webbook.nist.gov./** Accessed September 28, 2015
33. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, et al. 2014. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 42:D1083–90
34. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, et al. 2014. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.* 42:D1091–97
35. Lipinski C, Hopkins A. 2004. Navigating chemical space for biology and medicine. *Nature* 432:855–61
36. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. 2013. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* 135:7296–303
37. Bars I, Terning J. 2010. *Extra Dimensions in Space and Time*. New York: Springer-Verlag
38. Ruddigkeit L, Blum LC, Reymond J-L. 2013. Visualization and virtual screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* 53:56–65
39. Reymond J-L. 2015. The Chemical Space Project. *Acc. Chem. Res.* 48:722–30
40. Oprea TI, Gottfries J. 2001. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* 3:157–66
41. Engel T. 2006. Basic overview of chemoinformatics. *J. Chem. Inform. Model.* 46:2267–77
42. Varnek A, Baskin II. 2011. Chemoinformatics as a theoretical chemistry discipline. *Mol. Inform.* 30:20–32
43. Scifinder. 2015. Columbus, OH: Chem. Abstr. Serv. **https://scifinder.cas.org/**. Accessed September 28, 2015
44. Jinha AE. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publ.* 23:258–63
45. Powell JR. 2008. The quantum limit to Moore's law. *Proc. IEEE* 96:1247–48
46. Walter C. 2005. Kryder's law. *Sci. Am.* 293:32–22
47. Eldering CA, Sylla ML, Eisenach JA. 1999. Is there a Moore's law for bandwidth? *Commun. Mag. IEEE* 37:117–21
48. Sobel D. 2010. *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*. London: Bloomsbury
49. Misura KMS, Chivian D, Rohl CA, Kim DE, Baker D. 2006. Physically realistic homology models built with Rosetta can be more accurate than their templates. *PNAS* 103:5361–66
50. Kaufmann KW, Lemmon GH, DeLuca SL, Sheehan JH, Meiler J. 2010. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* 49:2987–98

51. Khatib F, Cooper S, Tyka MD, Xu K, Makedon I, et al. 2011. Algorithm discovery by protein folding game players. *PNAS* 108:18949–53
52. Khatib F, DiMaio F, Cooper S, Kazmierczyk M, Gilski M, et al. 2011. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nat. Struct. Mol. Biol.* 18:1175–77
53. Gilski M, Kazmierczyk M, Krzywda S, Zabranska H, Cooper S, et al. 2011. High-resolution structure of a retroviral protease folded as a monomer. *Acta Crystallogr. D* 67:907–14
54. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, et al. 2012. Increased diels-alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* 30:190–92
55. Bradley JC, Lancashire R, Lang A, Williams A. 2009. The Spectral Game: leveraging Open Data and crowdsourcing for education. *J. Cheminform.* 1:9
56. Du L, Robles AJ, King JB, Powell DR, Miller AN, et al. 2014. Crowdsourcing natural products discovery to access uncharted dimensions of fungal metabolite diversity. *Angew. Chem. Int. Ed.* 53:804–9
57. Martin SF, Falkenberg H, Dyrlund TF, Khoudoli GA, Mageean CJ, Linding R. 2013. PROTEINCHALLENGE: crowd sourcing in proteomics analysis and software development. *J. Proteomics* 88:41–46
58. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. 2014. Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins Struct. Funct. Bioinform.* 82:1–6
59. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. 2012. Wisdom of crowds for robust gene network inference. *Nat. Methods* 9:796–804
60. Bishop CM. 2006. *Pattern Recognition and Machine Learning*. New York: Springer
61. Smalheiser NR. 2002. Informatics and hypothesis-driven research. *EMBO Rep.* 3:702–2
62. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S. 2015. The rise of "big data" on cloud computing: review and open research issues. *Inf. Syst.* 47:98–115
63. Chen T, Zhao J, Ma J, Zhu Y. 2015. Web resources for mass spectrometry-based proteomics. *Genom. Proteom. Bioinform.* 13:36–39
64. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. 2012. XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* 84:5035–39
65. Rinehart D, Johnson CH, Nguyen T, Ivanisevic J, Benton HP, et al. 2014. Metabolomic data streaming for biology-dependent data acquisition. *Nat. Biotechnol.* 32:524–27
66. Rübel O, Greiner A, Cholia S, Louie K, Bethel EW, et al. 2013. OpenMSI: a high-performance web-based platform for mass spectrometry imaging. *Anal. Chem.* 85:10354–61
67. Fischer CR, Ruebel O, Bowen BP. 2016. An accessible, scalable ecosystem for enabling and sharing diverse mass spectrometry imaging analyses. *Arch. Biochem. Biophys.* 589:18–26
68. Malm E, Srivastava V, Sundqvist G, Bulone V. 2014. APP: an Automated Proteomics Pipeline for the analysis of mass spectrometry data based on multiple open access tools. *BMC Bioinform.* 15:441
69. Mohammed Y, Mostovenko E, Henneman AA, Marissen RJ, Deelder AM, Palmblad M. 2012. Cloud parallel processing of tandem mass spectrometry based proteomics data. *J. Proteome Res.* 11:5101–8
70. Muth T, Peters J, Blackburn J, Rapp E, Martens L. 2013. ProteoCloud: a full-featured open source proteomics cloud computing pipeline. *J. Proteomics* 88:104–8
71. Deutsch EW, Mendoza L, Shteynberg D, Slagel J, Sun Z, Moritz RL. 2015. Trans-Proteomic Pipeline, a standardized data processing pipeline for large-scale reproducible proteomics informatics. *Proteom. Clin. Appl.* 9:745–54
72. Slagel J, Mendoza L, Shteynberg D, Deutsch EW, Moritz RL. 2015. Processing shotgun proteomics data on the Amazon Cloud with the Trans-Proteomic Pipeline. *Mol. Cell. Proteom.* 14:399–404
73. Riffle M, Eng JK. 2009. Proteomics data repositories. *Proteomics* 9:4653–63
74. Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA. 2015. Making proteomics data accessible and reusable: current state of proteomics databases and repositories. *Proteomics* 15:930–50
75. Karger BL, Snyder LR, Horvath C. 1973. *Introduction to Separation Science*. New York: Wiley
76. Giddings JC. 1984. Two-dimensional separations: concept and promise. *Anal. Chem.* 56:1258A–70A
77. Frahm JL, Howard BE, Heber S, Muddiman DC. 2006. Accessible proteomics space and its implications for peak capacity for zero-, one- and two-dimensional separations coupled with FT-ICR and TOF mass spectrometry. *J. Mass Spectrom.* 41:281–88
78. Barner-Kowollik C, Gruendling T, Falkenhagen J, Weidner S. 2012. *Mass Spectrometry in Polymer Chemistry*. New York: Wiley

79. Canterbury JD, Yi X, Hoopmann MR, MacCoss MJ. 2008. Assessing the dynamic range and peak capacity of nanoflow LC–FAIMS–MS on an ion trap mass spectrometer for proteomics. *Anal. Chem.* 80:6888–97
80. Schneider B, Nazarov E, Covey T. 2012. Peak capacity in differential mobility spectrometry: effects of transport gas and gas modifiers. *Int. J. Ion Mobil. Spectrom.* 15:141–50
81. Merenbloom SI, Bohrer BC, Koeniger SL, Clemmer DE. 2007. Assessing the peak capacity of IMS–IMS separations of tryptic peptide ions in He at 300 K. *Anal. Chem.* 79:515–22
82. May JC, McLean JA. 2013. The influence of drift gas composition on the separation mechanism in traveling wave ion mobility spectrometry: insight from electrodynamic simulations. *Int. J. Ion Mobil. Spectrom.* 16:85–94
83. May JC, McLean JA. 2015. Ion mobility-mass spectrometry: time-dispersive instrumentation. *Anal. Chem.* 87:1422–36
84. Causon TJ, Hann S. 2015. Theoretical evaluation of peak capacity improvements by use of liquid chromatography combined with drift tube ion mobility-mass spectrometry. *J. Chromatogr. A* 1416:47–56
85. McLean JA, Ruotolo BT, Gillig KJ, Russell DH. 2005. Ion mobility-mass spectrometry: a new paradigm for proteomics. *Int. J. Mass Spectrom.* 240:301–15
86. Neue UD. 2005. Theory of peak capacity in gradient elution. *J. Chromatogr. A* 1079:153–61
87. Neue UD. 2008. Peak capacity in unidimensional chromatography. *J. Chromatogr. A* 1184:107–30
88. Moore AW, Jorgenson JW. 1995. Comprehensive three-dimensional separation of peptides using size exclusion chromatography/reversed phase liquid chromatography/optically gated capillary zone electrophoresis. *Anal. Chem.* 67:3456–63
89. Tia S, Herr AE. 2009. On-chip technologies for multidimensional separations. *Lab. Chip* 9:2524–36
90. Bruce JE, Anderson GA, Wen J, Harkewicz R, Smith RD. 1999. High-mass-measurement accuracy and 100 sequence coverage of enzymatically digested bovine serum albumin from an ESI-FTICR mass spectrum. *Anal. Chem.* 71:2595–99
91. Prentice BM, Chumbley CW, Caprioli RM. 2015. High-speed MALDI MS/MS imaging mass spectrometry using continuous raster sampling. *J. Mass Spectrom.* 50:703–10
92. Stauber J, MacAleese L, Franck J, Claude E, Snel M, et al. 2010. On-tissue protein identification and imaging by MALDI–ion mobility mass spectrometry. *J. Am. Soc. Mass Spectrom.* 21:338–47
93. Trede D, Schiffler S, Becker M, Wirtz S, Steinhorst K, et al. 2012. Exploring three-dimensional matrix-assisted laser desorption/ionization imaging mass spectrometry data: three-dimensional spatial segmentation of mouse kidney. *Anal. Chem.* 84:6079–87
94. Kind T, Fiehn O. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* 2:23–60
95. Kind T, Fiehn O. 2007. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.* 8:1–20
96. Annesley TM. 2003. Ion suppression in mass spectrometry. *Clin. Chem.* 49:1041–44
97. Kim S, Rodgers RP, Marshall AG. 2006. Truly "exact" mass: elemental composition can be determined uniquely from molecular mass measurement at ~0.1 mDa accuracy for molecules up to ~500 Da. *Int. J. Mass Spectrom.* 251:260–65
98. Savory JJ, Kaiser NK, McKenna AM, Xian F, Blakney GT, et al. 2011. Parts-per-billion Fourier transform ion cyclotron resonance mass measurement accuracy with a "walking" calibration equation. *Anal. Chem.* 83:1732–36
99. Scheltema RA, Kamleh A, Wildridge D, Ebikeme C, Watson DG, et al. 2008. Increasing the mass accuracy of high-resolution LC-MS data using background ions—a case study on the LTQ-Orbitrap. *Proteomics* 8:4647–56
100. Green FM, Gilmore IS, Seah MP. 2011. Mass spectrometry and informatics: distribution of molecules in the PubChem database and general requirements for mass accuracy in surface analysis. *Anal. Chem.* 83:3239–43
101. Nefedov AV, Mitra I, Brasier AR, Sadygov RG. 2011. Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *J. Proteome Res.* 10:4150–57
102. Yergey AL, Edmonds CG, Lewis IAS, Vestal ML. 2013. *Liquid Chromatography/Mass Spectrometry: Techniques and Applications*. New York: Springer

103. Norris JL, Caprioli RM. 2013. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chem. Rev.* 113:2309–42
104. Watrous JD, Dorrestein PC. 2011. Imaging mass spectrometry in microbiology. *Nat. Rev. Microbiol.* 9:683–94
105. Lanekoff I, Burnum-Johnson K, Thomas M, Cha J, Dey S, et al. 2015. Three-dimensional imaging of lipids and metabolites in tissues by nanospray desorption electrospray ionization mass spectrometry. *Anal. Bioanal. Chem.* 407:2063–71
106. Calligaris D, Caragacianu D, Liu X, Norton I, Thompson CJ, et al. 2014. Application of desorption electrospray ionization mass spectrometry imaging in breast cancer margin analysis. *PNAS* 111:15184–89
107. Giles K, Pringle SD, Worthington KR, Little D, Wildgoose JL, Bateman RH. 2004. Applications of a traveling wave-based radio-frequency-only stacked ring ion guide. *Rapid Commun. Mass Spectrom.* 18:2401–14
108. Baker ES, Clowers BH, Li F, Tang K, Tolmachev AV, et al. 2007. Ion mobility spectrometry—mass spectrometry performance using electrodynamic ion funnels and elevated drift gas pressures. *J. Am. Soc. Mass Spectrom.* 18:1176–87
109. Sleno L. 2012. The use of mass defect in modern mass spectrometry. *J. Mass Spectrom.* 47(2):226–36
110. Kendrick E. 1963. A mass scale based on $CH_2 = 14.0000$ for high resolution mass spectrometry of organic compounds. *Anal. Chem.* 35:2146–54
111. Hughey CA, Hendrickson CL, Rodgers RP, Marshall AG, Qian K. 2001. Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra. *Anal. Chem.* 73:4676–81
112. Marshall AG, Rodgers RP. 2004. Petroleomics: the next grand challenge for chemical analysis. *Acc. Chem. Res.* 37:53–59
113. Lerno LA, German JB, Lebrilla CB. 2010. Method for the identification of lipid classes based on referenced Kendrick mass analysis. *Anal. Chem.* 82:4236–45
114. Senko MW, Beu SC, McLafferty FW. 1995. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* 6:229–33
115. Wu CH, Yeh L-SL, Huang H, Arminski L, Castro-Alvear J, et al. 2003. The Protein Information Resource. *Nucleic Acids Res.* 31:345–47
116. Yao X, Diego P, Ramos AA, Shi Y. 2008. Averagine-scaling analysis and fragment ion mass defect labeling in peptide mass spectrometry. *Anal. Chem.* 80:7383–91
117. Toumi ML, Desaire H. 2010. Improving mass defect filters for human proteins. *J. Proteome Res.* 9:5492–95
118. Nefedov AV, Mitra I, Brasier AR, Sadygov RG. 2011. Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *J. Proteome Res.* 10:4150–57
119. Mitra I, Nefedov AV, Brasier AR, Sadygov RG. 2012. Improved mass defect model for theoretical tryptic peptides. *Anal. Chem.* 84:3026–32
120. Cuyckens F, Hurkmans R, Castro-Perez JM, Leclercq L, Mortishire-Smith RJ. 2009. Extracting metabolite ions out of a matrix background by combined mass defect, neutral loss and isotope filtration. *Rapid Commun. Mass Spectrom.* 23:327–32
121. Zhang H, Zhang D, Ray K, Zhu M. 2009. Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J. Mass Spectrom.* 44:999–1016
122. Zhu M, Ma L, Zhang D, Ray K, Zhao W, et al. 2006. Detection and characterization of metabolites in biological matrices using mass defect filtering of liquid chromatography/high resolution mass spectrometry data. *Drug Metab. Dispos.* 34:1722–33
123. Li X, Brownawell BJ. 2009. Analysis of quaternary ammonium compounds in estuarine sediments by LC–TOF-MS: very high positive mass defects of alkylamine ions as powerful diagnostic tools for identification and structural elucidation. *Anal. Chem.* 81:7926–35
124. Nagy K, Sandoz L, Craft BD, Destaillats F. 2011. Mass-defect filtering of isotope signatures to reveal the source of chlorinated palm oil contaminants. *Food Addit. Contam. A* 28:1492–500
125. Mason EA, McDaniel EW. 1988. *Transport Properties of Ions in Gases*. New York: Wiley
126. Lietz CB, Yu Q, Li L. 2014. Large-scale collision cross-section profiling on a traveling wave ion mobility mass spectrometer. *J. Am. Soc. Mass Spectrom.* 25:2009–19

127. Paglia G, Williams JP, Menikarachchi L, Thompson JW, Tyldesley-Worster R, et al. 2014. Ion mobility derived collision cross sections to support metabolomics applications. *Anal. Chem.* 86:3985–93
128. Wyttenbach T, Pierson NA, Clemmer DE, Bowers MT. 2014. Ion mobility analysis of molecular dynamics. *Annu. Rev. Phys. Chem.* 65:175–96
129. Zhong Y, Hyung S-J, Ruotolo BT. 2012. Ion mobility–mass spectrometry for structural proteomics. *Expert Rev. Proteom.* 9:47–58
130. Lapthorn C, Pullen F, Chowdhry BZ. 2012. Ion mobility spectrometry–mass spectrometry (IMS-MS) of small molecules: separating and assigning structures to ions. *Mass Spectrom. Rev.* 32:43–71
131. Lanucara F, Holman SW, Gray CJ, Eyers CE. 2014. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nat. Chem.* 6:281–94
132. Berant Z, Karpas Z. 1989. Mass-mobility correlation of ions in view of new mobility data. *J. Am. Chem. Soc.* 111:3819–24
133. Fenn LS, Kliman M, Mahsut A, Zhao SR, McLean JA. 2009. Characterizing ion mobility-mass spectrometry conformation space for the analysis of complex biological samples. *Anal. Bioanal. Chem.* 394:235–44
134. Harvey D, Crispin M, Bonomelli C, Scrivens J. 2015. Ion mobility mass spectrometry for ion recovery and clean-up of MS and MS/MS spectra obtained from low abundance viral samples. *J. Am. Soc. Mass Spectrom.* 26:1754–67
135. Li H, Bendiak B, Siems W, Gang D, Hill H Jr. 2013. Ion mobility-mass correlation trend line separation of glycoprotein digests without deglycosylation. *Int. J. Ion Mobil. Spectrom.* 16:105–15
136. May JC, Goodwin CR, Lareau NM, Leaptrot KL, Morris CB, et al. 2014. Conformational ordering of biomolecules in the gas phase: nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer. *Anal. Chem.* 86:2107–16
137. May JC, McLean JA. 2014. Lipid map. *Anal. Sci.* 0814. **https://theanalyticalscientist.com/issues/0814/data-visualization-infographics/**
138. Ceglarek U, Leichtle A, Brügel M, Kortz L, Brauer R, et al. 2009. Challenges and developments in tandem mass spectrometry based clinical metabolomics. *Mol. Cell. Endocrinol.* 301:266–71
139. Benton HP, Ivanisevic J, Mahieu NG, Kurczy ME, Johnson CH, et al. 2015. Autonomous metabolomics for rapid metabolite identification in global profiling. *Anal. Chem.* 87:884–91
140. Patti GJ, Yanes O, Siuzdak G. 2012. Innovation: metabolomics: the apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* 13:263–69
141. Johnson CH, Ivanisevic J, Benton HP, Siuzdak G. 2015. Bioinformatics: the next frontier of metabolomics. *Anal. Chem.* 87:147–56
142. Patti GJ, Tautenhahn R, Rinehart D, Cho K, Shriver LP, et al. 2013. A view from above: cloud plots to visualize global metabolomic data. *Anal. Chem.* 85:798–804
143. Gowda H, Ivanisevic J, Johnson CH, Kurczy ME, Benton HP, et al. 2014. Interactive XCMS online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal. Chem.* 86:6931–39
144. Goodacre R, Neal MJ, Kell DB. 1996. Quantitative analysis of multivariate data using artificial neural networks: a tutorial review and applications to the deconvolution of pyrolysis mass spectra. *Zentralbl. Bakteriol.* 284:516–39
145. Franceschi P, Wehrens R. 2014. Self-organizing maps: a versatile tool for the automatic analysis of untargeted imaging datasets. *Proteomics* 14:853–61
146. Patterson AD, Li H, Eichler GS, Krausz KW, Weinstein JN, et al. 2008. UPLC-ESI-TOFMS-based metabolomics and gene expression dynamics inspector self-organizing metabolomic maps as tools for understanding the cellular response to ionizing radiation. *Anal. Chem.* 80:665–74
147. Goodwin CR, Sherrod SD, Marasco CC, Bachmann BO, Schramm-Sapyta N, et al. 2014. Phenotypic mapping of metabolic profiles using self-organizing maps of high-dimensional mass spectrometry data. *Anal. Chem.* 86:6563–71
148. Goodwin CR, Covington BC, Derewacz DK, McNees CR, Wikswo JP, et al. 2015. Structuring microbial metabolic responses to multiplexed stimuli via self-organizing metabolomics maps. *Chem. Biol.* 22:661–70
149. Sherrod SD, McLean JA. 2016. Systems-wide high dimensional data acquisition and informatics using structural mass spectrometry strategies. *Clin. Chem.* 62:77–83