Accepted Manuscript

Title: Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects

Author: Maria Vinaixa, Emma L. Schymanski, Steffen Neumann, Miriam Navarro, Reza M Salek, Oscar Yanes

PII: S0165-9936(15)30083-2
DOI: http://dx.doi.org/doi:10.1016/j.trac.2015.09.005
Reference: TRAC 14511

To appear in: *Trends in Analytical Chemistry*

Please cite this article as: Maria Vinaixa, Emma L. Schymanski, Steffen Neumann, Miriam Navarro, Reza M Salek, Oscar Yanes, Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects, *Trends in Analytical Chemistry* (2015), http://dx.doi.org/doi:10.1016/j.trac.2015.09.005.

Mass spectral databases for LC/MS and GC/MS-based metabolomics: state of the field and future prospects

Maria Vinaixa[1,2,3], Emma L. Schymanski[4], Steffen Neumann[5], Miriam Navarro[1,2,3], Reza M Salek[6], Oscar Yanes[1,2,3*]

1. Centre for Omic Sciences, Universitat Rovira i Virgili, Avinguda Universitat 3, 43204 Reus, Spain.
2. Department of Electronic Engineering, Universitat Rovira i Virgili, Avinguda Paisos Catalans 26, 43007 Tarragona, Spain.
3. Metabolomics Platform, Spanish Biomedical Research Center in Diabetes and Associated Metabolic Disorders (CIBERDEM), Monforte de Lemos 3-5, 28029 Madrid, Spain.
4. Eawag: Swiss Federal Institute for Aquatic Science and Technology, Überlandstrasse 133, 8600 Dübendorf, Switzerland.
5. Leibniz Institute of Plant Biochemistry, Dept. of Stress and Developmental Biology, Weinberg 3, 06120 Halle, Germany.
6. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

* to whom correspondence should be addressed

Oscar Yanes, PhD.
Centre for Omic Sciences
Rovira i Virgili University
Avinguda Universitat, 3.
43204 Reus (Spain)
phone: +34 977776617
email: oscar.yanes@urv.cat

Highlights

- Mass spectral databases play a key role in metabolomics.
- We underline the advantages and limitations of public and commercial databases.
- The overlap of compounds in public and commercial databases is calculated.
- Future prospects for mass spectral databases are discussed.

Abstract

Mass spectrometry-based metabolomics is now widely used to obtain new insights into human, plant and microbial biochemistry, drug and biomarker discovery, nutrition research and food control. Despite this great shared interest, identifying and characterizing the structure of metabolites has become a major bottleneck for converting raw mass spectrometric data into biological knowledge. In this regard, comprehensive and well-annotated MS-based spectral databases play a key role towards converting raw spectral data into metabolite annotations and thus biological knowledge. The main characteristics of the mass spectral databases currently used in MS-based metabolomics, are reviewed in this paper, underlining the advantages and limitations of each. Extending this, the overlap of compounds with MS^n ($n \geq 2$) spectra from authentic chemical standards in most public and commercial databases has been calculated for the first time. Finally, future prospects for mass spectral databases are discussed in terms of the needs posed by novel applications and instrumental advancements.

Keywords: metabolomics, databases, mass spectrometry, liquid chromatography, gas chromatography, identification, mass spectral databases.

1. Introduction

Metabolomics is called upon to complement upstream biochemical information obtained from genes, transcripts and proteins, widening current genomic reconstructions of metabolism and improving our understanding of cell biology, physiology and medicine by linking cellular pathways to biological mechanism [1]. Two technological platforms are most commonly used to identify and quantify metabolites to achieve this: nuclear magnetic resonance (NMR) spectroscopy, and mass spectrometry (MS), often coupled to chromatographic techniques.

Unlike genes, transcripts or proteins, which are biopolymers encoding information as a sequence of well-defined monomers, namely nucleotides and amino acids, metabolites are chemical entities that do not result from a residue-by-residue transfer of information within

the cell. Instead, the extremely large diversity of metabolite structures in living organisms results directly from series of chemical transformations catalyzed mainly by enzymes on environmental or dietary resources. Identifying and characterizing the structure of metabolites has become one of the major bottleneck for converting raw spectrometric data into biological knowledge, preventing metabolomics from evolving as fast as the other "omic" sciences [2–4].

Identification of metabolites is still evolving within the community, with active discussion on how to define what constitutes a valid metabolite identification [5]. The Metabolomics Society, for instance, is currently assessing and developing an improved set of reporting standards [5–7]. The Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI; http://msi-workgroups.sourceforge.net) established so far four levels of identification[8]. Level 1 identification requires that at least two orthogonal molecular properties of the putative metabolite be confirmed with an authentic pure compound analyzed under identical analytical conditions. By contrast, for level 2 and 3, the comparison against literature and database data is sufficient, and therefore rather than identifications, only annotations are achieved. Level 4 refers to unknown compounds. Following this classification, the majority of metabolites reported in metabolomics literature correspond to the annotation of known chemical structures described in databases. Most of these annotations are based on physico-chemical properties (e.g., chromatographic retention time) and/or spectral similarity with public/commercial spectral databases [1]. This demonstrates the importance of having comprehensive and well-annotated MS-based spectral databases [9]. Conversely, relatively few studies comply with the standards of level 1 reporting [1,10–12].

Proper usage and development of MS-based spectral databases, therefore, is essential if metabolomics wants to reach the status of the other omic sciences [9]. Unfortunately, current databases are still far from containing experimental data from all known metabolites, despite efforts to increase and improve their content, one such example is the Metabolite Standards Synthesis Core (MSSC) initiative by NIH, aiming to generate new compound standards (http://www.metabolomicsworkbench.org/standards/nominatecompounds.php). The major limitation is the relatively small number of metabolites commercially available as pure standards, not to mention the large number of metabolites with unknown chemical structures that remain to be identified and characterized [12]. Additionally, the transferability of mass spectral databases, particularly MS/MS, between MS instruments can impose some limitations, restricting the structural assignments of metabolites by empirically matching spectral values from pure standard compounds. Despite these limitations, the use of reference spectral databases is still one of the best approaches to annotate the structure of known metabolites when full isolation and structure determination by NMR or X-ray

crystallography is not possible. Alternatively, novel computational tools that heuristically predict MS fragmentation patterns *in silico* have been developed to assist with identification of metabolites for which tandem MS data is not available yet in databases [13–19]. For electron ionization mass spectrometry (EI-MS), it has been shown that fragmentation spectra can be simulated with quantum chemical and molecular dynamics methods [20], although the runtime is still too large (several thousand CPU hours per molecule have been reported) to simulate spectra for many compounds.

Freely accessible and/or commercially available compound databases currently used in the field of metabolomics provide information on chemical structures, physico-chemical properties, spectral profiles, biological functions, and pathway mapping of metabolites. On the basis of these annotations Fiehn and colleagues [21] classify these databases within two categories: (i) pathway-centric databases such as KEGG [22], Reactome [23], Wikipathways [24] or Biocyc [25]; and (ii) compound-centric databases such as PubChem [26], ChemSpider [27], METLIN [28], MassBank [29], GMD [30] or HMDB [31]. While PubChem, ChemSpider and Chemical Abstracts Service (CAS) provide >60 million, >35 million and >100 million chemical compounds respectively, these databases are not typically used in metabolomics due to limited biological relevance of the vast majority of chemicals and the lack of mass spectral information. In contrast, some other compound-centric databases are also enriched with mass spectral information, which enables annotation of metabolites by matching mass spectral features of the unknown compounds to curated spectra of reference standards. Although these are much smaller repositories compared to PubChem or ChemSpider, mass spectral databases represent a first step towards converting raw spectral data into metabolite annotations and thus biological knowledge.

This review article is intended to provide an overview of the state of the art on mass spectral databases most commonly used for metabolite annotation in metabolomics. Mass spectral databases and search engines aiming to assist identification of metabolites through spectral matching have been around for several decades for EI-MS, with other ionization methods catching up over the past decade. In general, chromatographic and ionization techniques still determine the identification workflow and the most appropriate metabolite mass spectral reference database to be used. In subsequent sections we will describe the latest versions of the most widely used mass spectral databases for LC/MS and GC/MS based metabolomics. We focus on the strengths and weaknesses that these databases provide for the annotation of metabolites.

2. Overview of the LC/MS-based untargeted metabolomics workflow

LC/MS-based untargeted metabolomics typically involves comparing the relative abundances of metabolites in multiple samples without prior identification. By using liquid

chromatography coupled to quadrupole time-of-flight [28] or Orbitrap-based mass spectrometers [32], hundreds to thousands of peaks with a unique *m/z* value and retention time (*m/z*-RT pair) are routinely detected from biological samples in a profiling experiment. Each peak, termed a metabolite feature, is integrated over the LC time scale and compared between runs using computational tools such as the freely accessible XCMS [33], MetAlign [34] and MZmine packages [35,36], or vendor software such as MassHunter and/or Agilent Mass Profiler Professional (MPP, Agilent Technologies), Sieve™ (Thermo Scientific), DataAnalysis (Bruker Daltonics) or MarkerLynx™ (Waters Corporation). After selecting interesting features according to statistical criteria [37] these features can be characterized on the basis of their mass spectral information (accurate mass, isotopic pattern and fragmentation pattern) and retention time.

The exact mass (*i.e.*, mass-to-charge ratio, *m/z*) of a selected feature is often used to query compound-centric databases. Given the great redundancy of features that correspond to the same metabolite due to naturally occurring isotopes, adduct formation and in-source fragmentation, grouping and annotating these features should be considered to determine the neutral exact mass or even molecular formula of the underlying molecule. This helps reduce both the number of precursor ion queries in databases as well as the false annotations of adduct, isotope or fragment peaks. Open-source algorithms such as CAMERA [38], AStream [39], nontarget [40] or PUTMEDID [41] and commercially available software such as those mentioned above enable significant data reduction by annotating LC/MS-based peaks. However, as metabolomics data can be highly complex, due to co-elution in the LC, differences in the ion source parameters and mass resolving power of various instruments, misannotations can still prevent the correct identification of a large number of metabolites. If features are considered individually, the presence of adducts and isotopes should not be neglected and many database searches allow the specification of expected adducts when performing queries.

Major metabolite databases that support batch precursor ion-based searching are the Human Metabolome Database (HMDB) [31] and METLIN (more details below). Database hits provide only putative assignments that must be further validated by retention time matching and/or MS/MS analysis. In the absence of a pure standard analyzed under identical analytical conditions, MS/MS data searched against a reference MS/MS databases is typically the most conclusive evidence for validating and putatively annotating a metabolite feature using mass spectrometry.

Next, we describe the most commonly used mass spectral databases for LC/MS-based untargeted metabolomics. While we focus on LC/MS, it should be noted that all relevant information can be extrapolated to capillary electrophoresis coupled to mass spectrometry (CE-MS) [42–44]. As LC/MS and CE/MS approaches typically use the same ionization

source (*i.e.*, ESI), mass analyzer (*i.e.*, TOF, Orbitrap) and fragmentation technique (*i.e.*, CID, HCD), the MS/MS spectral information and databases apply to both approaches.

3. Mass spectral databases for LC/MS-based untargeted metabolomics

3.1. Human Metabolome Database (HMDB)

The HMDB version 3.6 (http://www.hmdb.ca/) is a comprehensive web-accessible electronic database containing detailed information on metabolites found in the human body. First introduced in 2007 [45], HMDB is currently the world's largest and most comprehensive organism-specific metabolomics database as of August 2015, accounting for nearly 42,000 metabolite entries [31]. HMDB brings together quantitative chemical, physical, clinical and biological data about all experimentally detected and biologically 'expected' human metabolites. 'Expected' metabolites refer to those that potentially can be detected in human samples, e.g. biofluids, because their biochemical pathways are known or human intake/exposure is frequent but actually remain yet to be experimentally verified. These include dipeptides, drugs, drug metabolites and food-derived compounds. Although there is a separate dedicated resource on food constituents called FoodDB (http://foodb.ca/), its mass spectral information is duplicated in the HMDB.

For each HMDB entry, the information is organized in the so-called MetaboCard layout format. This format includes different fields that contain clinical, chemical, spectral, biochemical and enzymatic data. Some of these fields are hyperlinked to other databases such as KEGG, PubChem, ChemSpider, MetaCyc, ChEBI [46], PDB, Swiss-Prot [47], and GenBank [48,49] and a variety of structure and pathway viewing applets [50]. Importantly, MetaboCards are downloadable in an "Extensible markup language" (XML) format, which can be easily parsed or imported into relational databases that provide advanced querying capabilities.

Mass spectral resources within HMDB are important database features that have significantly improved in the latest version (3.6) of the resource. For example, the number of authentic chemical standards containing mass spectral information has increased from ~3,000 (version 2.5) to ~9,500. Besides GC/MS spectral data, HMDB 3.6 contains ~3,500 metabolites with electrospray ionization (ESI) MS/MS spectral information, which correspond to ~8% of all entries. Of the 5,912 MS/MS spectra in HMDB, 3,453 have been imported from MassBank, in addition to those acquired using local instruments. MS/MS spectra in HMDB can be either low and/or high-resolution. The HMDB allows querying up to 150 precursor ions in batch mode within seconds ('MS Search' tool), and matching experimental MS/MS

data against its own mass spectral database ('MS/MS Search' tool) based on collision-induced dissociation spectra at low, medium and high energy. The MS and MS/MS search functions report a list of matches listed by mass error (e.g., delta ppm) or spectral similarity, respectively. HMDB 3.6 has also incorporated a MS/MS spectrum viewing tool with fragment peak assignments similar to those found in the METLIN database, this latter being based on MetFrag [51] predictions. All mass spectral data can be downloaded in mzML [52], a well-established and accepted open source format for MS data exchange gathering both spectral features and acquisition conditions.

On the downside, the lack of systematization regarding collision energies and instrument types in the MS/MS spectral database and that HMDB only allows an MS/MS search with a maximum precursor mass deviation of 20 Da are perhaps the main drawback of the HMDB. Lifting this restriction would allow to search for related metabolites (e.g. glycosides) using their spectral similarity, and to draw conclusions on the unknown.

3.2. METLIN

METLIN is a metabolite repository that was launched in 2004 as a web and freely accessible electronic database (http://metlin.scripps.edu/) to assist with metabolite research and to facilitate metabolite annotation through mass spectrometry analysis [53,54]. As of August 2015, the METLIN repository contains over 240,000 compounds that include endogenous metabolites from different organisms (e.g., plant, bacteria, and human) and exogenous compounds such as pharmaceutical drugs and other synthetic organic substances. Contrary to the HMDB, peptides consisting of three and four amino acids are also included in the METLIN database.

Importantly METLIN has systematically generated high-resolution MS/MS data of over 13,000 distinct authentic chemical standards at three different collision energies (10V, 20V and 40V), which correspond to ~5% of the total repository data. These data were collected using an ESI-quadrupole time-of-flight (Q-TOF) mass spectrometer from Agilent Technologies in both positive and negative detection modes, representing a total number of >68,000 high-resolution MS/MS spectra. This makes METLIN one of the largest spectral resources for MS-based metabolomics.

The METLIN web-based interface allows single and batch searching of up to 500 precursor ions (m/z) in a matter of seconds. The batch search reports a list of matches listed by mass error (e.g., delta ppm). Queries to the database can be performed through mass-to-charge data allowing for the selection of numerous adduct types, but also chemical formula, chemical names, CAS number and KEGG identifier are accepted (using Advanced search option). The results of the queries to the database can be downloaded as a CSV file, which contains the returned hits with information about potential adducts, mass error in parts per

million (ppm), metabolite names, molecular formula, CAS number, KEGG identifier and, importantly, whether the MS/MS spectra are available or not.

As of August 2015, the METLIN database includes spectral matching functionalities that allow investigators to upload their MS/MS data to automatically compare to MS/MS data stored in METLIN. The MS/MS search reports a list of matches scored by spectral similarity. Remarkably, the mass spectra in METLIN match quite well with those spectra generated from different MS instruments, among which are the AB SCIEX Triple TOF 5600, Agilent 6460 Triple Quad, Thermo Scientific Q-Exactive and Thermo Scientific LTQ Orbitrap Velos, either using collision-induced dissociation (CID) or higher-energy collisional dissociation (HCD) [47],[48]. Yet, it is important to clarify that the correlation of METLIN MS/MS data to MS/MS data acquired by using different instrument platforms was performed using only 23 metabolite standards and spectral matching were based on the X-Rank scoring system [55]. Identification of metabolites using MS/MS data can also be performed manually by accessing the 'fragment search' or 'neutral loss search' options. These tools provide a framework by which unknown metabolites can be chemically classified on the basis of characteristic fragments or neutral losses that are used as signatures for unique chemical functional groups (*e.g.,* fragments at 85.0289 m/z and 60.0813 m/z for acyl-carnitines).

The main disadvantage of the METLIN database is that none of the data is available for download, which prevents researchers from using the large amount of MS/MS data for developing novel tools and applications. Although laboratories with Agilent instruments can purchase METLIN with their vendor software, this is not updated as frequently as the website. The Agilent METLIN personal metabolite database (PMD) contains 9,083 unique compounds, of which 4,165 are metabolites and the rest are di- and tripeptides. The METLIN website currently has ~13,000 compounds with MS/MS spectra with a similar contribution of di- and tripeptides. Unfortunately, many adduct ions for common metabolites are missing. The application programming interfaces (APIs) for METLIN online have been disabled since 2011.

3.3. MassBank

MassBank is an open-community mass spectra repository designed for public sharing of reference mass spectra from authentic chemical standards for metabolite annotation (http://www.massbank.jp). MassBank data is contributed by consortium members mainly from Japan, but also from the EU, Switzerland, USA, Brazil and China [29].

MassBank contains spectral information and acquisition conditions of mass spectra derived from different MS setups, including different ionization techniques such as ESI (60% of the total dataset), EI (31%), CI (2%), APCI (1.6%) and MALDI (<1.5%), and high-resolution (qTOF, Orbitrap) and low-resolution (QqQ, ion-trap) mass analyzers from different

vendors. As of August 2015, MassBank contained nearly 19,000 MS1 (mainly EI) and 28,000 MS/MS spectra (mainly ESI-qTOF and ESI-ITFT), including >32,000 spectra in positive ionization mode and >10,000 in negative mode, thus ~43,000 spectra in total.

One distinctive feature of MassBank is the use of the so-called 'merged spectra', an artificial reference that accounts for ~2% of the total number of spectra within the database. Merged spectra were created by summarizing all different CID spectra for the same compound into a single spectrum [29]. This is aimed to make the identification of metabolites more independent of the instrument configuration and MS manufacturer.

Each entry is described with a compound name, chemical structure and experimental conditions, including the MS platform, chromatographic methods, retention time, precursor ion, high-resolution MS data, and links to other databases. Different software options are available for users who want to contribute MassBank records, including the Excel-based Record Editor [56] for manual record creation or Mass++ and RMassBank [57] in C++ and R, respectively for batch workflows. Software is available to set up a local MassBank server for either Windows or Linux. Further, the MassBank database is downloadable if the records are given a Creative Commons license [56], and users are able to write their own programs for accessing, customizing and utilizing it. MassBank allows access to its library by web API, and the code and supporting tools are hosted on GitHub (https://github.com/MassBank).

In addition, MassBank web browser functionalities are accessible using Java applet. The 'Spectrum Search' applet allows users to conduct MS^n spectra similarity search by providing experimental MS/MS spectra in single or batch mode ('Batch Service' function) as a list of *m/z* values and intensities, ionization mode and instrument type. The query spectrum is retrieved and displayed as a list showing the similarity score and number of identical product ions. It is also possible to graphically compare the query spectrum with the retrieved spectra. The 'Quick Search' feature allows searching for compounds by compound name, exact mass, molecular formula and chemical structure. Unfortunately, neither 'Quick Search' nor 'Peak Search' applets allow searches using *m/z* of precursor ions, while the batch search feature can only be performed by email. 'Peak Search Advanced' enables finding peaks by specifying an ion or neutral loss by molecular formulas instead of a numerical *m/z* value. Additional details can be found in the extensive massbank manual: http://www.massbank.jp/manuals/UserManual_en.pdf

Perhaps the main disadvantage of the MassBank database is that not all of the records are sufficiently curated and some entries are poorly or misannotated, while others contain noisy or poorly extracted spectra. Recent efforts in developing RMassBank have focused on creating cleaned-up and well annotated spectra for upload to MassBank starting from mzML files of standard substances [57], using web services such as the Chemical Translation

Service (CTS) [58]. Issues with the Java applets on some computer installations have also reduced the usability of MassBank in recent times.

Recently, NORMAN MassBank (http://massbank.eu/MassBank/) and MoNA (http://mona.fiehnlab.ucdavis.edu/) have been introduced as the European (2012) and North American (2015) MassBank servers, respectively.

3.4. LIPID MAPS & LipidBlast

The LIPID MAPS Structure Database (LMSD, http://www.lipidmaps.org/data/structure/) is a database encompassing structures and annotations of biologically relevant lipids. As of August 2015, the LMSD contained >40,000 unique lipid structures, making it the largest publically available lipid-only database in the world. The LMSD classification system comprises of eight lipid categories, each with its own subclassification hierarchy. All lipids in the LMSD have been assigned a LIPID MAPS ID. LMSD can be searched by lipid class, common name, systematic name or synonym, mass, InChIKey (*i.e.*, the hashed version of the standard International Chemical Identifier string [59]) or LIPID MAPS ID using the 'Quick Search' tool. Alternatively, queries using LIPID MAPS ID, systematic or common name, mass, formula, category, main class, subclass data, structure or substructure are accepted using one of the search interfaces in the LMSD database section. Each LMSD record contains an image of the molecular structure, common and systematic names, links to external databases, Wikipedia pages (where available), other annotations and links to structure viewing tools (e.g., gif image, ChemDraw, MarvinView and JMol). In addition, LMSD contains spectral information about authentic chemical standards for several lipid categories. These include retention time data, MS/MS spectra with annotations for principal product ions, and LC/MS/MS protocols, including experimental parameters used for acquisition of some (not all) MS/MS spectra (http://www.lipidmaps.org/data/standards/search.html). Most lipid standards are provided by Avanti® Polar Lipids, Inc. and to a lesser extent by Cayman Chemical. These include 197 fatty acyl, 21 glycerolipids, 242 glycerophospholipids, 18 sphingolipids, 38 sterol lipids, 9 prenol and 4 saccharolipid standards. Unfortunately, MS/MS spectra are only available as images, which hampers the ability to parse the spectral data.

LMSD contains several MS analysis tools to *in silico* predict possible lipid structures from precursor and/or product-ion MS experimental data [60]. Unfortunately, the precursor ion search is constrained by the selection of lipid families (glycerophospholipids, cholesteryl esters, glycerolipids, sphingolipids, cardiolipins or fatty acids), and only a single precursor ion can be queried at a time. Product-ion search allows matching a MS/MS peak list to a virtual database of "high probability" product ions, mostly side-chains and headgroups typically found in mammalian versions of glycerolipids, glycerophospholipids (including

cardiolipins) and sphingolipids [61]. The main disadvantage is that the MS/MS spectra of different lipid families are only predicted in negative or positive ionization mode, but not both. In addition, spectra are only available for one adduct per lipid (e.g., $[M+NH_4]^+$ for glycerolipids). Recently, LMSD supports free download of all the structures and annotations in the database.

A freely available computer generated (*in silico*) MS/MS database for lipid annotation called LipidBlast [62] was developed by Fiehn lab and released in 2013. LipidBlast contains 212,516 MS/MS spectra covering 119,200 compounds from 26 lipid compound classes. More than half of all LipidBlast compound structures were imported from the LMSD or generated using the LIPID MAPS Tools, covering 13 lipid classes of the most common glycerophospholipids and glycerolipids. Many bacterial and plant lipids are not covered by LMSD and were computer-generated in LipidBlast. Importantly, LipidBlast works with low- and high-resolution instruments covering 78,314 positive and 134,202 negative mode ionization *in silico* MS/MS spectra. Several adducts are also covered in LipidBlast including $[M+H]^+$; $[M+Na]^+$; $[M+NH_4]^+$; $[M-H]^-$; $[M–2H]^{(2-)}$; $[M+NH_4–CO]^+$; $[M+2Na–H]^+$; $[M]^+$; $[M–H+Na]^+$; $[M+Li]^+$.

Since the heuristic modeling of LipidBlast was mostly developed on the basis of ion trap MS/MS spectra using CID, the main limitation is that the *in silico* predictions of the database suffer from the “one third rule”, by which fragment masses below 28% of the precursor mass are not detected. Consequently, *in silico* MS/MS spectra are missing certain classes of potentially interesting and useful ions (e.g., fragment m/z 184.07 in glycerophospholipids).

3.5. NIST 14

Historically developed as an EI-MS database (see details below), the latest commercially available released version of the spectral database of the US National Institute of Science and Technology (NIST 2014) also contains 234,284 ESI MS/MS spectra of small molecules including authentic chemical standards of metabolites, lipids, biologically active peptides and all di-peptides and tryptic tri-peptides [63]. Specifically, 51,216 ion trap spectra from 8,171 compounds, and 183,068 CID spectra (q-TOF and triple quad) from 7,692 compounds are included (Table 1). Of note, many of the precursor ions for which there is MS/MS data correspond to common adducts formed during ESI (in addition to $[M+H]^+$ in positive ion mode and $[M-H]^-$ in negative mode), including $[M+H-H_2O]^+$, $[M+Na]^+$, $[M+NH_4]^+$, $[M+H-NH_3]^+$, $[2M+H]^+$, $[M-H-H_2O]^-$, $[2M-H]^-$, $[M-2H]^{2-}$ (Figure 1). Few other spectral databases offer a comparable number of MS/MS spectra from adducts. This is particularly useful because predominant adducts in the ESI spectrum vary from one metabolite to another as well as on the mobile phase used. One example is glucose-6-phosphate, which ionizes predominantly

as $[M+Na]^+$ or $[M+NH_4]^+$ using formic acid (0,1%) and ammonium enriched mobile phases, respectively (Figure 2A). Each precursor adduct, in turn, results in different MS/MS spectra (Figure 2B).

The small molecules section of the NIST 14 MS/MS database was acquired using different high- and low resolution instruments, over a range of energies, and in both collision cells (beam type) and ion traps (up to MS^3) in both positive- and negative ion mode, when appropriate. The instruments include; Thermo Scientific Orbitrap Elite, Thermo Scientific Orbitrap Velos, Agilent QTOF 6530, and Waters TQD QqQ, Micromass Quattro Micro QqQ, and Thermo Finnigan LTQ IT/ion trap [64].

The NIST 14 database is formatted in binary format suitable for stand-alone use or by the NIST MS Search software. The NIST MS Search software (version 2.2) can be used for searching (identifying) compounds from their experimental mass spectra and for browsing mass spectral libraries. Mass spectra are sorted by charge, molecular weight, formula, compound name and acquisition parameters. It also includes MS interpretation programs for analyzing mass spectra on the basis of chemical structure, molecular formula and isotopic patterns. Additional instrument-specific formats (e.g. Agilent ChemStation and MassHunter™) are available separately to permit library searching directly within other GC/MS or LC/MS data systems. NIST spectra can be exported in the text-based *.msp format (generated by the NIST search program), which can be read by most vendor software or parsed into workflows. The Lib2NIST converter also allows users to convert and add their spectra into the NIST database. As a result, some groups are now publishing small, openly-accessible libraries in the NIST format (http://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start) for others to use with the NIST MS Search software, which is available for free, unlike the NIST database itself.

Overall, the NIST 14 has become a well curated and comprehensive mass spectral database that should be considered for LC/MS-based metabolomics. Although the InChI chemical identifier has been added to each compound in NIST 14, unfortunately metabolites do not yet contain links to other metabolomic databases (e.g., KEGG, HMDB identifiers) to allow further biological characterization. However, the NIST database is connected in some way to most vendor software and is also available offline rather than via a web service.

3.6. mzCloud

Hosted by HighChem LLC, Slovakia, mzCloud encompasses an open community of academic and industrial partners who provide MS/MS and MS^n spectral trees that can be freely searched through an intuitive web-based interface (https://www.mzcloud.org/). Currently, mzCloud contains highly curated mass spectral information on 2,987 authentic chemical standards measured on Orbitrap instruments. These include raw and calibrated

MS spectra displaying annotated observed adducts (e.g., $[M+H]^+$, $[M-H_2O+H]^+$, $[M+Na]^+$), and MS/MS and MS^n spectra displayed at various collision energies using collision-induced dissociation (CID) and higher-energy collisional dissociation (HCD). Spectral peaks are structurally annotated using one or more of *in silico* prediction methods based on Mass Frontier™ 7.0 software, manual assignment by a mass spectrometrist and quantum chemical methods. mzCloud displays supplementary information and data sources along with experimental parameters and a collection of compound identifiers.

Importantly, mzCloud uses proprietary computational tools to provide substructural information about metabolites that are not present in the mzCloud database through the comparison of product ion spectra of structurally related compounds. Structural information is derived by utilizing previously characterized ion structures stored in mzCloud and matching them with unknown product ion spectra.

On the downside, the number of naturally occurring metabolites is still low in comparison with the other databases described above. Neither the MS/MS and MS^n spectra are available for download, nor a batch service is offered. Finally, the mzCloud.org web-based interface is based on Microsoft Silverlight, which gives rise to compatibility issues with some browsers as well as Mac and Linux systems. For Linux users, Pipelight (http://pipelight.net/cms/installation.html) executes the Silverlight runtime under the Wine emulator (https://www.winehq.org/) inside a Firefox plugin.

4. Overview of the GC/MS-based untargeted metabolomics workflow

GC/MS has been a long-standing approach used for metabolite profiling due to the widespread use of electron impact ionization (EI). EI is a hard ionization technique that has been historically standardized at 70 eV. Unlike soft ionization techniques such as ESI, EI is a highly reproducible ionization process across many different platforms. Hence, GC/MS experimental spectra collected in different labs can be easily compared to recorded EI database samples for metabolite annotation. However, spectral similarity between different substances, errors in the process of signal deconvolution as well as sources of analytical variability such as chemical derivatization of the samples or column aging/degradation, making it challenging for metabolites to be unequivocally annotated solely through EI spectral matching to a database of reference compounds [65]. Since Kovats introduced retention indices (RI) [66] based on aliphatic carbon numbers that ultimately led to the concept of mass spectral tags (MST), RI has been progressively included in mass spectral databases commonly used for GC/MS metabolomics [67,68]. In combination with EI spectral matching, RI greatly improves metabolite annotation. For example, the combination of AMDIS [69] and the NIST database allows users to perform deconvolution, RI calculation and spectral database searching of a whole sample in one step.

The common workflow for GC/MS metabolite profiling typically implies working with deconvoluted mass spectra and identified compounds instead of unidentified peaks, as previously described for LC/MS datasets, although an unbiased approach where the metabolites are not known before data analysis is also applied by many groups. In addition, when a mass spectrum does not match to any reference standard in the databases, the MST concept allows handling and referencing of yet unidentified metabolic components. MST detected across profiling studies can be systematically indexed and archived in public databases, which may facilitate later identification using pure authentic reference substances. Currently, databases such as GMD [30,70] or BinBase/FiehnLib [67] keep systematic MST annotations.

5. Mass spectral databases for GC/MS-based untargeted metabolomics

The databases commonly used in GC/MS-based metabolomics are the mass spectral database of the US National Institute of Science and Technology (NIST, http://www.nist.gov/srd/nist1a.cfm), the Golm Metabolome Database (GMD, http://gmd.mpimp-golm.mpg.de) and the Fiehn/Binbase library (http://fiehnlab.ucdavis.edu/db), which contain mainly TMS-derivatized metabolites. In addition, HMDB, MassBank and the Madison Metabolomics Consortium Database (MMCD) also contain EI spectra and retention index data. However, most of these data overlap with that contained in the NIST and GMD and will not be discussed in this review.

5.1. NIST 14

The NIST electron ionization (EI) mass spectral database is one of the most widely used and comprehensive mass spectral reference libraries. The latest commercially available released version is NIST 14, the successor of NIST 11. It consists of a fully evaluated and manually curated collection of 276,259 EI mass spectra from 242,477 unique compounds. The focus of NIST 14 has been mainly on adding authentic chemical standards of plant and human metabolites, drugs, and compounds of industrial and environmental importance, which indicates its strategic positioning in the metabolomics scene. Emphasis has been also put on spectra for a wider range of derivatization methods. In addition, NIST 14 includes 387,463 measured Kovats or Lee RI and corresponding GC methods, column conditions and literature citations for 82,868 compounds (>67% of these compounds have spectra in the EI database). For most compounds without measured RI data, predicted values are given[71], albeit with high error margins. Besides mass spectra, typical data include name, formula, molecular structure, molecular weight, CAS number, contributor name, list of peaks, synonyms, and measured retention index when available.

The NIST 14 EI database is formatted for stand-alone use using the NIST MS Search software, a platform-independent program that allows for various types of comparisons of an acquired unknown spectrum with the NIST database. Additionally, NIST 14 can also be queried via most vendor-specific developed software due to its established nature in the industry, such as Thermo Xcalibur® quantitation and MassHunter software. The NIST MS Search offers spectral interpretation features including molecular mass estimation and element count as well as substructure presence/absence [72].

Finally, NIST 14 database comes with an updated version of the AMDIS software (Automated Mass Spectral Deconvolution and Identification System) [69]. AMDIS has been designed by NIST to deconvolute and reconstruct "pure component" spectra from complex GC/MS total ion current chromatograms. Deconvoluted mass spectra are used for compound identification via NIST database matching, while substructure interpretation is also available (complementary to the MS Search) using the classifiers of Varmuza et al. [73,74]. Although AMDIS is freely available, the code is not. Importantly, AMDIS reads data files from most MS manufacturers and open data format such as mzXML/mzData and NetCDF.

5.2. The Golm Metabolome Database

The Golm Metabolome Database (GMD, http://gmd.mpimp-golm.mpg.de/) is a publicly available GC-EI mass spectral database developed and hosted by the Max Planck Institute for Molecular Plant Physiology (Golm, Germany) [30]. The spectra in the GMD are contributed by several consortium members and comprises GC/MS libraries obtained using either quadrupole or TOF mass detectors. All libraries can be freely downloaded and used. The GMD harbors information on 1,461 unique chemical standards. One metabolite typically has a set of different associated analytes due to different silylation products for instance, which leads to >26,500 spectra in total, >11,600 spectra linked to analytes, and >9,150 spectra linked to analytes tagged with an RI. Overall, GMD now contains >3,500 analytes and >2,000 metabolites associated with valid spectra. This information is organized in metabolite report cards (similar to the MetaboCard layout format in the HMDB database). It also includes information on >3,100 mass spectral tags, namely repeatedly observed mass ions with retention time behavior for which the compound is not yet identified.

All these data are used by the GMD to allow compound annotation by matching to GC/MS reference spectra and retention indices. The GMD permits mass spectra-based queries thereby facilitating the annotation process of user-obtained GC/MS spectra. The queries can be performed with or without RI constraints, both for GC and GCxGC columns. GMD also allows users to generate customized libraries as subsets of the existing GMD. Interestingly, GMD applies several scoring methods based on distance functions for spectra

comparison, however it is unclear which scoring method works best for identification. Compounds with similar spectra compared to the user-submitted query are tabulated along with sortable match-scores and a graphical display of the selected mass spectral hit.

Moreover, based on the comprehensive compound and associated spectra information content of the GMD, an automated, machine-learning-based annotation of "unknown" spectra has been implemented [75]. Employing decision trees trained in a cross-validation setting, mass spectra can be annotated for/with the presence or absence of 21 different functional groups that frequently occur in known metabolites, *e.g.,* amino-, alcohol-, or carboxylic acid moieties. Thereby, even spectra for which no matching reference compound or respective metabolite spectrum can be found in the database, a basic classification of the "unknown" compound can be generated allowing, for example, to identify the likely candidate compounds for subsequent identification steps.

Finally, the GMD also supports text-based queries to search for particular metabolites, reference compounds and respective chemical derivatives, sum formula, molecular weight, functional chemical groups, KEGG identifiers, and other properties. All compounds contained in the GMD are displayed with a broad array of chemical and physical property information and links to external database resources.

5.3. The Fiehn library

The Fiehn library (FiehnLib) [76] comprises of >2,200 EI and retention indices for over 1,000 primary metabolites below 550 Da, covering authentic chemical standards of lipids, amino acids, fatty acids, amines, alcohols, sugars, amino-sugars, sugar alcohols, sugar acids, organic phosphates, hydroxyl acids, aromatics, purines, and sterols as methoximated and trimethylsilylated mass spectra under EI ionization.

There are two different Fiehn libraries created using either quadrupole or TOF mass detectors. The FiehnLib Agilent, was created on a single quadrupole mass detector operated in full scan mode from m/z 50 to 650. The FiehnLib LECO was obtained on a Pegasus IV TOF instrument using EI ionization from m/z 85 to 500. In both cases, fatty acid methyl esters (FAMEs) were used as internal retention index markers. This results in a disadvantage because very few labs use FAME markers. Complex polynomial calculations have to be used to convert those back to the more common Kovats RI. The FiehnLib libraries can be used in conjunction with instrument vendor software such as Agilent's ChemStation or Leco's ChromaTOF software. Additionally, LECO FiehnLib is also used in combination with SetupX/BinBase data processing system [77], which handles sample organization and class assignments and provides downloadable web services for results.

6. Overlap of LC/MS databases

As is clear from the above sections, there are many MS databases to choose from and their coverage of compounds of interest is a key question. Here, the overlap of compounds with MS^n spectra from authentic chemical standards in most public and commercial databases has been calculated for the first time (Figure 3). The InChIKey, a hashed version of the full standard IUPAC International Chemical Identifier (InChI) was used to compare the compound lists from each database. Structural information was collected from HMDB, MassBank, GNPS (http://gnps.ucsd.edu/) and ReSpect (http://spectra.psc.riken.jp/) (hereafter "public" databases as all information was downloadable) and converted to InChIKeys where appropriate (see Supplemental File for details). Representatives from Agilent, mzCloud, NIST14 and Wiley MS kindly provided InChIKey lists for their databases, hereafter "commercial databases". As the various databases contain different stereoisomers of the same substances, the comparison was performed using the first block of the InChIKey, which encodes the skeleton of the compound (the second block contains the stereochemistry). A total of 27,622 unique InChIKeys (first blocks, hereafter implicit) were present across all databases. Among open databases totaling 7,127 InChIKeys, we show that only 18 compounds (<1%) have some kind of spectral data in the four databases (Figure 3A). GNPS with 2,731 unique InChIKeys (87%) out of 3,136 is the largest public mass spectral database, while MassBank also has a large proportion of unique InChIKeys (2,080). When comparing all open databases combined versus the individual commercial ones (Figure 3B), only 225 compounds out of 27,622 (<1%) have some kind of spectral data in all databases. Interestingly, 3,345 compounds (12%) are unique to the open databases. Agilent METLIN, with 6,045 (21.8%) out of 27,622 InChIKeys, contains the greatest percentage of unique compounds. It is important to note that Agilent METLIN is also the greatest contributor among all databases with a total of 9,083 InChIKeys, >50% of which are di- and tri-peptides. Altogether, the ratio of compounds in each database with any type of spectral data in two or more databases is >50%, with the exception of Agilent METLIN and GNPS that only overlap with other databases in ~35% of their compounds (Figure 3C). Altogether these results show a relatively low overlap of compounds among existing spectral databases, which explains (and justifies) why most users currently query multiple databases.

7. Conclusions and future perspectives

It has been very positive for MS-based metabolomics that the number and quality of spectral databases has increased so rapidly over the last 5 years. However, this growth brings other problems that will need to be addressed soon to allow for genuine progress in metabolomics to be made (Table 2). Two major issues are evident, which could be best addressed by coordinated and centralized actions in the future.

(i) Only 5-10 % of the known metabolites reported in compound-centric databases such as HMDB and METLIN have annotated spectral data. To address this issue, a significant rise in MS, MS/MS and MS^n spectra from authentic chemical standards should best be tackled through an international initiative bringing together organic chemistry and metabolomics groups, and likely involving both the academic sector and commercial companies.

(ii) Despite the undeniable presence of naturally-occurring unknown metabolites (*i.e.*, not discovered previously) from non-targeted metabolomic studies, it is unclear whether this phenomenon is overestimated due to errors in adducts/fragments annotation and chemical/background noise [78]. To partially address this problem, a newer frontier of metabolomic databases characterized by well-annotated mass spectra containing all adduct and fragment species for reference substances is needed. In this regard, NIST14 and mzCloud have started to generate MS^n spectra of adducts in positive and negative mode. Moreover, saving full scan (MS1) spectral data from authentic chemical standards showing the diversity of adduct formation would enable the development of computational methods to deal with the annotation of mzRT features in LC/MS-based untargeted metabolomic studies.

In the current context, though, there are two opposite trends in spectral databases. The first is that in addition to human expertise, more and more computational mass spectrometric methods are being used to improve the quality of reference accurate mass spectra. This includes the signal processing and filtering to remove co-isolated peaks, automatic annotation of formulas to fragment peaks, re-calibration of spectra or even the annotation of fragment structures. In addition to the enriched information, all of these steps serve as additional quality control of the spectra, including detection of e.g. fragment peaks that can not be explained with a formula that is a subset of the parent ion. This additional information provides new ideas for the interpretation of mass spectra [57,64]. Furthermore, the development of extensive web services for compound databases such as PubChem and ChemSpider, as well as translational web services such as CACTUS (http://cactus.nci.nih.gov/) and the Chemical Translation Service (CTS) [79] mean that the more extensive annotation of compound information can now be semi-automated, also opening up a world of automatic curational opportunities whereby user-contributed spectra can potentially be improved retrospectively [57,64,65]. Although these improvements should facilitate spectra comparison and metabolite annotation, the spectral similarity and scoring functions are diverse and poorly described in most databases. For instance, MassBank appears to use modified cosine similarity, METLIN uses modified X-Rank, HMDB has forward and reverse fit, GMD can use up to five functions to filter the library search hits in regard to the specific distance measure (e.g., Euclidean, Hamming, Jaccard). Despite the

scoring system being a central algorithm of database search [55,65,80–83], mass spectra comparison is not a standardized process and can easily lead to different hits depending on which database is used.

The opposite trend is the collection of huge sets of spectra from regular experiments with very little preprocessing, let alone human curation efforts. In that case, typically the majority of the spectra will be unknowns, and can be expected to include co-isolation artifacts. This concept was already pioneered by the BinBase [77,84] system in particular for GC/MS spectra. Later, the MassBank-derived Bio-MassBank (http://bio.massbank.jp/) collected LC/MS spectra from different species, while massbank.eu has started collecting environmentally-relevant unknown spectra. GNPS [85] is to-date the largest collection of unknown MS/MS spectra, and includes massive computational efforts to cluster these by spectral similarity. These clusters, or bins in the BinBase system, can be annotated with similar reference spectra for further interpretation.

The role of standardization of spectral data related to the role of standardization in metabolomics is an issue under discussion, with pros and cons. While spectral standardization within a particular database is certainly helpful (e.g. mzCloud, METLIN), the diversity in acquisition is also beneficial for metabolite annotation as it can highlight different and common fragmentation processes across analytical conditions. Public databases have typically been collected in many different laboratories applying a multitude of different analytical methods, reflecting the analytically diverse nature of the metabolomic community. Therefore, standardization of spectra acquisition using one particular ionization source, mass analyzer and/or fragmentation technique would probably only be useful for a small percentage of groups. Instead, high quality data curation and advanced search features are equally or more as important as standardization, to enable users to choose the data they wish to use.

It is paramount that as much data is made openly available as possible, so that computational experts can access and interpret the data. Scientists should be encouraged to publish their spectra in an open and exchangeable electronic format to enhance the dissemination of information. Similar strategies exist already for all other “omics” techniques and data/meta-data repositories are now opening up for metabolomics (e.g., MetaboLights [86], Metabolomics Workbench [87]). The advantage of the big data approach and the collection of all information is the computational opportunities that arise to start characterizing common background, noise and interferences, which are perhaps otherwise uninteresting to publish but prone to cause misannotations in a non-target identification approach. Common laboratory contaminants, matrix substances and similar substances could thus be rapidly annotated as such and the focus can return to the true substances of interest.

Repositories of mass spectra from isotopically labelled experiments will soon become a reality. While isoMETLIN was released recently [88], this is so far restricted to the indexing of exact masses of variously-labeled species and does not contain any labelled spectra to date. In this regard, the prediction of labeled spectra from non-labelled spectra will need to be explored. This will require better computational tools for predicting and annotating spectra. While the prediction of some lipid spectra has been reasonably successful (as these are made up of consistent body pieces), prediction success is still a way off for most metabolite classes, where *in silico* fragmenters still predict many more fragments than are actually detected. It remains to be seen whether the quantum chemistry methods fare better.

Finally, with the advent of new instrumental setups in GC/MS that include GC coupled to high-resolution, high-accuracy mass analyzers such as quadrupole-Orbitrap [89,90] or quadrupole-TOF, existing libraries (e.g., NIST) may need to be revisited due to differences in mass accuracy and the relative intensities of fragment ions that can affect mass spectra similarity scores. In addition, no extensive collection of spectra are available yet for GC coupled to different chemical ionization (CI) MS [91] techniques, which will likely see increase in use in metabolomics.

Acknowledgements

We thank the financial support from the Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), an initiative of Instituto de Investigacion Carlos III (ISCIII), and the Spanish Ministry of Economy and Competitiveness grant SAF2011-30578. Miriam Navarro is the recipient of a MINECO fellowship (BES-2012-052585), under the SAF2011-30578 project. This work was supported in part by the SOLUTIONS project, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No. 603437. Finally, we are very thankful to Don Li and Emma Rennie from Agilent Technologies, Stephen Stein and Dmitrii Tchekhovskoi from NIST, Robert Mistrik from mzCloud, Mingxun Wang from UCSD (GNPS), Herbert Oberacher (MSforID, sold by Wiley) and Tobias Schulze (UFZ, massbank.eu) for providing their compound lists for the spectral overlap calculations, as well as all contributors to open databases and the reviewers for their helpful suggestions.

REFERENCES

[1] G.J. Patti, O. Yanes, G. Siuzdak, Innovation: Metabolomics: the apogee of the omics trilogy., Nat. Rev. Mol. Cell Biol. 13 (2012) 263–9. doi:10.1038/nrm3314.

[2] D.S. Wishart, Advances in metabolite identification., Bioanalysis. 3 (2011) 1769–82. doi:10.4155/bio.11.155.

[3] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, et al., Proposed minimum reporting standards for chemical analysis, Metabolomics. 3 (2007) 211–221. doi:10.1007/s11306-007-0082-2.

[4] D.S. Wishart, Computational strategies for metabolite identification in metabolomics., Bioanalysis. 1 (2009) 1579–1596. doi:10.4155/bio.09.138.

[5] D.J. Creek, W.B. Dunn, O. Fiehn, J.L. Griffin, R.D. Hall, Z. Lei, et al., Metabolite identification: are you sure? And how do your peers gauge your confidence?, Metabolomics. 10 (2014) 350–353. doi:10.1007/s11306-014-0656-8.

[6] L. Sumner, Z. Lei, B. Nikolau, K. Saito, U. Roessner, R. Trengove, Proposed quantitative and alphanumeric metabolite identification metrics, Metabolomics. 10 (2014) 1047–1049. doi:10.1007/s11306-014-0739-6.

[7] R. Salek, M. Arita, S. Dayalan, T. Ebbels, A. Jones, S. Neumann, et al., Embedding standards in metabolomics: the Metabolomics Society data standards task group, Metabolomics. 11 (2015) 782–783. doi:10.1007/s11306-015-0821-8.

[8] R. Goodacre, D. Broadhurst, A.K. Smilde, B.S. Kristal, J.D. Baker, R. Beger, et al., Proposed minimum reporting standards for data analysis in metabolomics, Metabolomics. 3 (2007) 231–241. doi:10.1007/s11306-007-0081-3.

[9] W.B. Dunn, A. Erban, R.J.M. Weber, D.J. Creek, M. Brown, R. Breitling, et al., Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics, Metabolomics. 9 (2013) 44–66. doi:10.1007/s11306-012-0434-4.

[10] J. Kalisiak, S.A. Trauger, E. Kalisiak, H. Morita, V. V. Fokin, M.W.W. Adams, et al., Identification of a new endogenous metabolite and the characterization of its protein interactions through an immobilization approach, J. Am. Chem. Soc. 131 (2009) 378–386. doi:10.1021/ja808172n.

[11] Z. Wang, E. Klipfell, B.J. Bennett, R. Koeth, B.S. Levison, B. Dugar, et al., Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease., Nature. 472 (2011) 57–63. doi:10.1038/nature09922.

[12] T. Kind, M. Scholz, O. Fiehn, How Large Is the Metabolome? A Critical Analysis of Data Exchange Practices in Chemistry, PLoS One. 4 (2009). doi:10.1371/journal.pone.0005440.

[13] M. Heinonen, H. Shen, N. Zamboni, J. Rousu, Metabolite identification and molecular fingerprint prediction through machine learning, Bioinformatics. 28 (2012) 2333–2341. doi:10.1093/bioinformatics/bts437.

[14] M. Gerlich, S. Neumann, MetFusion: Integration of compound identification strategies, J. Mass Spectrom. 48 (2013) 291–298. doi:10.1002/jms.3123.

[15] L. Li, R. Li, J. Zhou, A. Zuniga, A.E. Stanislaus, Y. Wu, et al., MyCompoundID: Using an evidence-based metabolome library for metabolite identification, Anal. Chem. 85 (2013) 3401–3408. doi:10.1021/ac400099b.

[16] F. Allen, A. Pon, M. Wilson, R. Greiner, D. Wishart, CFM-ID: A web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra, Nucleic Acids Res. 42 (2014). doi:10.1093/nar/gku436.

[17] F. Allen, R. Greiner, D. Wishart, Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification, Metabolomics. 11 (2014) 98–110. doi:10.1007/s11306-014-0676-4.

[18] L. Ridder, J.J.J. Van Der Hooft, S. Verhoeven, R.C.H. De Vos, R. Van Schaik, J. Vervoort, Substructure-based annotation of high-resolution multistage MSn spectral trees, Rapid Commun. Mass Spectrom. 26 (2012) 2461–2471. doi:10.1002/rcm.6364.

[19] L. Ridder, J. van der Hooft, S. Verhoevem, R.C.H. de Vos, R.J. Bino, J. Vervoort, Automatic Chemical Structure Annotation of an LC − MS^n Based Metabolic Profile from Green Tea, Anal. Chem. 85 (2013) 6033–6040.

[20] S. Grimme, Towards first principles calculation of electron impact mass spectra of molecules., Angew. Chem. Int. Ed. Engl. 52 (2013) 6306–12. doi:10.1002/anie.201300158.

[21] O. Fiehn, D.K. Barupal, T. Kind, Extending biochemical databases by metabolomic surveys, J. Biol. Chem. 286 (2011) 23637–23643. doi:10.1074/jbc.R110.173617.

[22] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, M. Kanehisa, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 27 (1999) 29–34. doi:10.1093/nar/27.1.29.

[23] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, et al., Reactome: A knowledgebase of biological pathways, Nucleic Acids Res. 33 (2005). doi:10.1093/nar/gki072.

[24] T. Kelder, M.P. Van Iersel, K. Hanspers, M. Kutmon, B.R. Conklin, C.T. Evelo, et al., WikiPathways: Building research communities on biological pathways, Nucleic Acids Res. 40 (2012). doi:10.1093/nar/gkr1074.

[25] P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, et al., Expansion of the Biocyc collection of pathway/genome databases to 160 genomes, Nucleic Acids Res. 33 (2005) 6083–6089. doi:10.1093/nar/gki892.

[26] Y. Wang, J. Xiao, T.O. Suzek, J. Zhang, J. Wang, S.H. Bryant, PubChem: A public information system for analyzing bioactivities of small molecules, Nucleic Acids Res. 37 (2009). doi:10.1093/nar/gkp456.

[27] H.E. Pence, A. Williams, Chemspider: An online chemical information resource, J. Chem. Educ. 87 (2010) 1123–1124. doi:10.1021/ed100697w.

[28] Z.-J. Zhu, A.W. Schultz, J. Wang, C.H. Johnson, S.M. Yannone, G.J. Patti, et al., Liquid chromatography quadrupole time-of-flight mass spectrometry characterization of metabolites guided by the METLIN database., Nat. Protoc. 8 (2013) 451–60. doi:10.1038/nprot.2013.004.

[29] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, et al., MassBank: A public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (2010) 703–714. doi:10.1002/jms.1777.

[30] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmüller, et al., GMD@CSB.DB: The Golm metabolome database, Bioinformatics. 21 (2005) 1635–1638. doi:10.1093/bioinformatics/bti236.

[31] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, et al., HMDB 3.0-The Human Metabolome Database in 2013, Nucleic Acids Res. 41 (2013). doi:10.1093/nar/gks1065.

[32] W. Lu, M.F. Clasquin, E. Melamud, D. Amador-Noguez, A.A. Caudy, J.D. Rabinowitz, Metabolomic analysis via reversed-phase ion-pairing liquid chromatography coupled to a stand alone orbitrap mass spectrometer, Anal. Chem. 82 (2010) 3212–3221. doi:10.1021/ac902837x.

[33] C.A. Smith, E.J. Want, G. O'Maille, R. Abagyan, G. Siuzdak, XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification, Anal. Chem. 78 (2006) 779–787. doi:10.1021/ac051437y.

[34] A. Lommen, Metalign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing, Anal. Chem. 81 (2009) 3079–3086. doi:10.1021/ac900036d.

[35] M. Katajamaa, J. Miettinen, M. Oresic, MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data., Bioinformatics. 22 (2006) 634–6. doi:10.1093/bioinformatics/btk039.

[36] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data., BMC Bioinformatics. 11 (2010) 395. doi:10.1186/1471-2105-11-395.

[37] M. Vinaixa, S. Samino, I. Saez, J. Duran, J.J. Guinovart, O. Yanes, A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data, Metabolites. 2 (2012) 775–795. http://www.mdpi.com/2218-1989/2/4/775.

[38] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Anal. Chem. 84 (2012) 283–289. doi:10.1021/ac202450g.

[39] A. Alonso, A. Julià, A. Beltran, M. Vinaixa, M. Díaz, L. Ibañez, et al., AStream: An R package for annotating LC/MS metabolomic data, Bioinformatics. 27 (2011) 1339–1340. doi:10.1093/bioinformatics/btr138.

[40] CRAN - Package nontarget, (n.d.). http://cran.r-project.org/web/packages/nontarget/index.html (accessed May 18, 2015).

[41] M. Brown, D.C. Wedge, R. Goodacre, D.B. Kell, P.N. Baker, L.C. Kenny, et al., Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets, Bioinformatics. 27 (2011) 1108–1112. doi:10.1093/bioinformatics/btr079.

[42] A. Valdés, V. García-Cañas, C. Simó, C. Ibáñez, V. Micol, J.A. Ferragut, et al., Comprehensive Foodomics Study on the Mechanisms Operating at Various Molecular Levels in Cancer Cells in Response to Individual Rosemary Polyphenols, Anal. Chem. 86 (2014) 9807–9815. doi:10.1021/ac502401j.

[43] C. Ibanez, C. Simo, V. Garcia-Canas, A. Gomez-Martinez, J.A. Ferragut, A. Cifuentes, CE/LC-MS multiplatform for broad metabolomic analysis of dietary polyphenols effect on colon cancer cells proliferation, Electrophoresis. 33 (2012) 2328–2336. doi:10.1002/elps.201200143.

[44] C. Ibanez, A. Valdes, V. Garcia-Canas, C. Simo, M. Celebier, L. Rocamora-Reverte, et al., Global Foodomics strategy to investigate the health benefits of dietary constituents, J Chromatogr A. 1248 (2012) 139–153. doi:10.1016/j.chroma.2012.06.008.

[45] D.S. Wishart, D. Tzur, C. Knox, R. Eisner, A.C. Guo, N. Young, et al., HMDB: the Human Metabolome Database, Nucleic Acids Res. 35 (2007) D521–6. doi:10.1093/nar/gkl923.

[46] K. Degtyarenko, P. De matos, M. Ennis, J. Hastings, M. Zbinden, A. Mcnaught, et al., ChEBI: A database and ontology for chemical entities of biological interest, Nucleic Acids Res. 36 (2008). doi:10.1093/nar/gkm791.

[47] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, et al., The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. 31 (2003) 365–370. doi:10.1093/nar/gkg095.

[48] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, et al., GenBank, Nucleic Acids Res. 41 (2013). doi:10.1093/nar/gks1195.

[49] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, B.A. Rapp, D.L. Wheeler, GenBank., Nucleic Acids Res. 28 (2000) 15–18. doi:doi: 10.1093/nar/28.1.15.

[50] E.P. Go, Database resources in metabolomics: An overview, J. Neuroimmune Pharmacol. 5 (2010) 18–30. doi:10.1007/s11481-009-9157-3.

[51] S. Wolf, S. Schmidt, M. Muller-Hannemann, S. Neumann, In silico fragmentation for computer assisted identification of metabolite mass spectra, BMC Bioinformatics. 11 (2010) 148. doi:10.1186/1471-2105-11-148.

[52] L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, et al., mzML--a community standard for mass spectrometry data., Mol. Cell. Proteomics. 10 (2011) R110.000133. doi:10.1074/mcp.R110.000133.

[53] C.A. Smith, G. O'Maille, E.J. Want, C. Qin, S.A. Trauger, T.R. Brandon, et al., METLIN: a metabolite mass spectral database., Ther. Drug Monit. 27 (2005) 747–751. doi:10.1097/01.ftd.0000179845.53213.39.

[54] T.R. Sana, J.C. Roark, X. Li, K. Waddell, S.M. Fischer, Molecular formula and METLIN personal metabolite database matching applied to the identification of compounds generated by LC/TOF-MS, J. Biomol. Tech. 19 (2008) 258–266.

[55] R. Mylonas, Y. Mauron, A. Masselot, P.-A. Binz, N. Budin, M. Fathi, et al., X-Rank: A Robust Algorithm for Small Molecule Identification Using Tandem Mass Spectrometry, Anal. Chem. 81 (2009) 7604–7610. doi:10.1021/ac900954d.

[56] MassBank | Download, (n.d.). http://www.massbank.jp/en/download.html (accessed May 18, 2015).

[57] M.A. Stravs, E.L. Schymanski, H.P. Singer, J. Hollender, Automatic recalibration and processing of tandem mass spectra using formula annotation, J. Mass Spectrom. 48 (2013) 89–99. doi:10.1002/jms.3131.

[58] The Chemical Translation Service – a web-based tool to improve standardization of metabolomic reports | InChI Trust, (n.d.). http://www.inchi-trust.org/the-chemical-translation-service-a-web-based-tool-to-improve-standardization-of-metabolomic-reports/ (accessed May 18, 2015).

[59] A. McNaught, The IUPAC international chemical identifier : InChI-A new standard for molecular informatics, Chem. Int. 28 (2006) 12–14. doi:10.1016/j.hrthm.2010.04.035.

[60] E. Fahy, M. Sud, D. Cotter, S. Subramaniam, LIPID MAPS online tools for lipid research, Nucleic Acids Res. 35 (2007). doi:10.1093/nar/gkm324.

[61] E. Fahy, D. Cotter, M. Sud, S. Subramaniam, Lipid classification, structures and tools, Biochim. Biophys. Acta - Mol. Cell Biol. Lipids. 1811 (2011) 637–647. doi:10.1016/j.bbalip.2011.06.009.

[62] T. Kind, K.H. Liu, Y. Lee do, B. DeFelice, J.K. Meissen, O. Fiehn, LipidBlast in silico tandem mass spectrometry database for lipid identification, Nat Methods. 10 (2013) 755–758. doi:10.1038/nmeth.2551.

[63] N. US Department of Commerce, NIST Standard Reference Database 1A v14, (n.d.). http://www.nist.gov/srd/nist1a.cfm.

[64] X. Yang, P. Neta, S.E. Stein, Quality control for building libraries from electrospray ionization tandem mass spectra, Anal. Chem. 86 (2014) 6393–6400. doi:10.1021/ac500711m.

[65] S. Stein, Mass spectral reference libraries: An ever-expanding resource for chemical identification, Anal. Chem. 84 (2012) 7274–7282. doi:10.1021/ac301205z.

[66] A.J. Lubeck, D.L. Sutton, Kovats Retention Indices of Selected Hydrocarbons through C10 on Bonded Phase Fused Silica Capillaries, J. High Resolut. Chromatogr. Chromatogr. Commun. (1983) 328–332. http://www.wfu.edu/chemistry/courses/jonesbt/334/RetentionIndex.pdf (accessed May 18, 2015).

[67] T. Kind, G. Wohlgemuth, Y. Lee do, Y. Lu, M. Palazoglu, S. Shahbaz, et al., FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry, Anal Chem. 81 (2009) 10038–10048. doi:10.1021/ac9019522.

[68] V.I. Babushok, P.J. Linstrom, J.J. Reed, I.G. Zenkevich, R.L. Brown, W.G. Mallard, et al., Development of a database of gas chromatographic retention properties of organic compounds., J. Chromatogr. A. 1157 (2007) 414–21. doi:10.1016/j.chroma.2007.05.044.

[69] N.I. of S. and T. (NIST), Deconvolution, Automated Mass Spectral. “Identification System software (AMDIS ver. 2.1.).,” 2005.

[70] J. Hummel, J. Selbig, D. Walther, J. Kopka, The golm metabolome database: A database for GC-MS based metabolite profiling, Top. Curr. Genet. 18 (2007) 75–95. doi:10.1007/4735_2007_0229.

[71] S.E. Stein, V.I. Babushok, R.L. Brown, P.J. Linstrom, Estimation of Kováts Retention Indices Using Group Contributions, J. Chem. Inf. Model. 47 (2007) 975–980. doi:10.1021/ci600548y.

[72] S.E. Stein, Chemical substructure identification by mass spectral library searching, J. Am. Soc. Mass Spectrom. 6 (1995) 644–655. doi:10.1016/1044-0305(95)00291-K.

[73] K. Varmuza, W. Werther, Mass Spectral Classifiers for Supporting Systematic Structure Elucidation, J. Chem. Inf. Model. 36 (1996) 323–333. doi:10.1021/ci9501406.

[74] W. Werther, H. Lohninger, F. Stancl, K. Varmuza, Classification of mass spectra, Chemom. Intell. Lab. Syst. 22 (1994) 63–76. doi:10.1016/0169-7439(94)85018-6.

[75] J. Hummel, N. Strehmel, J. Selbig, D. Walther, J. Kopka, Decision tree supported substructure prediction of metabolites from GC-MS profiles, Metabolomics. 6 (2010) 322–333. doi:10.1007/s11306-010-0198-7.

[76] T. Kind, G. Wohlgemuth, D.Y. Lee, Y. Lu, M. Palazoglu, S. Shahbaz, et al., FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry., Anal. Chem. 81 (2009) 10038–48. doi:10.1021/ac9019522.

[77] K. Skogerson, G. Wohlgemuth, D.K. Barupal, O. Fiehn, The volatile compound BinBase mass spectral database., BMC Bioinformatics. 12 (2011) 321. doi:10.1186/1471-2105-12-321.

[78] M. Brown, W.B. Dunn, P. Dobson, Y. Patel, C.L. Winder, S. Francis-McIntyre, et al., Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics, Analyst. 134 (2009) 1322–1332. doi:10.1039/b901179j.

[79] G. Wohlgemuth, P.K. Haldiya, E. Willighagen, T. Kind, O. Fiehn, The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports., Bioinformatics. 26 (2010) 2647–8. doi:10.1093/bioinformatics/btq476.

[80] F.W. McLafferty, D.A. Stauffer, S.Y. Loh, C. Wesdemiotis, Unknown identification using reference mass spectra. Quality evaluation of databases., J. Am. Soc. Mass Spectrom. 10 (1999) 1229–40. doi:10.1016/S1044-0305(99)00104-X.

[81] H. Horai, M. Arita, T. Nishioka, Comparison of ESI-MS Spectra in MassBank Database, in: 2008 Int. Conf. Biomed. Eng. Informatics, IEEE, 2008: pp. 853–857. doi:10.1109/BMEI.2008.339.

[82] H. Oberacher, M. Pavlic, K. Libiseller, B. Schubert, M. Sulyok, R. Schuhmacher, et al., On the inter-instrument and inter-laboratory transferability of a tandem mass spectral reference library: 1. Results of an Austrian multicenter study., J. Mass Spectrom. 44 (2009) 485–93. doi:10.1002/jms.1545.

[83] S.E. Stein, D.R. Scott, Optimization and testing of mass spectral library search algorithms for compound identification., J. Am. Soc. Mass Spectrom. 5 (1994) 859–66. doi:10.1016/1044-0305(94)87009-8.

[84] O. Fiehn, G. Wohlgemuth, M. Sholz, Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata, in: DILS, 2005: pp. 224–239. doi:10.1007/11530084.

[85] K.R. Duncan, M. Crüsemann, A. Lechner, A. Sarkar, J. Li, N. Ziemert, et al., Molecular Networking and Pattern-Based Genome Mining Improves Discovery of Biosynthetic Gene Clusters and their Products from Salinispora Species, Chem. Biol. 22 (2015) 460–71. doi:10.1016/j.chembiol.2015.03.010.

[86] R.M. Salek, K. Haug, P. Conesa, J. Hastings, M. Williams, T. Mahendraker, et al., The MetaboLights repository: curation challenges in metabolomics., Database (Oxford). 2013 (2013). doi:10.1093/database/bat029.

[87] R.M. Salek, C. Steinbeck, M.R. Viant, R. Goodacre, W.B. Dunn, The role of reporting standards for metabolite annotation and identification in metabolomic studies., Gigascience. 2 (2013) 13. doi:10.1186/2047-217X-2-13.

[88] K. Cho, N. Mahieu, J. Ivanisevic, W. Uritboonthai, Y.-J. Chen, G. Siuzdak, et al., isoMETLIN: a database for isotope-based metabolomics., Anal. Chem. 86 (2014) 9358–61. doi:10.1021/ac5029177.

[89] A.C. Peterson, J.-P. Hauschild, S.T. Quarmby, D. Krumwiede, O. Lange, R.A.S. Lemke, et al., Development of a GC/Quadrupole-Orbitrap mass spectrometer, part I: design and characterization., Anal. Chem. 86 (2014) 10036–43. doi:10.1021/ac5014767.

[90] A.C. Peterson, A.J. Balloon, M.S. Westphall, J.J. Coon, Development of a GC/Quadrupole-Orbitrap mass spectrometer, part II: new approaches for discovery metabolomics., Anal. Chem. 86 (2014) 10044–51. doi:10.1021/ac5014755.

[91] N. Strehmel, J. Kopka, D. Scheel, C. Böttcher, Annotating unknown components from GC/EI-MS-based metabolite profiling experiments using GC/APCI(+)-QTOFMS, Metabolomics. 10 (2013) 324–336. doi:10.1007/s11306-013-0569-y.

[92] LipidSearch™ Software-Thermo Scientific, (2013). http://www.thermoscientific.com/content/tfs/en/product/lipidsearch-software.html.

[93] mzCloud - advanced mass spectral database, (n.d.). https://www.mzcloud.org/.

[94] Herbert Oberacher, Wiley Registry of Tandem Mass Spectral Data, MS for ID, (2012).

[95] R.J.M. Weber, E. Li, J. Bruty, S. He, M.R. Viant, MaConDa: a publicly accessible mass spectrometry contaminants database., Bioinformatics. 28 (2012) 2856–7. doi:10.1093/bioinformatics/bts527.

[96] Y. Sawada, R. Nakabayashi, Y. Yamada, M. Suzuki, M. Sato, A. Sakata, et al., RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database, Phytochemistry. 82 (2012) 38–45. doi:10.1016/j.phytochem.2012.07.007.

Figure Captions

Figure 1: Pie charts representing the distribution of adducts and number of MS/MS spectra recorded in NIST 14 database using Thermo Finnigan Elite Orbitrap (ESI-HCD) or Agilent 6530 Q-TOF (ESI-CID) operating in either positive or negative ionization mode.

Figure 2: (A) Mass spectra of glucose 6-phosphate (KEGG cpd: C00092) under different analytical conditions (*i.e.*, mobile phases). (B) MS/MS spectra from different precursor ion species corresponding to observed adducts.

Figure 3: (A) Venn diagram showing the overlap between open mass spectral databases (HMDB, MassBank, GNPS and ReSpect). (B) Venn diagram showing the overlap between five commercial databases (Agilent METLIN PMD, mzCloud, NIST14 and Wiley MS) and open databases described in A. (C) Number and percentage of unique and shared compounds (*i.e.*, InChIKey) with MS^n ($n \geq 2$) data in each database in relation to all eight resources.

Tables

Table 1. MS/MS data from electrospray ionization in NIST 14**.**

Mass Analyzer	**# spectra**	**# compounds**	**# ions**
Ion trap	~39,000	~6,000	~39,000
qTOF (CID)	~42,000	~3,000	~4,000
QqQ (CID)	~27,000	~1,000	~3,000
Orbitrap (HCD)	~69,000	~3,000	~6,000
Ion trap with FTMS	~5,000	~2,500	NA

Table 2. Summary of the most widely used mass spectral databases in metabolomics

Database	**Pros**	**Cons**
HMDB [31]	- Public - Mass spectral data on ~9,500 chemical standards. - Spectral data is downloadable	- Mixed collision energies and instrument types.
METLIN [53]	- Public - Curated mass spectral data on > 13,000 chemical standards. - Over 63,500 high-resolution MS/MS spectra.	- Only QTOF data - Spectral data is not downloadable.
LipidSearch [92]	- Over 1.5 million lipid ions and their predicted fragment ions. - Includes lipid adduct ions and MS^n fingerprints. - Data are stored in XML files.	- Commercial license required - Developed for Orbitrap technology. - *In silico* generated MS/MS library. - Overlap with LipidBlast is unclear
LipidBlast [62]	- Over 200,000 tandem mass spectra covering 25 lipid classes. - Publicly available	- *In silico* generated library using heuristic modeling of tandem mass spectra.

	- Spectral data is downloadable.	- "One third rule" limitation: developed with mostly ion-trap tandem mass spectra. - Does not allow batch search of precursor ions. - Overlap with LipidSearch unclear.
LipidMaps [60]	- Over 40,000 unique lipid structures. - Spectral data is downloadable.	- MS/MS spectra only predicted in negative or positive ionization mode. - MS/MS spectra only available for one adduct per lipid.
mzCloud [93]	- Public - Highly curated MS/MS and MS^n spectral information. - Spectral peaks are structurally annotated.	- Low number of metabolites. - Spectral data is not downloadable. - Only Orbitrap spectra.
Wiley 10^{th} [94]	- Largest mass spectral library commercially available. - 719,000 spectra (>950,000 spectra if combined with NIST 14). - Over 638,000 compounds (>760,000 compounds if combined with NIST 14). - Compatible with most instrument manufacturers.	- Commercial license required. - Only 70 eV EI mass spectra. - Beyond metabolomic applications.
MaConDa [95]	- Public database of ~200 contaminants in mass spectrometry. - Theoretical and experimental spectral records detected across several MS platforms. - Downloadable	- Has no MS/MS data.
MassBank [29]	- Public - Mass spectra from different MS setups. - Roughly 19,000 MS1 and 28,000 MS2 and MS^n spectra. - Spectral data is downloadable.	- Not sufficiently curated.
NIST14 [63]	- 234,284 ESI MS/MS spectra of 9,344 chemical standards.	- Commercial license. - Lack of additional identifiers to

	- Large number of MS/MS spectra from adducts. - MS/MS spectra recorded using multiple high- and low-resolution instruments. - Curated collection of 276,259 EI mass spectra from 242,477 unique compounds. - 387,463 measured Kovats or Lee retention index information from 82,337 chemical standards.	external database resources.
GMD [30]	- Public - Over 2,500 EI mass spectra and retention index information. - Spectral data is downloadable.	- Data derived primarily from plant materials.
FiehnLib [67]	- Over 2,200 EI and retention indices for over 1,000 metabolites.	- Commercial license with Agilent Technologies and Leco Corporation. - Data derived primarily from plant materials.
ReSpect [96]	- Public - More than 9,000 MS/MS spectra corresponding to > 3,600 metabolites: ~38% literature data ~12% QTOF MS/MS ~50% QqQ MS/MS - Merged spectra (same as MassBank) - Curated record data - Downloadable	- Only QTOF and QqQ MS data - Mainly phytochemicals (plant metabolomics) - High degree of redundancy with MassBank
GNPS [85]	- Public - 8,853 MS/MS spectra - MS/MS of adducts - MS/MS of unidentified structures - Downloadable	-Very few spectra in negative ionization - Limited spectrum information - No spectral clean-up/noise removal - “gold standard” is not comparable with reference databases