Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography–High-Resolution Tandem Mass Spectrometry and NMR Profiling, in Silico Databases, and Chemometrics

Jean-Luc Wolfender, Jean-Marc Nuzillard, Justin van Der Hooft, Jean-Hugues Renault, Samuel Bertrand

Accelerating metabolite identification in natural product research: toward an ideal combination of LC-HRMS/MS and NMR profiling, *in silico* databases and chemometrics

Jean-Luc Wolfender,[1] Jean-Marc Nuzillard,[2] Justin J. J. van der Hooft,[3] Jean-Hugues Renault,[2] Samuel Bertrand[4,5]

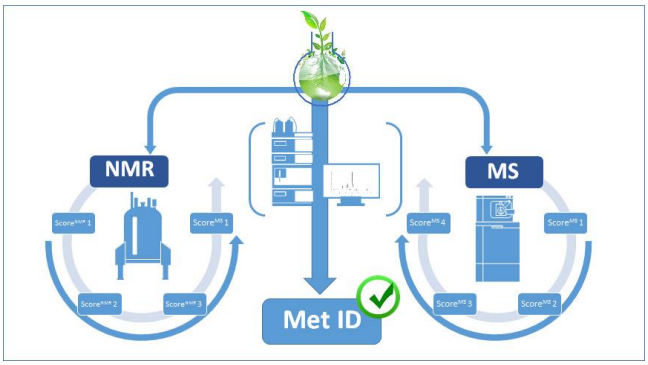
[1] School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, CMU, 1, Rue Michel Servet, 1211 Geneva 4, Switzerland

[2] Institut de Chimie Moléculaire de Reims, UMR CNRS 7312, Université de Reims Champagne Ardenne, France

[3] Bioinformatics Group, Wageningen University, Wageningen 6708 PB, The Netherlands

[4] Groupe Mer, Molécules, Santé-EA 2160, UFR des Sciences Pharmaceutiques et Biologiques, Université de Nantes, France

[5] ThalassOMICS metabolomics facility, Plateforme Corsaire, Biogenouest, 44035 Nantes, France



Abstract

The rapid innovations in metabolite profiling, bioassays and chemometrics have led to a paradigm shift in natural product (NP) research. Indeed, having partial or full structure information about possibly "all" specialized metabolites and an estimation of their levels in plants or microorganisms provides a way to perform pharmacognostic or chemical ecology investigations from a new and holistic perspective. The increasing amount of accurate metabolome data that can be acquired on massive sample sets, notably through data-dependent LC-HRMS/MS and NMR profiling, allows the mapping of natural extracts at an unprecedented level of precision.

Most progress made recently in accelerating metabolite identification has been pushed by the need for metabolomics to have tools that provide a confident annotation of the biomarkers highlighted as the results of data mining through multivariate analysis, often on important datasets of complex samples. Historically, NP chemists have been involved in the unambiguous *full de novo* identification of unknown compounds from complex natural biological matrices. This process is classically performed by the tedious isolation of pure bioactive NPs through comprehensive bioactivity-guided isolation workflows involving orthogonal chromatographic steps at the preparative level. Increasingly advanced metabolomics metabolite profiling methods are of strategic importance in dereplication workflows in NP research as well as for the full metabolome composition assignment of relevant organisms from both drug discovery and chemical ecology perspectives.

In this review, we describe the latest developments in metabolite profiling by both LC-MS and NMR-based methods and related databases from a natural product chemist perspective. We assess the current possibilities and limits of such methods and the workflows for manual and automated NP annotations by equally treating the MS and NMR approaches that are both key for the "as confident as possible" NP annotation in crude natural extracts. We also propose future lines of development in the field that are important for NP research but are also generally needed for metabolite annotation in metabolomics because NPs represent perfect candidate compounds for identification due to their intrinsic structural complexity and chemodiversity across organisms. This review does not aim to provide a comprehensive survey of all metabolite profiling applications made in NP research to date. Typical case studies are discussed, and an update of a selection of the latest advanced original studies and numerous specialized reviews is made with links to tools and DBs regarded as useful for their current or future usage in NP research. Evaluations of what can be readily implemented and what is still required for confident NP structural elucidation are made, especially concerning access to generic structural and spectral DBs as well as the use of orthogonal detection methods for improved confidence in metabolite annotation.

Keywords: Database, Dereplication, *in silico*, LC-MS, Metabolite annotation, Metabolite profiling, Metabolomics, Natural products, NMR

1 Introduction

Natural product research aims to characterize specialized metabolites from various living organisms and assess their biological properties from either a chemical ecology or drug discovery viewpoint.

These "**specialized**" metabolites, which are oftentimes referred to as "**secondary metabolites**", are non-essential to sustain the life of a given organism but necessary for its survival in a given environment, in contrast to "**primary metabolites**", including amino acids, lipids, and carbohydrates, which are necessary for physiology purposes.[1] Specialized metabolites are small-molecular-weight molecules (typically < 1500 Da) mainly involved in processes like defense against other biotic as well as abiotic agents, or used as attractants for reproduction purposes due to the sessile lifestyles of many organisms, such as plants and microorganisms.[2] Throughout this review, we will refer to such metabolites as **natural products** (NPs).

NPs from all living organisms have evolved and diversified for increased fitness within a specific environment. This has resulted in plants and microorganisms achieving the synthesis of distinct sets of NPs. Such pressure of evolution has generated the huge chemodiversity of NPs in nature. In plants, for example, and as stated by Pichersky and Lewinsohn, the total number of NPs found in the kingdom by far exceeds the capacity of any one plant genome to encode the necessary enzymes, and just as a plant lineage acquires the ability to elaborate new specialized compounds during evolution, it also loses the ability to produces others.[2]

NPs thus include a large and diverse group of compounds from a variety of sources, mainly plants, bacteria, and fungi, from terrestrial and aquatic biotopes. They play significant biological roles in all organisms and have evolved to interact with enzymes, receptors and ion channels. Some are active in living cells and are even from different organisms, able to cross cell membranes and interfere with enzymes or even act against parasites.[3] Due to their co-evolution in natural systems, NPs are therefore encoded to be bioactive and of high interest in the drug discovery field. They have long been used as medicines, and today, they continue to be a reservoir of potential drugs.[3-4] Thus, NPs and their related structures serve as essential sources of new chemical entities for the pharmaceutical industry due to their immense variety of functionally relevant compounds.[5]

The **chemical space** encompassed by NPs is very large,[6] and more than 250,000 NPs have been reported to date in the dictionary of natural products (DNP).[7] The DNP surveys literature data of all NPs characterized worldwide as the result of the isolation work and full *de novo* identification of varied organisms with their taxonomic origin.[7] It is important to note that the DNP mainly focuses on plant resources, implying that when the microbial chemical space is added, the number above can easily quadruple. The chemical space is characterized by a "multi-dimensional descriptor space" in which NPs

can be associated with a wide range of "descriptors" and "properties", such as their molecular mass, lipophily (their affinity for a lipidic environment), compound class, and the topological features of their molecular structure.[8] A measurement of the chemodiversity of NPs can be obtained by evaluating the size of the chemical space visualized by the principal component analysis (PCA) of sets of molecular descriptors. This space has been shown to be much larger for NPs than those occupied by new chemical entities coming from combinatorial chemistry.[9] Both drugs and NPs cover similar parts of chemical space, demonstrating the potential of many NPs to become leads for drug discovery.[10] The methods used to navigate this chemodiversity space have constantly evolved. For example, ChemGPS-NP assists in compound selection and prioritization, property description and interpretation, cluster analysis and neighborhood mapping, as well as the comparison and characterization of large compound data sets.[11]

Notwithstanding the potential of NPs to become effective drugs, the drug discovery workflow that leads from crude natural extracts to well-characterized bioactive NPs as hits and then as lead compounds is considered complex, slow, costly, and often not compatible with the pace of high-throughput screening campaigns. This explains in part why many pharmaceutical industries slowed down and then terminated most of their NP-oriented research programs in the early 2000s.[3, 12]

This difficulty of working with NPs is in part related to the very high complexity of the biological matrices (**natural crude extracts**) in which they are embedded, which in turn causes their chemical richness. NPs as pure active ingredients are typically extracted from plants and microorganisms with solvents of different polarities (usually hydroalcoholic mixtures, methanol, ethyl acetate or methylene chloride). Each of these extracts typically contains 10's of main natural products and 100's or 1,000's of less abundant ones. In addition, polar extracts are dominated by primary metabolites, mainly saccharides, and lipophilic extracts are dominated by various types of lipids and pigments. These compounds are part of the metabolome but are often not of interest for bioactivity. There are, however, exceptions; for example, polysaccharides are known to exhibit immuno-stimulating effects[13] and the high number of bioactive suggesting their functional roles.[14] Potentially all NPs, even minor constituents, may have interesting biological properties (e.g., potent defense toxins, hormones). Their identification requires using metabolite profiling methods that are able to work over a **large dynamic range** and generate information-rich spectral data for their full or partial identification.[15]

The identification of bioactive NPs from such complex matrices is classically performed in **pharmacognosy** (search for bioactive compounds form natural sources) by **bioactivity-guided isolation approaches**. Here, crude extracts exhibiting given biological activity are fractionated by a combination of preparative chromatographic methods. All fractions are submitted to bioassays, and those fractions continuing to exhibit activity are carried through further isolation and purification steps

until pure active ingredients are obtained. These ingredients are then fully characterized by a combination of NMR, HR/MS and chiroptical spectroscopic methods (*i.e.* sensitive to molecular chirality, such as electronic circular dichroism – ECD – and vibrational circular dichroism – VCD) until their structure and absolute configuration are obtained.[16] This process is slow but effective and has led to major breakthroughs in NP research, such as the discovery of artemisinin (a sesquiterpene lactone containing an unusual peroxide) isolated from *Artemisia annua*. Artemisinin has become a reference drug for the treatment of malaria, and its discovery by Professor Tu Youyou led her to be awarded the Nobel Prize of Medicine in 2015.[17]

To rationalize the process that yields interesting active ingredients, the metabolite profiling of crude extracts and dereplication prior to isolation has been underway for years in classical NP research.[18] **Dereplication** is the process of differentiating novel compounds from those that have already been studied.[12, 19] Since its appearance in 1990, dereplication has significantly evolved over the last decades. It has been used in different workflows ranging from major compound identification and the acceleration of activity-guided fractionation up to the chemical profiling of collected extracts.[18]

In parallel to these advances in metabolite profiling, the field of metabolomics appeared at the beginning of the millennium for life science applications[20] and experienced exponential growth until today, as is the case for other omics approaches. **Metabolomics** is defined as a non-selective, universally applicable, comprehensive analytical approach for the identification and quantitation of metabolites in a biological system. This area of research strives to obtain complete metabolite fingerprints, detect differences between metabolites and generate hypotheses to explain these differences.[19] Metabolomics is practically considered the **large-scale analysis of metabolites** of a given organism during various physiological states,[21] but it also extends to the comparative comprehensive metabolite profiling for **deep/full metabolome** analyses for **chemotaxonomic investigations** and **NP prioritization** studies in drug discovery. Tools in metabolomics have tremendously evolved over the last decade because such an unbiased data-driven approach has served many fields of life sciences and has also strongly influenced various aspects of NP research, notably in giving additional dimensions to dereplication. These developments were partly driven by the progress made in the acquisition techniques of the metabolite profiles in complex biological matrices in both the MS and NMR fields in terms of sensitivity, resolution and throughput, but also more recently by the introduction of *in silico* and chemometric associated methods.[22] It was recognized that traditional analysis methods only slightly dipped into the complete pool of molecules present in complex mixtures, thereby leaving a large amount of "dark matter."[23] These unknowns potentially represent much-needed novel bioactive molecules that could, for example, be used to combat antibiotic resistance.

In metabolomics, **putative or partial metabolite identification** from metabolite profiles or fingerprints of complex extracts is referred to as "**annotation"**. Today, this process still represents a major bottleneck in metabolomics because annotation is often not unambiguous, and only putative or partial assignments can be made. When compared to biological fluid metabolomics, this aspect is even more striking in NP research because NP chemodiversity is very extensive, and sample compositions vary substantially based on the organisms that are screened. In contrast, in biological fluid analysis, redundant metabolites are often profiled, and in this case, quantitation aspects are key to observing sometimes minute but significant changes in profiles (e.g., those related to disease or diet changes).

To assess the **level of confidence of metabolite annotation,** different reporting standards for identification have been defined by metabolomics researchers. This resulted in a four-level system ranging from Level 1 (identified compound) *via* Levels 2 and 3 (putatively annotated compounds and compound classes) to Level 4 (unidentified or unclassified metabolites that can be differentiated based on analytical data).[24]

In typical NP research, unknown metabolites must be **fully characterized *de novo* after isolation**.[25] This also occurs redundantly for known NPs when the dereplication process is not sufficiently efficient. The additional full assessment of their absolute configuration by chiroptical methods and sometimes X-ray crystallography is also often required because 3D structural characterization is key to understanding ligand target interactions in pharmacological investigations of NPs.[26]

Full characterization clearly provides the high-quality unambiguous identification of metabolites but is time-consuming and often not worth the effort, especially when known compounds are redundantly characterized and their spectroscopy data have been published several times. Moreover, the isolation and full identification of minor compounds requires large amounts of biological material. Additionally, the pace of this work is not compatible with the pace at which HTS screening campaigns are performed on extracts for drug discovery purposes.[3]

The rapid high-quality identification of NPs is not only necessary for rationally characterizing active ingredients but also increasingly needed for obtaining detailed exhaustive composition information for herbal products used in traditional medicine, nutraceutical products or botanicals with claimed clinical efficacy.[27] This can facilitate linking composition with possible efficacy, screening for possible toxic NPs, and establishing composition trends for given therapeutic usage in evidence-based approaches.[28] It is also needed to support quality control studies that increasingly rely on fingerprints rather than on single marker determination for herbal products.[29] This need for accurate composition determination is even more complex when studying drugs used by traditional Chinese medicine (TCM), where multiherb preparations are often used.[30]

Altogether, only a few “complete” workflows exist that can take the researcher from raw data to performing the robust annotations and identifications of metabolites in complex mixtures; in particular, when investigating completely novel bioresources, existing tools often fail to identify reliable candidate molecules. Moreover, their usage is not always straightforward for NP researchers, which hampers their interpretation of downstream results. Thus, there is a strong need for metabolite profiling methods and data mining workflows that provide a much higher confidence of NP identification with reliable annotation scores that can be achieved with a high degree of automation.

In this review, we will describe state-of-the-art metabolite profiling and data analysis methods based on both LC-MS and NMR profiling that are currently used in NP research and metabolomics or related fields or can be implemented. In particular, their coverage in terms of full metabolome analysis will be discussed; however, this review will focus on the structure elucidation of NPs from complex mixtures. The present spectral NMR and MS/MS databases (DBs) suitable for NP annotation will be surveyed together with the different tools that can be used to generate searchable spectra generated *in silico* from structural NP DBs. Various recent workflows that can lead to annotation will be described and assessed, especially in terms of their usage/implementation in NP research and their ease of operation and level of automation for natural product chemists. Because both MS and NMR data are important for the characterization of NPs, their corresponding dereplication workflows will be treated equally. Current approaches integrating both MS and NMR analytical dimensions will be highlighted, and ideas for progress that can be made for better practical integration will be provided. Future prospects that may also come from the addition of orthogonal methods to LC-MS, such as collisional cross section (CCS) obtained with ion mobility measurements or retention time (RT) predictions, will be put in perspective. Finally, we will share our views on the development that is needed in terms of the contextualization (e.g., taxonomical data) of the metabolomics data generated from NP extracts. The creation of a novel method to combine scores from different and, if possible, orthogonal spectral/physico-chemical information for more reliable annotation and the expansion of *in silico* candidate DBs through predictions of structural variation by exploiting knowledge of natural biosynthesis pathways will also be discussed. We will end by presenting our views and recommendations on the most important and exciting avenues toward more efficient and automated large-scale metabolite annotation and identification workflows.

2 General description of annotation/dereplication strategies and databases

2.1 Annotation/dereplication strategies

A pure compound isolated in a typical NP bio-guided workflow needs to be identified. At this stage, the NP can either have been previously reported and is a “known NP” or has never been described and is an “unknown NP”. In the former case, effort must be directed to reuse already available knowledge from the literature,[31] while in the latter case, the elucidation of its structure must be carried out *de novo*. The rapid identification of known molecules, previously defined as dereplication, is normally the process demanding the least effort, provided that the corresponding spectral data are available and easily searchable. The same applies not only to isolated compounds but also to all NPs within complex mixtures, such as crude extracts or enriched fractions, for which spectral data can be acquired.

In this context, spectroscopic methods are used to obtain indirect partial information about the structures of all metabolites (Figure 1). A compound isolated from two biological matrices is expected to produce identical spectral descriptions under identical conditions. However, in cases of complex mixtures or enriched fractions, such spectral descriptions may overlap, thus complexifying the dereplication process. The question of the extent to which identical spectral signals constitute a proof of compound identity is still pending.[32] The answer depends on the spectroscopic method with which the comparison is achieved; often, a single type of spectral data is not sufficient. The collection of analytical results from various techniques thus provides a more reliable identification tool than the results from a single one. The choice of such methods results from a balance between various criteria, such as accessibility, ease of sample preparation, sensitivity, degree of reproducibility, of information content, ability to provide a quantitative response, etc. Two analytical methods, namely, nuclear magnetic resonance spectroscopy (NMR) and high-resolution mass spectrometry (HRMS) - usually hyphenated with liquid chromatography (LC) as LC-HRMS - are commonly used (see sections 3 and 4) for the rapid annotation of NPs in extracts.

Figure 1

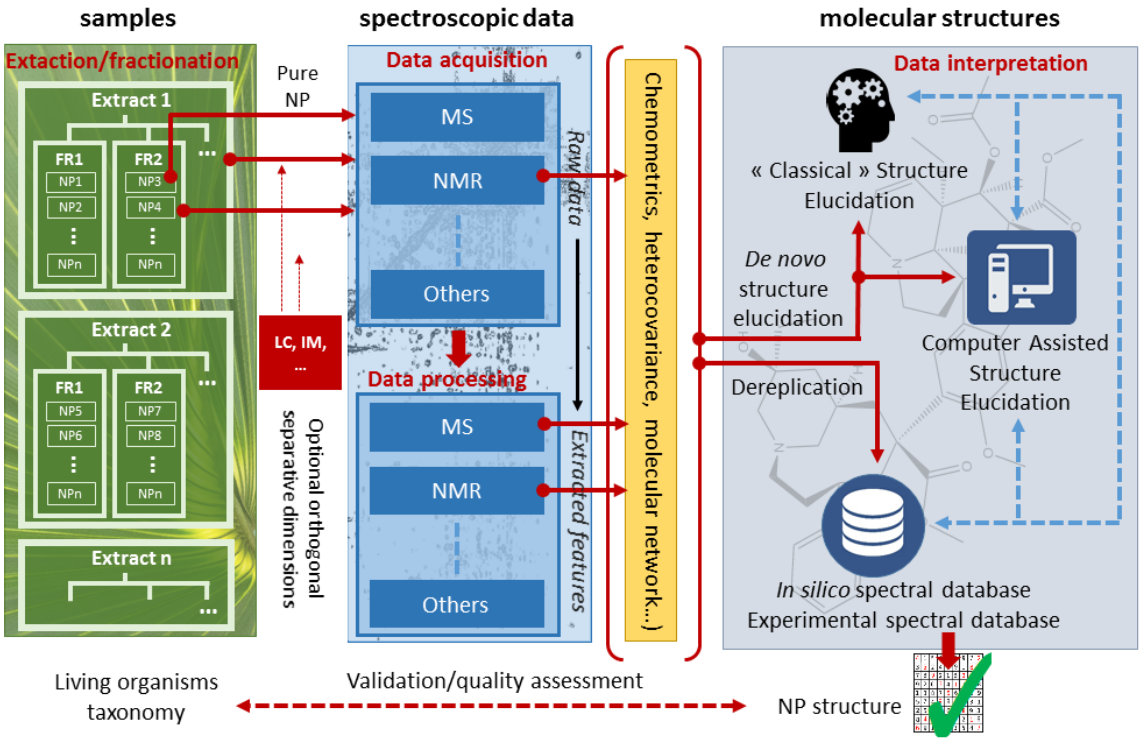


Figure 1. Schematic workflow of structure elucidation/dereplication in natural product chemistry. The principal task consists of connecting the space of samples such as extracts, chemically simplified fractions (FRs), or isolated compounds (NPs) (left panel) and the space of molecular structures (right panel). Extracts are obtained by different extraction processes that lead to complex mixtures of NPs with given physicochemical properties according to the nature of the solvent used. Fractions and pure NPs are obtained after single or multiple preparative chromatographic steps. This task is achieved by a combination of physicochemical spectroscopic methods, mainly MS and NMR (central panel). "Others" indicate additional methods, e.g., X-ray diffractions for pure NPs, LC–ECD for fractions or extracts. When a mixture of NPs is submitted to spectroscopic analysis, often an orthogonal analytical separation method is used prior to spectral acquisition (liquid chromatography, ion mobility, etc.). The space of physicochemical spectroscopic data is divided into two subspaces: (i) acquired "raw data" (e.g., FID time domain data in NMR, LC–MS raw data files) and (ii) "processed data" (e.g., NMR spectra expressed in Hz/ppm calculated by the Fourier transformation FIDs, peak picking of MS features, and combination of related MS and MS/MS spectra in LC–MS). Molecular structure determination results from data interpretation through two different strategies: de novo structure elucidation or dereplication. The latter is generally computer-assisted for the database search step, whereas the de novo approach resorts from manual or computer-assisted strategies interpretation of the spectroscopic data. Connection between NP mixtures and the space of molecular structures may involve the use of chemometrics for deconvolution purposes for finally generating composition information on extracts or fractions. Consistency of the structural data generated are checked based on taxonomy and the known biosynthetic pathways of the organisms studied.

Ideally, the set of spectroscopic data acquired for every newly described compound, along with structural data, should be preserved for future comparison purposes. Such raw spectral data should be as close as possible to the acquisition format, provided that this format is exchangeable. Initiatives have recently been proposed to the scientific community to preserve data in this way,[33] as for example free induction decay signals (FIDs) in case of NMR. Most often, however, only extracted spectral parameters are preserved, such as chemical shifts, coupling constants and relative peak area values for NMR. Spectral parameters are supposed to be sufficiently informative to represent the original data and therefore allow the identification of compounds based on the comparison of their parameter

values.[34] Spectroscopy-based dereplication thus reduces the search for already known compounds whose spectral parameters match those of an unknown compound. Ideally, each newly reported compound should have, at least, its well-defined and preferably computer readable molecular structure, spectral data and spectral parameters preserved in a **publicly accessible DB** so that dereplication can be carried out with minimal effort.

Ideally, for NP dereplication, each DB should contain searchable structure information and link to its biological sources as well as all possible types of spectral information; if possible, these data should be deposited in raw format and universally readable by open-source software. However, in reality, this situation is far from ideal, as most existing DBs display partial structural/spectral information and are often only restricted to a limited number of NPs (Table 1).

Table 1. *NP databases which can be used for dereplication purposes are listed in alphabetical order. This list was compiled from previously published ones.[19, 35] Some databases that focus on the constitutive metabolome and lipids were added in this table when accurate spectral information is provided. The columns provide a large range of DB properties that facilitate the reader to choose the correct DBs for dereplication of metabolites in their sample type with specific spectral data. Also, the presence of in silico (simulated) spectral data is indicated that could assist in tentative structural assignments. Finally, the availability of an automated search (such as API) is indicated; something relevant for uptake of the DB in a computational framework for automated metabolite annotation.*

Name	Type of NP	Number of entries	Publically available	Complementary information*						Experimental data	Simulated data	SOAP or API for automated search
				MS/MS spectra	UV spectra	NMR spectra	Biological sources	Pathway	Functional info			
AntiBase[36]	Specialized metabolites	>40,000	No	Yes	Yes	^{13}C	Yes	No	Yes	Yes	Yes	No
BMRB[37]	Constitutive and specialized metabolites	1,800	Yes	No	No	Yes	No	No	No	Yes	No	Yes
CH-NMR-NP[38]	Specialized metabolites	30,500	Yes	No	No	Yes	No	No	No	Yes	No	No
ChEBI[39]	Constitutive and specialized metabolites	55,000	Yes	No	No	No	Yes	No	Yes	No	No	Yes
DEREP-NP[40]	Molecular	65	Yes	No	No	No	No	No	Yes	No	No	No

	features											
Dictionary of Marine Natural Products[41]	Marine specialized metabolites	>25,000	No	No	Yes	No	Yes	No	Yes	No	No	No
Dictionary of Natural Products[7]	Specialized metabolites	275,000	No	No	Yes	No	Yes	No	No	No	No	No
E. coli Metabolome Database (ECMDB)[42]	Constitutive and specialized metabolites	3,800	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No
FooDB[43]	Metabolites present in food stuff	26,600	Yes	Yes	No	Yes	Yes	No	No	Yes	Yes	No
Global Natural Product Social (GNPS)[44]	Specialized metabolites	-	Yes	Yes	No	No	Yes	No	No	Yes	No	No
Golm Metabolome[45]	Constitutive and specialized metabolites	3,500	Yes	No	No	No	Yes	No	No	Yes	No	Yes
Human Metabolome database (HMDB)[46]	Constitutive metabolites and lipids	42,000	Yes	Yes	No	No	Yes	No	No	Yes	No	Yes
KNApSAcK[47]	Specialized metabolites	51,000	Yes	No	No	No	Yes	No	No	No	No	Yes
Madison Metabolomics Consortium Database[48]	Constitutive and specialized metabolites	20,000	Yes	No	No	Yes	Yes	No	No	Yes	No	No
MarinLit[49]	Specialized metabolites	50,000	No	No	Yes	^{1}H	Yes	No	No	No	No	No
MassBank[50]	Various metabolites	-	Yes	Yes	No	No	No	No	No	Yes	No	Yes
European MassBank (NORMAN	Various metabolites	-	Yes	Yes	No	No	No	No	No	Yes	No	Yes

MassBank)[51]												
MetaboLights[52]	Constitutive and specialized metabolites	26,000	Yes	Yes	No	Yes	Yes	Yes	No	Yes	No	No
MetaCyc[53]	Constitutive and specialized metabolites	15,000	Yes	No	No	No	Yes	Yes	Yes	No	No	Yes
MetIDB[54]	Flavonoids	5,700	Yes	No	No	Yes	No	No	No	Yes	Yes	No
MetLin[55]	Constitutive and specialized metabolites	240,000	Yes	Yes	No	No	No	No	No	Yes	No	No
MFSearcher[56]	Constitutive and specialized metabolites with a focus on flavonoids	-	Yes	No	No	No	No	No	No	No	No	Yes
MoNA[57]	Various metabolites	70,000	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes
MZedDB[58]	Constitutive and specialized metabolites	-	Yes	No	No	No	No	No	No		No	Yes
NANPDB[59]	Specialized metabolites from native organisms from Northern Africa	4,500	Yes	No	No	No	Yes	No	Yes	No	No	No
Natural Product NMR DB[38]	Specialized metabolites	30,500	Yes	No	No	^{13}C	No	No	No	Yes	No	No
NIST[60]	Constitutive and specialized metabolites	13,800	No	Yes	No	No	No	No	No	Yes	No	No
NPatlas[61]	Specialized molecules from microorganisms	>20,000	Yes	No	No	No	Yes	No	Yes	No	No	No
Norine[62]	Non ribosomal peptides	1,200	Yes	No	No	No	Yes	No	No	No	No	Yes

NPCARE[63]	Specialized metabolites with a focuss on antitumoral compounds	-	Yes	No	No	No	No	Yes	No	No	No	No
Plant Metabolic Network (PMN)[64]	Constitutive and specialized metabolites	4,500	Yes	No	No	No	No	Yes	Yes	No	No	Yes
ResPect[65]	Specialized metabolites	3,500	Yes	Yes	No	No	No	No	No	Yes	Np	Yes
Seaweed Metabolite Database[66]	Specialized metabolites from seaweeds	1,000	Yes	No	No	No	Yes	No	No	No	No	No
StreptomeDB[67]	Specialized metabolites from Streptomyces spp.	>4,000	Yes	Yes	Yes	No	Yes	No	Yes	No	Yes	
SuperNatural II[68]	Constitutive and specialized metabolites	326,000	Yes	No	No	No	No	Yes	Yes	No	No	No
UNPD-ISDB[69]	Specialized metabolites	170,000	Yes	Yes	No	No	No	No	No	No	Yes	Yes
Yeast Metabolome database (YMDB)[70]	Constitutive and specialized metabolites	16,000	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	No

*complementary data is usually not available for all DB entries.

The role of these DBs is not limited to the acceleration of dereplication but constitutes the basis for the development of *spectral parameter prediction*, which in turn is needed for dereplication and *de novo* structure elucidation. Such tools may then enlarge existing experimental spectral DB by adding simulated spectral (*in silico* data) (see sections 2.2 and 3.2). Today, it is relatively easy to find catalogs of NP structures but more difficult to find DBs in which each structure is associated with corresponding spectral data tables or raw genuine spectroscopic information. Table 1 provides some of the DBs containing spectral data for either only NPs or those that are parts of larger sets of compounds.

In such DBs, information about the biological origin of NPs provides additional information for metabolite annotation. As far as the phylogenic classification of living organisms reflects genetic peculiarities and the production of specialized metabolites is under genetic control, preserving access to the relationship between a compound and the classification information about the source organism is key. Chemotaxonomic considerations can indeed be decisive in the identification process of an unknown compound. Ideally, an NP DB should provide relationships between structures, spectroscopic properties, and a compound's biological origin (as reported in Table 1). Similarly, it is important to provide information about the biosynthetic pathway of fully *de novo* characterized NPs. Ideally, this should help fill the gap between NP structures, their biosynthetic origin and spectral data.[71] Interestingly, recent metabolic pathway reconstruction in the genus *Penicillium* has revealed that NP pathways within DBs were much more available than expected.[72]

As far as the search for **biological activity** is the underlying motivation for undertaking NP research, associating biological activity with compound structures is also an important feature of NP DBs. Indeed, NPs with specific biological activity can raise interest in the scientific community, and having the means to establish a link between various NP structures and their activities, if well documented, may provide ways to infer the potential pharmacological effects of given compounds and thus considerably extend the significance of a given metabolic profiling investigation.[73]

2.2 Toward *in silico* databases?

Currently, no single exhaustive DB exists where all reported NPs are listed with exhaustive metadata (Table 1). However, increasing numbers of structural DBs have evolved and begun to aggregate experimental and simulated spectra (*in silico* data), while some spectral DBs are enriched with more metadata (i.e., structures reported by SMILES and InChIs code[74] with derived information, such as molecular formulas (MFs), accurate mass, molecular mass, related bibliographic resources, and phylogeny information about the source organism). In this respect, for microorganism metabolite profiling, the NPatlas initiative is worth mentioning because the DB aims to bring together structural and functional information for a wide range of microbial NPs.[61] Access to searchable experimental spectral data remains a key aspect of achieving high-confidence annotation. However, based on these requirements, available NP DBs have various limitations, of which the impact differs from DB to DB:

- **Limitation 1 - DB lack of exhaustivity:** None of the available NP DBs are exhaustive, and several of them need to be browsed to determine all reported NPs for a given organism of interest, which renders this step time-consuming. An alternative solution is to perform searches in more exhaustive DBs that are not limited to NPs and encompass compounds of synthetic origin (ChemSpider,[75] PubChem[76] or SciFinder[77]). Such an approach generally rapidly yields a

comprehensive list of candidate molecules. Most of them are, however, irrelevant for NP scientists, and in this case, the selection of coherent NP candidates is again very time-consuming.

- **Limitation 2 - DB inconsistency:** Differences in chemical identifiers between the various DBs exist when they are compared.[74b] For example, the same name may be linked to different structures in various DBs.[78] Such inconsistencies may drastically increase the time spent gathering coherent information and establishing a final list of plausible candidate structures.
- **Limitation 3 - DB long-term availability:** Maintaining publicly available DBs has a cost, and the lack of funding may lead to their disappearance, as was the case for the Universal Natural Product Data Base (UNPD)[79] or the MoTo DB.[80] Compiling information from the literature to create such DBs comes with a cost and explains why more persistent and exhaustive DBs need to be purchased, such as the Dictionary of Natural Products[7] or AntiBase.[36] The usage of those DBs in combination with open-source workflows is thus problematic. Recently, many open-source DB initiatives have emerged, as is the case for the GNPS platform, which encompasses many DBs relevant to NP scientists and can be searched and downloaded as well.[44]
- **Limitation 4 - DB querying versatility:** All DBs have different intrinsic structural organizations and thus various ways to query them. Some DBs can be searched by MW, accurate mass, molecular formula (MF), substructure and/or spectral features. These differences hamper the performance of simultaneous streamlined searches in a range of DBs.
- **Limitation 5 - Data acquisition inconsistency:** The spectral data reported within DBs may vary significantly. It is important to keep in mind that for MS, fragmentation spectra are strongly dependent on the instrument and fragmentation settings (see section 3.1). Similarly, in NMR, spectral information, even if they are much more consistent across platforms, may vary significantly based on the nature of the solvent and the intensity of the magnetic field used for measurements. This complexifies spectral comparisons; therefore, the spectra of a given compound should ideally be measured in different conditions, and the sample preparation and acquisition parameters should be well documented.

As a consequence of these limitations, searching a DB for known NPs is impaired and the data mining process is considerably slowed down; in contrast, the acquisition of metabolite profiling data is usually rapid. Therefore, we advocate the need to improve the NP DB search efficiency. Although the existence of organism-specific DBs is useful (such as ECMDB[42] for *E. coli*),[35a] it would be highly advantageous if a common template could make their use easier.[73] This effort was already initiated with some DBs from the Wishart group (ECMDB,[42] HMDB,[46] YMDB[70] and FooDB[43]), which share similar architectures.

Another historical issue is that NPs that have been described decades ago may have well-reported

melting points, optical rotations, and UV and IR spectral peak positions but very few details, if any, on their NMR and MS spectra. Furthermore, NMR and MS data are presently scattered in scientific journals in formats that do not favor any kind of automatic extraction for DB-building purposes. Unfortunately, all of these data, if available in the literature, are not always reported in DBs. This, unfortunately, limits the ability to achieve efficient annotation. Therefore, there is a huge gap between the number of NPs reported in the largest DBs (> 300 000 NPs) and the number of NPs with reported spectra.[81]

Thus, the idea of predicting spectral features has emerged as a way to rationally expand spectral DBs from the structures of NPs providing generic tools for compound annotation.[82] Spectral parameter prediction can be undertaken either from first principles (in other words, *ab initio*, by quantum mechanical methods)[83] or from the mathematical (or chemometric) analysis of a set of high-quality, carefully curated reference data. The latter approach has been proven to be very efficient in terms of accuracy and computing time, while the former is more universal and can be applied even in the absence of reliable reference data. Considering the widely used chemometric approach to prediction, each description of the experimental spectroscopic data of a new compound is a possible source of improvement for a prediction algorithm that in turn will facilitate the dereplication or *de novo* structure elucidation of other compounds. Filling the holes of missing or low-quality experimental spectral data with predicted data results in so-called *in silico* DBs. The constitution of such a DB requires "only" a list of molecular structures and a prediction algorithm. Obviously, this prediction technique is strongly dependent on the spectroscopic method, and specific approaches will be detailed in dedicated sections of this article. Currently, such simulated spectral DBs are currently being used to complement existing DBs (i.e., ECMDB,[42] HMDB,[46] YMDB[70] and FooDB[43]). This combined approach allows for a drastic expansion of available spectral information, thus potentially yielding improved annotation.[84]

3 The LC-HRMS/MS side

3.1 Global overview of metabolite profiling by LC-HRMS/MS

3.1.1 UHPLC with different-phase chemistries (polar/non-polar)

Microorganism or plant samples can be profiled by mass spectrometry-based methods either directly or in hyphenation with chromatographic methods. Deciphering the composition of biological samples based only on MS is effective, especially with the recent advances in HRMS and ultra-HRMS. [19] This can be achieved by the direct infusion of samples in the MS interface (DIMS), particularly by using nano-infusion devices.[85] Alternatively, samples can also be analyzed without performing any extraction step using ambient mass spectrometry methods (AMS), such as desorption electrospray ionization DESI or direct analysis in real time (DART).[86]

Metabolite fingerprints can also be acquired from plant or microorganism surfaces using either a laser or an ion beam in imaging MS experiments, which provide not only compositional information but also the spatial locations of metabolites by rastering across an entire surface.[87] All of these methods provide either single MS and/or MS/MS spectra on potentially all metabolites (as none of them are filtered by a chromatography step) and usually have high throughput. Without a chromatography step, the effects of ion suppression become even more pronounced with such direct MS methods, and a significant number of analytes may not be ionized. Such direct MS profiling methods are effective but will not be further reviewed in detail here because they have drawbacks for the structural elucidation process, such as isomers that are not detected separately. Indeed, to obtain a comprehensive survey of the highest possible number of metabolites, i.e., "deep metabolome analysis", MS methods that are hyphenated with a chromatography step are mostly used because natural extracts, in particular, oftentimes hold numerous structural isomers that need to be separated with an orthogonal method to MS for deconvolution. The orthogonal chromatography dimension also helps with "feature reduction" because it allows for deconvolution by assigning MS features to single metabolites, which facilitates the metabolite identification process.[25] As mentioned earlier, this step also considerably reduces the ion suppression issues that frequently occur when dealing with complex biological matrices, such as plant or microbial extracts.[88]

Liquid chromatography mass spectrometry (LC-MS) is thus by far the most commonly used method for the metabolite profiling of NPs. For a recent review of its applications to various plant extract metabolites, the reader is referred to Ganzera *et al*.[89] Gas chromatography (GC) also plays an important role in both the analysis of volatile compounds found in the essential oils of plant extracts[90] and the profiling of primary metabolites after their derivation in metabolomics studies.[91] Capillary electrophoresis (CE)[92] can also be used but is more rare, as it is mainly restricted to charged or ionizable metabolites and is often not used in hyphenation with MS. Due to space restrictions, GC-MS and CE-MS methods will not be discussed here in detail. Interested readers are referred to selected references.[93]

Today, the state-of-the-art LC-MS for metabolite profiling consists of the hyphenation of **ultra-high-pressure liquid chromatography** (UHPLC) with HRMS mass spectrometer detectors that can provide MS and MS/MS spectra with high sensitivity, mass resolution, accuracy, and throughput. To improve the quality of the generated data, the chromatographic and mass spectrometric dimensions must both be optimized, and none of these dimensions should be neglected from neither a mass spectrometry nor a chromatographic viewpoint.

In UHPLC, the reduction of the support particle size (i.e., sub-2 µm) has allowed for a shorter analysis

time and higher separation efficiency at the price of dedicated LC systems that can withstand pressures greater than 400 bars. Since its introduction more than a decade ago, considerable increases in peak capacity, sensitivity and reproducibility have been achieved.[94] A peak capacity exceeding 800 could be attained on typical plant extracts over gradients exceeding an hour, thus demonstrating its use for natural extracts.[95] Sufficient peak capacities in metabolomics applications where throughput has been maximized could, however, be reached in runtimes of less than 10 minutes.

As mentioned before, NPs from plants and microorganisms exhibit a very wide chemodiversity, and it is thus not practically feasible to develop generic metabolite profiling that can encompass the vast majority of metabolites in a single analytical run. From a chromatographic perspective, various stationary phase chemistries must be considered to achieve the best selectivity, while from the MS side, different ionization methods of fragmentation and energies need to be used to generate the richest possible sets of structural information. The majority of studies in natural products research deal with generic linear gradients on reversed-phase (RP) chromatography and electrospray in both positive and negative ion modes, and electrospray (ESI) is the preferred ionization method.[25] Such an approach is efficient but mainly compatible for the analysis of NPs of medium polarity, and this has to be kept in mind if a full naturally extracted metabolome is to be characterized or if expected biomarkers do not have physicochemical properties that are compatible with standard RP phase analysis.

An extensive study on a large set of 120 representative natural compounds belonging to 22 chemical classes was recently performed using LC-HRMS, using 59 different analytical conditions, including 29 columns (RPLC, HILIC and mixed-mode) and several mobile-phase pH conditions to determine what conditions were sufficient for the retention and detection of the vast majority of NPs. This demonstrated that four RPLC conditions were suitable to retain and detect 89% of the set of representative NPs. HILIC offered extended and complementary retention to RPLC for polar compounds, but no universal conditions that would detect the complete set of selected NPs could be highlighted over the entire set of tested columns. The HILIC mode, in particular, was found to be very limited in the extension of the panel of NPs detected.[96]

Supercritical fluid chromatography (SFC) hyphenated to MS is potentially a complementary method to RPLC for profiling extracts with metabolites spread over a large chemical space and has been claimed to be compatible with a far broader chemical spectrum. In addition to being a green chemistry method for the isolation of NPs at the industrial scale, it is also a powerful analytic profiling tool that is now compatible with the use of a sub 2-µm column in a mode of acquisition that has been defined as "convergence chromatography".[97] For ultra-high-pressure SFC (UHPSFC), a set of NPs covering more than 18 logP units on 15 different stationary phases was evaluated, and it was demonstrated that the

technique performed well for the analysis of almost 90% of the selected compounds.[98] For example, this method was used in an offline multidimensional mode with RPLC (UHPSFC C_{18} x UHPLC C_{30}) for the analysis of apolar carotenoids, and chlorophyll characterization in different sweet bell peppers allowed for the determination of 115 different compounds.[99] This latter example indicates that 2D-LC techniques bring orthogonal dimensions in the chromatographic separation of NPs prior to MS detection, thus allowing additional previously non-separated species to be detected. Such an LCxLC mode has been applied for the profiling of natural extracts, but its applications are still scarce compared to those of comprehensive GCxGC methods, which represent the gold standard in the profiling of essential oils.[100] This is probably due to mobile-phase incompatibility problems and the complexity of setups that restrict easy automation for acquisition over a large number of varied samples.[101]

Because natural extracts ideally require additional selectivity in chromatography due to their intrinsically convoluted nature (i.e., the close co-elution of isomeric species), an interesting orthogonal dimension to RPLC is **ion mobility** (IMS). IMS is a post-ionization separation technique that separates ions on a millisecond timescale based on their shape, size and charge. This can be performed using a chamber filled with an inert gas in either a drift tube or travelling wave devices (in low-pressure drift tube IMS, the electric field is only applied in a small region of the drift tube).[102] Such a technique is well compatible for hyphenation between the LC and the MS and adds selectivity to the separation of compounds prior to MS detection. This is usually achieved on fast-scanning instruments such as a time-of-flight (TOF) detectors, and recently, coupling with Orbitrap detectors has also been shown to be feasible.[103] IMS-MS can also introduce additional features, such as collision cross section values (CCS). CCS is also a unique physicochemical property of the analytes that can be used to improve annotation, as discussed in section 7.1. LC-IMS-MS is still an emerging technique for natural extract profiling, but it has been successfully applied to the separation of alkaloids[104] and has shown very interesting potential for the analyses of polyphenols after metabolization.[105]

Regardless of the level of metabolite separation achieved on the chromatographic side, on the MS side, the acquisition should aim for the recording of the highest possible number of MS and MS/MS spectra with the best possible quality. The recording of denoised deconvoluted spectra with high-resolution ion statistics and a sufficient S/N ratio is indeed key for the unambiguous detection of molecular ion species (MS) and the generation of rich fragmentation information (MS/MS). The generation of such MS raw data is a prerequisite for carrying metabolite identification to the best possible level.[106]

The structural information that can be recorded by MS during NP identification are the molecular

weight, the MF, and the fragmentation pattern of each molecular ion species, which is structure-specific and fragmentation energy/ionization mode-dependent: different collision energies yield different numbers of mass fragments with different abundances, and both ionization modes oftentimes have different preferred fragmentation paths that are activated by collision. Such mass spectrometry data are usually very complementary to NMR, as discussed in section 6.

In a typical LC-MS metabolite profiling analysis using atmospheric pressure ionization methods, NPs will be ionized in positive or negative modes, or both, depending on their physicochemical properties.[19] This process is mostly performed by an electrospray interface (ESI), which provides the ionization of the analytes in solution and after final transfer of the ion in the gas phase. Alternatively, ionization can take place in the gas phase after the evaporation of the analytes, as is the case for atmospheric pressure chemical ionization (APCI).[25] This process may simply lead to protonated $[M+H]^+$ or deprotonated $[M-H]^-$ molecules but will also often generate different adducts (e.g., addition of salt, solvents, dehydration, dimerization) that complexify the MS spectra. The presence of several adducts can, however, be important for the unambiguous determination of the accurate mass of unknown NPs. Additionally, the comparison of positive ionization (PI) and negative ionization (NI) MS spectra may help to unambiguously determine the molecular weight (MW). The reader can refer to very interesting tutorial to learn more on these aspects with practical tips and tricks emphasized.[107] When acquiring spectra with high resolution, it is possible to retrieve information about the MF, generally considered as the first important step towards structural identification. Today, a mass accuracy of less than 5 ppm can routinely be obtained on most benchtop instruments, while mass accuracies of less than 1 ppm can be achieved on state-of-the-art platforms, which has drastically reduced the number of possible MFs to consider.[108] Even when such mass accuracies (<1 ppm) are obtained, MF determination remains a difficult task if the number of atoms is substantial and the MW of the analytes is high.[109] This is especially true for NPs that have MWs exceeding 500 Dalton. A good way to reduce the number of possible MFs is to apply heuristic filtering, as discussed below.[110] In NP profiling studies, the MF annotation obtained for an average MW (ca. 400 DA) on a modern instrument in combination with heuristic filtering is generally reliable provided that the determination of the MW is correct. Established software tools such as SIRIUS also include isotopic patterns when available which can further increase confidence in MF assignments.[111]

3.1.2 Untargeted MS/MS acquisition

In untargeted LC-MS metabolomics and deep metabolome analyses, the automated acquisition of MS/MS spectra for all detected metabolites (MS^2) is a prerequisite to generate the necessary complementary structural information to MF.[106] **Tandem mass spectrometry** (MS/MS) is indeed an invaluable experimental tool for providing analytical data to support the identification of small

molecules.[112] The primary strategies used to achieve this task are **data-dependent acquisition** (DDA) and **data-independent acquisition** (DIA).[113] DDA is a mode of acquisition responsive to the signals detected in a given sample: MS/MS spectra are acquired if the selected criteria (e.g., threshold, charge of the detected compound, and dynamic exclusion list) are met in a full scan survey to ensure that the highest possible number of non-redundant dependent MS/MS scans are acquired. This process is efficient and automatically generates MS/MS spectra for each selected molecular ion species and co-eluting species in a narrow *m/z* window detected in the MS survey scan in decreasing order of intensity. However, it is limited by the number of cycles that are necessary to acquire the MS/MS-dependent scan. Depending on the frequency of acquisition of a given instrument, the number of DDA MS/MS spectra will vary, and low-abundance ions in MS will not be sampled in the recording of their corresponding MS/MS spectra, thus limiting the coverage of metabolite annotation for a given sample.

DDA has been successfully applied to plant metabolomics studies, for example, for the metabolite profiling of large series of Euphorbiaceae extracts in the frame of the prioritization of bioactivity in conjunction with MN analysis.[114] This acquisition mode has been applied in most NP studies where annotation was based on molecular networking (see section 3.3).[115]

A significant improvement in traditional DDA was recently proposed in the form of **data-set-dependent MS/MS** (DsDA). The principles of such an acquisition mode are based on repeated injections of a given sample, and the feedback between data processing and data acquisition can allow for the approximately real-time optimization of MS/MS acquisition parameters to obtain nearly complete MS/MS sampling coverage. Using such an approach on a complex mixture with a quadrupole time-of-flight mass spectrometer (Q-TOF), it was demonstrated that the DsDA approach generates significantly more MS/MS events than traditional DDA by temporally isolating data processing from acquisition, thereby maximizing the data acquisition time during the chromatographic gradient by minimizing the competing processing time.[112] It will be interesting to follow future developments in this direction.

Data-independent approaches (DIA), in theory, offer greater MS/MS coverage than DDA, typically at the expense of selectivity or sensitivity.[116] DIA is not biased toward the detection of the most abundant ions in a full scan spectrum because it does not use a selection step prior to fragmentation.[19] For the MS/MS recording of all ions present at any time in the chromatographic separation, DIA can be performed by simultaneously broad-banding all ions (entering the MS at a single chromatographic time point) or by multiplexing a full *m/z* range into smaller *m/z* isolation windows.[117] Various DIA acquisition modes exist depending on the manufacturer. In the so-called "MS to the E" (MS^E), the simultaneous acquisition of MS spectra at low and high collision energies occurs over the entire mass range. In all-ion fragmentation (AIF), multiplexed MS/MS data-independent acquisition (MSX-DIA) occurs. In

Sequential Window Acquisition of all Theoretical mass spectra (SWATH), all ions entering a given mass range are fragmented in *m/z* window increments.[19] This increased coverage comes at a price: the spectra produced by these acquisition modes are convoluted and, in contrast to DDA, the challenge lies in producing clean deconvoluted MS/MS spectra linked to their precursor ions. This is done by associating all ions with overlapping elution profiles at a given retention time (RT) or, more recently, with the usage of LC-IMS-MS, by taking into consideration all ions having the same CCS value and coming from a single LC peak.[118]

An example of a typical UHPLC-HRMS/MS metabolite profiling analysis by DDA and DIA from an artificial mixture of 5 herbs is presented in Figure 2.[119] Up to 18,000 MS/MS spectra were recorded in the DIA versus ca 2,600 in the DDA mode over a generic reversed phase gradient, as displayed by MS-Dial.[120] All spectra (HRMS, MS/MS in DDA and DIA modes) recorded for isoginkgetin a flavonoid present in the extract of the well-known medicinal herb *Ginkgo biloba* are displayed. The comparison of these spectra shows that some characteristic fragments (not all) are common between the DDA and DIA modes and match those obtained in a simulated *in silico* spectra of isoginkgetin obtained by CFM-ID.[121] It has to be noted that richness of fragment information generated is dependent on the type of NP scaffolds analysed. Such information in addition to the MF formula assignment already allows a significant reduction of structural candidates. In this case, the isoginkgetin standard was also analysed in the same conditions and formally identified.

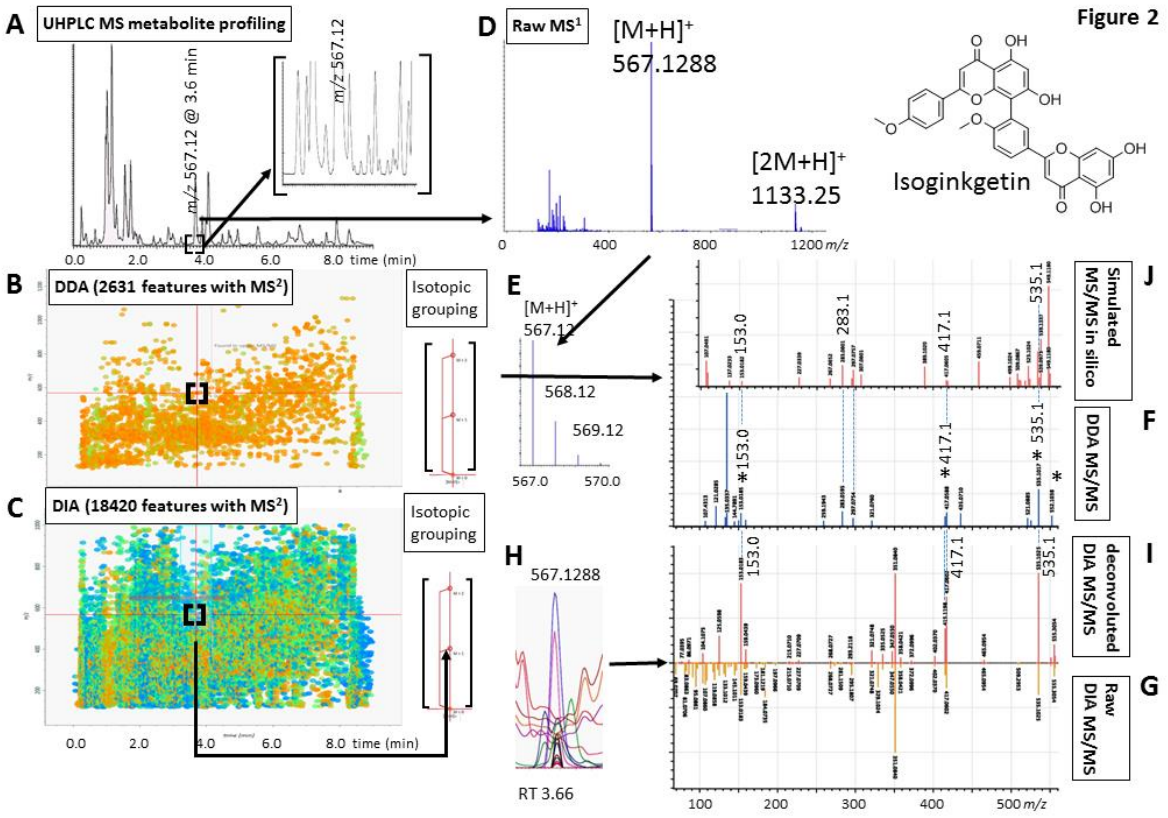


Figure 2. Example of data acquisition in a typical UHPLC–HRMS/MS metabolite profiling of plant extracts in both data dependent (DDA) and data independent (DIA) MS/MS modes. The analysis of a mixture of five plants extracts presenting a broad chemodiversity is shown.119 The HRMS and MS/MS spectra are displayed for a specific feature m/z 567.17 at 3.66 min corresponding to the [M + H]+ of the biflavonoid isoginkgetin (accurate mass, 566.1213; formula, C32H22O10) present in the extract of Gingko biloba, one the extracts of the mix. (A) UHPLC–ESI-HRMS metabolite profile acquired in the in PI mode (m/z 150–1200) on an Orbitrap mass spectrometer on a C18 column (50 mm × 2.0 mm i.d.; 1.7 µm) with a generic acetonitrile gradient 5–95% in 8 min for a broad profiling over a large NP polarity range. (B) Visualization in the form of an ion map (m/z vs RT) of all features having an associated MS/MS with the MS-Dial software120 (2631 features with MS2). On the attached panel isotopic clustering of MS-Dial of all features corresponding to a given analyte. (C) Same plot as part B for data acquired in the DIA mode (18 420 features with MS2) showing an increased coverage of fragmentation data compared to DDA. (D) HRMS spectrum recorded at the apex of the LC-peak at RT 3.66 displaying the features m/z 567.12 [M +H]+ of isoginkgetin as well as its dimer m/z 1133.25 [2M +H]+ obtained with ESI in positive ionization mode. (E) Zoom on the feature [M + H]+ of isoginkgetin showing the accurate mass and isotopic pattern of [M + H]+. This information is necessary to ascertain the corresponding molecular formula (MF): C32H22O10 (calculated for C32H22O10 566.1213, Δppm = 0.5). (F) DDA MS/MS spectrum of the precursor ionm/z 567.12 automatically selected during profiling and fragmented withHCDat three different normalized collision energies (NCE 15, 30, and 45) on the Orbitrap MS analyzer. (G) Raw DIA MS/MS spectrum at RT 3.66 min. All ions are fragmented at three different NCE. (H) Superposition of all ion traces of the fragment ions for selection of ions coeluting in LC at the apex of RT 3.66 min for the deconvolution of the raw DIA MS/MS spectrum in part F. (I) Deconvoluted DIA spectrum at RT 3.66 min associated with the feature m/z 567.12 eluting at the same retention time in the full scan spectra in part D. Comparison of the spectra in parts G and I allows a comparison of the DIA deconvolution, the DDA MS/MS spectrum in part F does not requires deconvolution since a specific precursor ion is selectively selected. (J) In silico simulated spectrum of isoginkgetin obtained by CFM-ID by input of its SMILES structural string. To be noted in the selected example a few fragments are common between the different MS/MS modes and the in silico spectra generated. The richness of fragment information generated is dependent on the type of NP scaffolds analyzed. Such information in addition to the MF formula assignment already allows a significant reduction of structural candidates. In this case, the isoginkgetin standard was also analyzed under the same conditions and its identity was confirmed.

DIA results in more complex datasets but allows more comprehensive information to be obtained than DDA. It also permits the retrospective mining of data because all *m/z* are fragmented within the LC separations and there is no discrimination due to precursor ion selection. However, DIA applications

to NP metabolite profiling studies are still scarce, whilst this mode has been successfully used for qualitative/quantitative proteomics studies.[122] The application of the MS^E mode on a QTOF platform was found to be effective for the characterization of alkaloids of *Hydrastis canadensis*. The spectral data quality obtained using this method was comparable to that obtained by conventional DDA.[123] More recently, the use of DIA with MS^E scan mode enabled the characterization of more than 110 phenolics in green and red oak-leaf lettuce cultivars by UHPLC-QTOF/MS.[124] In DIA, the SWATH mode was mainly applied to proteomic studies; to our knowledge, it has not yet been reported for the microbial metabolite profiling of plants but was successfully applied to small metabolites, such as pesticides in food[125] or drug metabolites.[126] Thus, DIA has very large potential.[127] To date, however, most untargeted metabolome annotations are made from datasets acquired by DDA that provide good-quality spectra with more limited coverage. This is likely a result of the use of deconvolution algorithms that are not yet sufficiently fully able to handle complex metabolomics data sets, although metabolomics deconvolution software tools are emerging now.[120]

3.1.3 Collision-induced (CID) fragmentation and normalization of fragmentation

An important point common to all MS/MS spectra is the influence of the energy given to ions on the fragmentation patterns in **collision-induced (CID) MS/MS** experiments. The extent of analyte fragmentation is indeed dependent on the nature of the analytes, and similar energies of collision across MS platforms are difficult to standardize, unlike in GC-EI-MS[128], where all obtained EI-MS spectra are normalized to a value of 70 ev worldwide, resulting in similar fragmentation patterns for a given volatile.

A typical CID MS/MS experiment is easy to perform and ubiquitous in the MS field: precursor ions are selected and then accelerated to obtain their kinetic energy; they are then allowed to collide with neutral molecules, which results in charged fragments recorded in the MS/MS spectra. The fragmentation spectra change according to the MS platform used, and interested readers can refer to the following review for an excellent overview.[113]

Normalizing energies in CID is a very challenging task. On the one hand, it depends on the internal excitation level of the ions related to the temperature of the ion source, the transfer ion optics and the thermalization of precursor ions prior to the CID. On the other hand, the type of CID process differs, as it will either represent the acceleration to a well-defined kinetic energy distribution prior to collision with a bath gas (in the case of a triple quadrupole instrument, QQQ) or the resonance excitation of selected precursor ions for some set time period by radio-frequency (in the case of ion traps, IT). In an IT system, fragments are off-resonance and will not be further excited once formed, while in QQQ, they may undergo consecutive collisions.[129] Thus, the nature of the fragmentation impacts how a given

MS/MS spectrum changes with increasing energies. Optimizing such energies has been thoroughly investigated, particularly for proteomics, where clear fragmentation rules for deducing peptide sequences have been established.[130] It is much more challenging to propose generic methods for NPs because of their very large chemodiversity and extended physicochemical properties that correspond to very different ionization and fragmentation behaviors.

CID MS/MS spectra thus differ across MS platforms, limiting the efficiency of the search for experimental MS/MS spectra against the MS/MS spectra recorded with different energies. Often, ion intensities will differ, and specific fragments will appear only when a given energy is reached, while at too-elevated energies, other diagnostic fragment ions may disappear. As an example, variations in the MS/MS spectra on different platforms of the fragmentation of isomeric flavonoid *C*-glycosides were compared on TOF versus IT analyzers.[131] Most modern mass spectrometers and acquisition software now allow for a "stepped collision energy" functionality where, during acquisition, several MS/MS spectra of different collision energies are merged to form one recorded spectrum.[132] This can yield information-rich spectra, but choosing the right energies to combine remains challenging because not all molecules perform optimally under one given condition.

One proposed solution to avoid missing characteristic fragments is to ramp collision energies at distinct individual voltages, which are then merged into a single MS/MS spectrum, much like the stepped collision energy functionality mentioned earlier.[113] For improved and/or more generic matching with DBs, it is possible to build libraries of the resulting merged or consensus MS/MS spectra. The concept of **consensus spectra** has already been adopted for quality control in building libraries from ESI-MS/MS data at defined energies. This can be achieved by grouping similar spectra into a one consensus spectrum where peak intensity values are taken as the *m/z* median values of the underlying spectra, and outliers can be rejected.[128] The consensus DDA MS/MS spectra of frequently occurring metabolites were also used to improve the annotation of untargeted quantitative metabolomics analysis in DIA MS/MS.[133]

Instead of trying to determine all relevant experimental parameters by adjusting the CID conditions and determining the excitation energy from there, another approach is to calibrate the instrument by using thermometer ions with well-known fragmentation energetics. To do this, benzylpyridinium ions have been used.[129] If the good implementations of such thermometer ions were available on a wide range of instruments, this could be used to help evaluate fragmentation energies across platforms, which is especially helpful when interrogating DBs acquired with specific parameters that can then be easily compared. Until then, it is important to take good note of the fragmentation conditions of a query in a spectral DB and then manually assess if the match is as expected.

3.1.4 ELSD and CAD for semi-quantitative estimations

Another challenge in the acquisition of metabolite profiling data, which is slightly different than the above identification challenges but also important in relation to the possible biological roles of NPs, is the retrieval of **semi-quantitative information** about the abundance of a given metabolite in the metabolite profiles of NP extracts. Because ionization with atmospheric pressure ionization (API) methods is molecule-dependent, it is not possible to extract such information from the ion abundance data of precursor ions. However, the obtained abundances can still be used in differential metabolomics to evaluate the fold changes of given biomarkers across samples with similar matrix compositions. In particular, in ESI, it is not possible to correlate the responsiveness of small molecules to ESI-MS simply with a single parameter.[134] Multivariate analyses are required to evaluate the ESI response, and some attempts have shown that correlations between different molecular descriptors with respect to the solvent pH and instrumental configuration are possible on a series of nitrogen-containing compounds.[135] This is promising but not yet applicable to general NP research because complex extracts are typically composed of a large variety of molecules. In addition to such efforts in untargeted metabolomics, the prediction of ESI response factors based on NP structures for semi-quantitative estimation is still far from achievable. The quantification of NPs in a given extract is, however, important information and would be ideal to implement to focus the annotation efforts toward the main constituents to thus be able to extract more meaningful data in chemotaxonomy or bioactive NP prioritization studies. For the semi-quantitative estimation of metabolite levels, additional detection methods that have response factors independent of the chemical properties of the analyte are needed.[136] This is the case for the evaporative light scattering detector (ELSD) or Corona-Charged Aerosol Detection (CAD),[137] which represent interesting alternatives with lower sensitivity compared to MS and exhibit linearity issues in their response.[138] Both methods have already been used to profile extracts concomitant with MS.[139] Their integration with MS is still scarce in NP research and requires increasing the amount of extracts injected and defining efficient splitting strategies to divert most of the LC flow in ELSD and CAD without altering the LC resolution in the MS detector.[136] In most LC-MS settings, the UV photodiode array detector (PDA) is coupled online with the MS detector. PDA detectors are very useful for the metabolite profiling of NPs because chromophore-containing NP classes, such as polyphenols, can be well discriminated based on their characteristic UV PDA spectra.[140] A “low” UV wavelength (203-215 nm) can be used for the broad generic detection of all NPs sharing at least a weak chromophore. The detection thresholds are, however, limited by the presence of buffer modifiers. In most cases, formic acid is used but has a higher UV cutoff than TFA, which is ideal for HPLC-UV detection; however, it results in MS signal suppression.[141]

3.1.5 Multistage mass spectrometry fragmentation and fragmentation trees

Next to an optimized acquisition mode, good-quality dereplication and annotation require good-quality precursor ion information to obtain the correct MF assignments and their associated denoised MS/MS spectra. However, even with high-quality MS/MS spectra the metabolite annotation process can still be very tedious in part because the relations between mass fragments cannot always be accurately determined or inferred from MS/MS data. To gain additional structural information about mass features and overcome the abovementioned bottleneck, with ion trap analyzers or ion-trap–orbitrap combinations, MS^3 or MS^n (MS to the "n", as in multistage MS fragmentation where n can be 4 or 5) can also be performed to further enhance fragmentation for deeper annotation using spectral tree approaches.[142] In fact, multistage fragmentation records the fragmentation paths present in molecules that undergo CID fragmentation. By systematically fragmenting increasingly smaller fragments, an extensive spectral tree can be built that can be compared, as a whole or in parts, to other spectral trees. This method was successfully used to discriminate 121 fragmented polyphenol standards and to observe common fragmentation paths in related flavonoid species.[142a] Alternatively, the relations between fragments and thus fragmentation pathways can be inferred from MS/MS spectra, which result in fragmentation trees. These fragmentation trees can form the basis for successful annotation strategies, as discussed later in this review.[143] The main difference between spectral trees and fragmentation trees is that the first are experimentally obtained (acquired) from multistage fragmentation data, whereas the latter are computed from MS/MS spectra based on the MFs assigned to the precursor and mass fragments.

These in-depth fragmentation approaches have the benefit of obtaining detailed fragmentation information and thus can be used to obtain more structure-specific information; however, they come at the cost of more acquisition time, and the number of software tools that can efficiently handle and exploit multistage fragmentation data remains small. For example, the annotation tool MAGMa was the first to exploit the hierarchical information present in MS^n spectral trees.[144]

3.1.6 Processing raw data

Once MS and MS/MS data have been acquired, ideally on the largest possible number of NPs in a given extract, they must be processed prior to entering the workflow for metabolite annotation. Mass spectrometry data come off the equipment in various so-called "raw" formats, which are typically vendor-specific. These raw data refer to the file formats in which the mass spectrometry data are stored; they typically include information about the analytical procedures followed, as well as scans with spectral information, i.e., masses and intensities. To analyze an experiment, e.g., to determine and prioritize the novel chemistries in a set of biological extracts, one must transform the raw data into tables with "features" representing metabolites and their intensities across different samples.

Only then, the prioritization of relevant metabolites can be performed by applying statistical approaches. Many tools have become available to perform this step.[145] Typically, necessary steps include peak-picking (recognizing LC-MS peaks within a sample) and, for differential metabolomics that allow prioritizing differential MS features to annotate, peak alignment can be performed across samples, for which appropriate thresholds for noise signals and signal variations must be chosen. Moreover, if acquired, MS/MS data must be connected to the right LC-MS peaks and the corresponding precursor ions.

Most vendors offer vendor-specific analysis toolkits that are able to visualize LC-MS runs, process mass spectrometry data, and perform metabolite annotations. Although the quality of many of these tools is good, it is expensive (and, for many labs, impossible) to reproduce specific data analysis steps when using proprietary software tools belonging to specific instruments. Furthermore, this results in data that are difficult to share between labs, to extend DBs with, and to improve annotation with in a generic manner. However, there is a strong need within the NP research community for data sharing and reuse. Therefore, various tools have been introduced in recent years that work using the open-source principle. For example, when using data mining strategies[132, 146] (see sections 3.1.2 and 3.1.3), conversion into an open format is a prerequisite. In this way, data can be effectively reused and reanalyzed by the scientific community.

The latest non-textual standard format is mzML,[147] which takes its key strengths from its predecessors mzXML and mzData. Although it has now been around for more than 5 years, its predecessors are still also in use. Alternatively, there are text-based formats, such as the Mascott Generic File format (MGF) and the NIST-originated MSP file format, which are widely used in proteomics as well.[148] These text-based formats have the benefit of being human-readable; however, despite being called “formats”, their exact contents and syntax may vary slightly, which can sometimes cause tools to hick-up during file parsing. It is therefore recommended to check the contents of converted and downloaded files for any obvious conversion artefacts because, they may impair MS annotation accuracy.[149]

A key step in using recent data analysis tools is thus the conversion of raw data into an open format. Currently, the most widely used tool available to convert raw data formats into open formats is ProteoWizard.[150] This is a very versatile tool that works from both the command line as well as through a graphical user interface (GUI), i.e., as an application that can usually be controlled with a mouse. There are some limitations, i.e., the vendor-to-open format conversions can only be performed on a Windows platform. However, most open formats can be converted into each other on both Windows and OSX platforms.

Once these data are converted, a plethora of tools is available to convert chromatographic spectral

data into tables of “spectral features” (i.e., unique mass and RT combinations) with mass intensity information (i.e., area under the curve of the extracted peak) or to perform data mining with MS/MS data (see section 4.4). If LC-HRMS/MS experiments were performed, the mass spectral scans store not only the mass fragments and intensities but also information about the precursor ions. Ideally, one MS/MS scan is uniquely tied to one mass feature discovered by the LC-MS peak-picking software tool; however, in practice, not all mass features are selected for fragmentation by DDA analysis, and some are selected multiple times. Thus, good matching between full scan and fragmentation spectra is important for the optimal use of these data.

In addition, statistical suites can perform comparative analysis to identify significantly enriched features in particular data sets when chemical markers or biomarkers must be identified in differential metabolomics studies.[151]

After MS and eventually MS/MS data are exported, mined (see below) and/or statistically analyzed, potentially interesting biomarkers or new bioactive NPs can be discovered. It is, however, important to go back to the raw data to assess the LC-MS characteristics of the mass features belonging to the identified biomarker. Indeed, the peak-picking process may alter the genuine raw data, and such checks are still needed to verify, for example, the co-elution of biomarkers or NPs in extracts by extracting selected features. Moreover, if independent data sets are available, checking for the presence of the identified mass feature can also strengthen and support this discovery.[151]

3.2 Dereplication based on LC-HRMS/MS

Directly after metabolite profiling, the process of metabolite annotation, which is based on the untargeted HRMS and MS/MS data that are acquired, usually begins. From a general viewpoint, such annotation heavily relies on DB searching; however, depending on the goal behind the profiling, these strategies may differ and follow each other up. Clearly, targeted metabolomics approaches do not require such DB searches, and if the samples are well known and contain target NPs, targeted acquisition is favored; this would also enable quantification, if required. In untargeted experiments, however, the first and fastest strategy to achieve the accurate identification of observed peaks is based on an in-house DB query. This first dereplication strategy of (1) “**Targeted manual dereplication**” involves comparing the obtained RT and MS information with those of previously isolated compounds that were analyzed using the same conditions. This strategy is rather simple and has the advantage of identifying various molecules with a very high degree of certainty (level 1 in metabolomics, see section 5). However, this strategy is limited to analyses of previously studied and known compounds. This has been done, however, at a rather large scale, as exemplified by the WEIZMASS library of NPs.[152] When a metabolite is newly observed at a laboratory in a given profile, other advanced LC-MS-based

annotation strategies are instead needed to allow any NPs detected in a mixture to be identified.

These advanced strategies can be grouped into three categories, which correspond to (2) "**Automated annotation against known compounds**", (3) "**Suspect analysis using known spectral features**" and (4) "**Extensive annotation of all compounds with generation of structural hypothesis for unknowns**". These strategies differ drastically considering the exhaustiveness of the DBs queried and their overall workflow. The first strategy is querying the DBs containing MS fragmentation spectra. The second strategy requires knowledge of the expected spectral features of the molecules of interest and then aims to find all related molecules in the mass spectral data to further annotate their structures, and the last strategy is based on searching, as exhaustively as possible, the NP DBs that mostly contain structures (most of which lack spectral data). In addition, while the first strategy focuses on the information available in DBs, the second one eventually tries to interpret all spectral information obtained from the LC-HRMS/MS profile as thoroughly as possible to reduce its structural possibilities. If successful, this latter strategy will end with only one structure for a given peak. Unfortunately, as mentioned earlier, the success of this annotation is directly related to the presence of the compounds in the DBs and the effective ranking of candidate molecules in the case of multiple possible structures.

The first strategy, (2) **"Automated annotation against known compounds"**, was developed considering the difficult and time-consuming nature of the last strategy. This strategy is highly focused and consists of the direct comparison of the acquired MS/MS spectra with a DB containing fragmentation spectra. In addition, filtering based on precursor ion masses allows one to rapidly focus on accurate spectra. This search, however, must be achieved within the different experimental MS/MS DBs that are usually available online (see Table 1), as is the case, for example, with ReSpect DB.[65] Such an approach remains very time-consuming due to the large number of DBs to query. However, many of them can be exported mostly as .MGF files, as it is the case for GNPS[44], which allows for fast and automated searches using dedicated tools (i.e., TREMOLO[153], GNPS[44]). Such a strategy efficiently highlights MS/MS spectra with accurate matches; however, most of the queried spectra often remain unannotated, mostly due to insufficient DB coverage, particularly for NPs. The current trend to overcome such limitations is the usage of extended *in silico* DBs instead of restricted experimental ones. For example, Allard *et al.* created the ISDB based on a large NP DB containing only structures.[79] All corresponding MS/MS spectra were simulated using a CFM-ID *in silico* MS fragmenter[154], and this generated a massive DB of more than 170 000 spectra, which can be used for improved annotation.[69] To date, this remains the widest available *in silico* DB.[113] Such simulated spectral DBs are currently being used to complement existing experimental DBs (i.e., ECMDB,[42] HMDB,[46] YMDB[70] and FooDB[43]). This combined approach allows one to drastically expand the amount of available spectral information, thus yielding an improved selection of appropriate annotation.[84]

(3) **"Suspect analysis using known spectral features"** was introduced to reduce the search space in the extensive LC-HRMS/MS data sets to molecules of interest, such as the derivatives of molecules in waste water.[155] In natural extracts, this can be done by restricting the search of NP structures to a given species or genus (literature/DB search) and generating *in silico* a restricted set of corresponding MS/MS fragments and a corresponding list of MF.[156] Using known spectral features, such as masses or mass fragments from expected molecules (suspects), the data are queried for spectral data that contain those features. Then, the annotation process can be dedicated to a subset of the entire data set that can usually be related to certain compound classes, thereby saving both time and effort. Tools such as MS2Analyzer have capitalized on this idea to find specific user-defined mass fragments, neutral losses, or mass differences in spectral data.[157] Such strategies are thus effective but are reliant on the pre-knowledge of fragmentation behavior and/or well-curated NP databases with well-documented biological sources that are currently not widely available for large amounts of NPs.

The third strategy, (4) **"Extensive annotation of all compounds with generation of structural hypothesis for unknowns"**, consists of a step-by-step annotation following a more historical workflow (which was previously used when no large MS/MS DBs were easily available). This allows us to obtain more information, even though accurate annotations or identifications cannot always be achieved. The general workflow consists of 5 steps for each peak of a chromatogram with its associated MS and MS/MS spectra:[19]

- **Step (1)** consists of the interpretation of MS spectra to search for adducts, isotopes and neutral losses. This analysis is generally not too complicated to perform manually for a few spectra, as ESI mostly produces molecular ion species that appear in the form of single or multiple adducts, such as $[M + H]^+$, $[M + Na]^+$, $[M + H + CH_3CN]^+$ (if acetonitrile is used as the solvent), and $[M + H - H_2O]^+$ in positive ion mode (PI) or $[M - H]^-$, $[M + HCO_2]^-$ and $[M - H + CO_2]^-$ in negative ion mode (NI).[158] This step is crucial for the determination of the MW of the detected molecule and the accurate determination of its mass. Additionally, the comparison of different ionization modes (PI or NI) may also help to unambiguously determine the nature of the molecular ions recorded and their accurate masses. If automatic peak-picking is performed prior to such spectral interpretation, which is clearly important for higher-throughput metabolomics studies, some tools can be used to automate this interpretation, which use a larger combination of adducts to propose the most likely molecular ions. Some examples of these tools are CAMERA,[159] mz.unity[160] or the adduct/isotope/complex search algorithms from MZmine 2.[161] Unfortunately, sometimes, only one ion is observed, which corresponds to an unidentified adduct. In such cases, all typically observed adducts should be considered for the next step.

- **Step (2)** consists of the MF determination of the detected ion based on the MS information of its mass and spectral accuracy, heuristic filters and MS fragmentation pattern consistency.[109-110, 162] This latter principle relies on the possibility of determining the MF of all fragments observed in MS/MS within the limit of the MF of the detected precursor ion.[163] This strategy clearly improves the MF determination accuracy, particularly when MS^n is achieved (at least MS^3).[163] Many types of software can be used to ascertain MF, such as those dedicated to specific MS instruments or more generic ones (e.g., Sirius[111, 164] or MZmine 2[161]), most of which take MS/MS data into consideration. Such tools, however, perform even better if the possible atoms present in the ionized molecules are accurately set; these are mainly CHONPS for many of the NPs but can also be CHO only, for example, for polyphenols.[165] Considering the isotopic patterns of some atoms can also drastically improve their detection, for example, the isotopic distributions of Br, Cl and S show clear, unusually intense [M+2] isotopes, which allow their detection and can therefore be added to the "possible atom list" used during MF determination.[166] The detection of such halogenated compounds is currently mostly achieved manually; however, it is possible to highlight all halogenated peaks within an LC-HRMS chromatogram automatically.[167]
- **Step (3)** consists of searching the MF within available DBs (Table 1) to obtain putative annotations (a list of possible structures). Due to the large number of available DBs, this step is one of the most time-consuming steps in the entire process. To speed this process up, the use of more generic DBs, such as PubChem,[76] can be considered; however, the number of putative annotations related to NPs is mixed within many synthetic compounds, which can possibly result in the more complex determination of the accurate annotation. It is interesting to note that step (2) may sometimes be skipped by searching directly within the DBs for an accurate mass (after correcting it based on the detected/considered adduct). Such a faster strategy, however, may lead to a larger number of putative annotations. This search can be restricted to NPs only when performed in the DNP, which is a proprietary DB that is commonly used in NP laboratories.[7]
- **Step (4)** consists of reducing the number of putative annotations based on taxonomical information.[81, 168] It is possible to reduce the number of selected structures in later steps based on the biological matrices from which the compounds were obtained.[78, 156, 168a] For example, it is obvious that, in the case of fungal extract profiling, plant reported metabolites matching MS should probably receive a lower candidate score (or even not be taken into consideration) for the annotation. However, such comparisons remain largely manual, even though this

information is available in some DBs (Table 1). Efforts to include chemotaxonomy weighting are currently in progress to automate such tasks.[78, 169]

- **Step (5)** consists of using the acquired MS/MS spectra to determine the most likely structure among those generated after steps 1-4. First, it is of high interest to look for available spectra within DBs or literature data. Historically, when no fragmentation spectra were available, the manual interpretation of an acquired MS/MS spectrum was used to help with its structural determination.[170] However, recent developments in *in silico* fragmentation[171] have allowed for the determination of the appropriate annotations among all hypothetical structures. CFM-ID[154, 172] generates *in silico* fragmentation spectra that can be compared with an acquired MS/MS spectrum. Such a tool systematically breaks up molecules into possible fragments using various algorithms for manual or automated comparisons with experimental data. Various other tools, including MAGMA,[172] MetFrag,[173] and MS-Finder,[174] search structural databases for possible candidate molecules and then search for possible fragments in them that match the experimental data; then, they use different scoring algorithms to rank the found candidates. It is important to note that most of these tools consider the fragmentation of $[M+H]^+$ or $[M-H]^-$ adducts (in which $[M+H]^+$ is usually more relevant due to the larger library of positively charged ion fragmentation, which represents a larger training set for fragmentation algorithm development); thus, fragmentations related to other adducts may not be accurately determined. Additionally, similar approaches querying structural databases for spectral to structural consistency without *in silico* MS fragmentation (using tools such as CSI:fingerID[143] or ChemDistiller[175]) represent efficient alternatives. Strategies for the *in silico* interpretation of MS/MS spectra were recently reviewed in detail by Hufsky *et al*.[171a] and Blaženović *et al*.[171b]

This last annotation strategy can also begin with a direct query of all MS/MS by MN approaches and filtering with MF information, as discussed in detail in section 3.3.

The entire dereplication workflow described above is clearly a highly time-consuming process that needs to be improved by integrating these different steps into a pipeline that is at least semi-automated. Even if such a workflow is close to being accessible to experts, it always provides many lists of possible candidates that need to be manually curated. Thus, even if the *in silico* annotation workflow efficiency is solved in terms of “calculation duration” and “good candidate ranking,” its curation steps will still need to be performed, which will represent another time-consuming manual task, particularly if the goal is deep metabolome annotation.

Currently, the number of reported NPs is far less than the size of the theoretical NP chemical space.[176]

Therefore, annotation using traditional approaches is generally impaired by the absence of plausible candidate structures, i.e., many MS/MS spectra with associated MF do not yield any plausible database candidates.[81] To overcome this limitation, various research teams have started to create structural DBs with extended chemical spaces while following rules that take chemical consistency into consideration. For example, Jeffryes *et al.* expanded the chemodiversity of KEGG[177] or YMDB[70] by applying expert-curated enzymatic reaction rules by creating the MINEs DB, which contains more than half a million structures.[178] Such biosynthetic rules were recently proposed as a DB of chemical modification references, i.e., RetroRules,[179] which is able to propose lists of biosynthetically relevant compounds based on entries of structures (e.g., SMILES). A similar strategy could be adopted in combination with automatic *in silico* spectral match, such as that in MetWorks,[180] which tries to propose new structures based on accurate, biosynthetically relevant chemical modification in combination with spectral similarities to a known compound entry. Alternative approaches were developed by other teams working with lipids, who took advantage of the consistency of lipid structures to create extended DBs containing a large variety of compounds with similar skeletons but very large variations in fatty acid moieties. Such an approach can be either applied using a specific *in silico* MS/MS DB (i.e., LipidBlast[181]) or using preliminarily defined interpretation rules (i.e., LipidXplorer[182]). To a larger extent, even if NPs represent a wide chemical space, many of them share common scaffolds. As such scaffolds generally share common MS fragmentation mechanisms, precise interpretations of their fragment ions should lead to an accurate structural determination.[183] Hence, the abovementioned strategies used to expand the available set of candidate molecules following logical fragmentation rules should also make sense for many classes of NPs. In fact, such a strategy for the NP identification of unknown polyphenols or glycoalkaloids was recently applied in plant extracts to evaluate its potential.[184]

However, when considering the above annotation strategies, it is important to keep in mind that, as mentioned for the MF determination, using orthogonal information usually improves the annotation process (see section 7). In this context, it remains important to consider UV-visible PDA spectra when they are available, as they can be very informative by indicating the presence of characteristic chromophores.[139a, 185] As discussed in section 6, MS/NMR correlation approaches may also be very efficient, i.e., if they are generally limited to the main NPs in extracts. Ultimately, hyphenated strategies to NMR (see section 6.1) or organic synthesis approaches can yield certainties about metabolite structures. Methods to correctly filter the searched structural space or the usage of the orthogonal detection of MS and MS/MS can thus be efficiently used to improve the degree of confidence of metabolite annotation (see section 7.1).

3.3 Clustering of molecules into molecular families by Molecular Networking

The full elucidation of molecular structures is a common challenge when analyzing complex mixtures

with mass spectrometry. As discussed in section 3.1, tremendous advances in technology have now resulted in information-rich mass spectrometry data files in which, for a complex extract, more than thousands LC-MS features with their MS/MS data are typically recorded in a LC-MS/MS profile.

To find known and novel chemistries, comparing the MS/MS spectra within and between extracts is the cornerstone of mass spectral analysis. The main reason for this is that spectral similarities are often representative of structural similarities.[183] This can be understood as follows: minor modifications to a molecule, i.e., the addition of a methyl group, leave the overall structure intact. As this typically does not impact the major fragmentation pathways that together result in a fragmentation spectrum, structural similarities can thus be inferred based on spectral similarities. The following paragraphs explain how this concept was adapted in NP research and how computational tools were developed to allow its use at a large scale.

It has long been recognized that NPs share scaffolds that result in a group of structurally similar NPs, i.e., an NP **molecular family**. Historically, such scaffolds were highlighted by the manual search of common fragment ions across different MS/MS spectra.[186] Moreover, not only entire molecules or larger scaffolds but also smaller parts, i.e., substructures of those scaffolds, often result in similar fragmentation patterns, even when they are acquired from different molecules. However, until recently, it was not possible to exploit this concept at a larger scale.

Fragmentation spectra can now be compared within a complete LC-HRMS/MS profile and between various profiles. Various strategies have been employed to compare fragmentation spectra at a large scale by comparing mass fragments[187], neutral losses,[188] or both.[189] Thus, these approaches provide novel ways for researchers to represent and visualize LC-HRMS/MS data by exploiting different aspects of the similarities between fragmentation spectra.

Molecular Networking (MN)[69, 81, 132, 187a, 190] represents the most widely used tool to cluster molecules into molecular families based on their MS/MS spectral similarities, i.e., the more peaks that two MS/MS spectra share, the more similar they are. This is based on the so-called “modified cosine function” that i) looks at shared peaks between two spectra within a user-defined threshold, ii) considers the intensities of mass fragments, and iii) considers the difference between the two parent masses by shifting mass fragments within that difference and checking for improved matches between the mass fragments (see Figure 3). After comparing all MS/MS spectra to each other, each spectral combination is given a cosine score ranging from 0 (completely different) to 1 (identical).

Figure 3

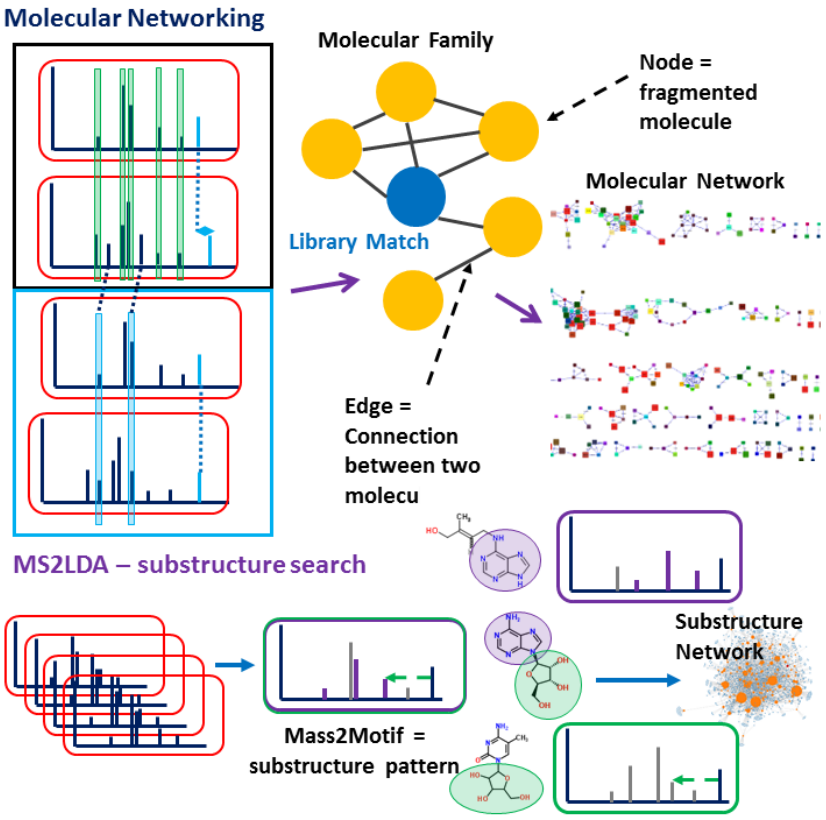


Figure 3. (Top left to right) Key Molecular Networking concepts: all MS/MS spectra within a sample or across different samples are compared based on the similarity of mass fragments. The parent mass difference (indicated by blue dotted arrow) is considered by shifting both spectra with this m/z value and checking for improved or additional matches that will add to the final similarity score (left panel, bottom). Using all these comparison scores, Molecular Families can be formed (middle panel), where fragmented molecules are the nodes and connections (edges) are present when the similarity score is above a user-defined threshold. Moreover, different layers of information can be displayed on the nodes and edges, such as where molecules result in a library match from reference MS/MS spectra (node in blue). Applying this to all the fragmented molecules typically results in a Molecular Network (MN) consisting of larger and smaller Molecular Families (right panel) as well as unconnected molecules (singletons). (Bottom left to right) Conceptualized MS2LDA substructure search: MS2LDA starts with a large set of MS/MS spectra from one or multiple samples (left panel) and then searches for Mass2Motifs (middle panel). These Mass2Motifs consist of often co-occurring mass fragments and/or neutral losses that can then be annotated by substructures. In the middle panel, one mass fragment-based (purple) and one neutral loss based (green) Mass2Motif are exemplified, where each Mass2Motif is present in a spectrum where its corresponding substructure is indicated in the structure on the left. Both Mass2Motifs are present in one MS/MS spectrum and in the corresponding structure on the right, and both corresponding substructures show how in this case the complete structure can be built from its substructures. Finally, by collecting all of the connections between fragmented molecules and Mass2Motifs, asubstructure network can be formed that connects Mass2Motifs/substructures (orange circles) with fragmented molecules (blue squares), as displayed on the right end.

Then, a “network representation” can be built as follows: fragmented molecules serve as the “nodes”, and they have connectivity (“edges”) to other nodes if they share sufficient spectral similarity. After collecting all nodes and edges, a network can be drawn (see Figure 3). In practice, a researcher can influence the “network topology” using different thresholds and filters, i.e., an edge can be drawn only if the cosine score between two nodes is above a user-set threshold (typically between 0.55 and 0.7) and there is a minimum amount of shared fragment peaks (typically between 3 and 7 depending on the type of molecules in the extract). Very recently, a self-organized topology was proposed as an alternative to networks. The MetGem software allows to generate locally and in a reduced amount of

time a double visualisation of spectral similarity. One is based on the classical cosine score establishment and constitutes the classical "molecular network", in parallel a different representation based on the t-SNE calculations (t-distributed stochastic neighbor embedding) allows to highlight relations between unrelated clusters in the molecular network visualisation.[191] The platform is open-source and is designed to be evolutive. It should become a tool of choice for quick and local molecular networks generation as-well as for the establishment of alternative spectral similarity visualisation within MSMS datasets.

The selection of the correct networking parameters is important when performing MN to ensure that i) the right peaks are correlated with each other, ii) most of the noise is eliminated without losing the signal, as this affects the cosine score, and iii) the cosine threshold is set appropriately. The resulting MN will then contain "Molecular Families", i.e., groups of fragmented molecules that are interconnected, as well as "singletons", i.e., molecules without any connected neighbors (Figure 3). The discovered connections can then be exploited: for example, when one molecule is known, (part of) its structure can be inferred to be connected to other molecules in the network, thereby extending the identification and annotation of metabolites "into the unknown".

The large-scale clustering of molecules based on their spectral data has inspired several annotation tools that exploit the network topology.[69, 84b] Recently, **the Global Natural Product Social Molecular Networking (GNPS**) platform facilitated the performance and sharing of MN analyses.[44] GNPS is a web-based platform that connects both a data storage repository and metabolomics analysis tools and acts as a graphical user interface (GUI) for a number of tools. It is currently used in more than 140 different countries by more than 40,000 different users handling 10^9 spectra, which is indicative of its large scale. Because GNPS also contains a variety of public spectral libraries, including both reference MS/MS spectra and user-annotated spectra, fragmented molecules can be matched to those libraries. Subsequently, positive matches can "travel or propagate" through the MN, as explained above. Recently, peptidic natural product and other natural product annotation software were also made available through GNPS.[192] Annotations from such tools can be further exploited in the MN as well since they also "propagate" within Molecular Families.[84b]

Clustering molecules into molecular families based on their MS/MS spectral similarities has yielded many novel insights.[69, 132, 193] These examples show how large complex data sets can be "digitized", which allows researchers to quickly form hypotheses about which molecules are structurally linked and how they are distributed across samples (i.e., strains, species, or locations). For example, indexing the *Pseudomonas* specialized metabolome revealed the novel related cyclic lipopeptide bananamides **1**-**3**[193e], and comparisons of more ca. 300 Euphorbiaceae species revealed a common trend in their

composition and allowed for the targeted isolation of new bioactive NPs.[114]

In another context, a recent study showed how drug screening in urine samples can be enriched using an untargeted MN-based approach: antihypertensive drugs and metabolized drugs were found as families within complex urine extracts.[132] This yielded information about drug adherence and drug metabolism as well as over-the-counter drug intake, in contrast to targeted LC-MS/MS assays, which look for specific drug molecules and do not monitor drug metabolism and over-the-counter drugs. The importance of data mining in future drug discovery approaches was recently reviewed.[194]

3.4 Substructure search by MS2LDA

NPs share not only large scaffolds that encompass almost the entire molecule but also smaller common building blocks such as sugars, amino acids, or *O*-methyl groups that can also be part of those scaffolds. Having information about which building blocks are present in NPs speeds up the process of structural elucidation, as lists of candidate molecules can be shortened based on this knowledge. MN clusters molecules based on their (larger) scaffolds into informative molecular families, whereby smaller substructures that molecules have in common are often overlooked by MN. In addition, moieties that are typically recognized as "neutral losses", i.e., the difference between a precursor ion and a mass fragment, such as deoxyhexose or ribose in the case of NP glycosides, are not considered in many molecular clustering approaches. Finally, clustering methods typically force molecules into one cluster (i.e., molecular family), whereas many molecules contain more than one recognizable scaffold or building block. Therefore, a novel tool called MS2LDA has emerged, which, inspired by a text-mining approach, searches in an unsupervised manner for substructure fingerprints in MS fragmentation data.[189b]

The LDA in MS2LDA stands for Latent Dirichlet Allocation.[195] This technique was originally developed for text documents, where co-occurring words are oftentimes grouped into "topics" that humans can interpret and annotate, i.e., a collection of "club", "manager", and "stadium" would typically be described as a "football"-related topic. When a large amount of text documents is mined with LDA, the documents are "decomposed" into one or more topics, and it becomes easier to assign them to different categories. By monitoring the appearance of topics over time, one can pick up on "trending topics". MS2LDA is the first implementation of LDA in metabolomics that decomposes fragmented molecules (the documents) into groups (the topics) of mass fragments and neutral losses (the words).

MS2LDA exploits the realization that the same structural features often result in similar fragmentation patterns by searching for co-occurring mass fragments and/or neutral losses in MS/MS spectra (Figure 3). For example, this resulted in a set of over 80 different fragmentation patterns, or so-called Mass2Motifs, that were discovered in reference MS/MS spectra from the GNPS and MassBank DBs

and could thus be validated; additionally, the Mass2Motifs of several amino acids and the loss of carboxylic acid were recognized. Moreover, the losses of hexose, deoxyhexose, and pentose were discovered by MS2LDA and subsequently validated using the structures of GNPS and MassBank entries. Figure 3 exhibits how two substructures present in adenosine can be recognized by their fragmentation patterns in MS2LDA. We also see how two other molecules only share one substructure with adenosine, namely, the adenine moiety or the sugar moiety. It is important to realize that MS2LDA recognizes adenosine as both a purine and glycoside moiety. Thus, when a researcher applies MS2LDA to a glycoside-containing extract, the Mass2Motifs for diverse glycoside moieties such as hexose or deoxyhexose can be expected, which is very helpful in the structural elucidation process. Finally, a network can be constructed in which molecules are connected to all substructures (Mass2Motifs) that were found so that, for example, different types of glycosides and adenine-containing molecules are connected by a shared Mass2Motif (Figure 3).

To facilitate its use, MS2LDA.org was built; here, substructure searches in MS fragmentation data can be performed and the resulting model can be inspected, Mass2Motifs can be annotated, and analysis results can be shared.[196] Since its creation, this web application has been used by over 70 different users who have run more than 300 analyses on diverse sample types, such as bacterial, beer, urine, fecal, plant, and organic matter extracts.

Another MS2LDA implementation is MS2LDA+, where multiple files are subjected to MS2LDA at once.[197] In this technique, Mass2Motifs are coupled between different samples, and based on the "Mass2Motif prevalence", one can quickly observe which substructures are more abundant in each sample. For example, an MS2LDA+-based PCA based on two groups of urine and beer extracts exposed urine- and beer-related substructures. Moreover, after annotating the drug-related Mass2Motifs in urine samples, the prevalence of Mass2Motifs quickly revealed which urine extracts contained particular drug metabolites.

Finally, work is in progress to integrate MN and MS2LDA analyses.[198] In plant metabolomics, large molecular families of related NPs are present, and by mapping Mass2Motifs onto MN subfamilies, the specific modifications of diterpenoid molecules could easily be tracked. To conclude, clustering molecules based on their spectral similarities has been proven to be a successful way of obtaining insights into large amounts of spectral data so that hypotheses can be quickly formed and novel chemistries can be prioritized.

3.5 Toward user-friendly interfaces

To be used by a large community, all of these approaches need to be accessible and easy to use by NP chemists. However, it is clear that many of the steps described in the LC-HRMS part (part 3) rely on a

large variety of tools that can be used together or separately. In this review, these tools are divided into various groups depending on their use in the LC-HRMS/MS data analysis workflow.

Initially, it is important to consider tools that are able to identify LC-MS peaks or MS/MS features from a raw data format (or a generic format after their conversion by vendor software or ProteoWizard[150]). Various tools can be used to extract reliable LC-MS/MS peaks by focusing on LC-MS peak detection. Such tools are mainly XCMS (an R command line interface[199] or a web-based interface[200]) or MZmine2.[161, 201] This latter software possesses an efficient GUI interface, which is clearly an added value in comparison to any command-line tools. In addition, both tools are able to extract the MS/MS data of all peaks from DDA MS/MS acquisition for further analyses of the information in the dereplication pipeline or MN. Alternatively, for DIA MS/MS acquisitions, MS-Dial[120] represents an efficient alternative to extract MS/MS spectra. In parallel to software focused on LC-MS peak detection, which can be considered a more traditional approach, other tools are focused on the direct extraction of MS/MS spectra from raw data. These are generally integrated directly with the creation of MN, as is the case for GNPS.[44] This latter tool represents an easy-to-use web interface; however, the direct use of MS/MS information remains less efficient in creating easily interpretable MN, and using preliminary peak-picking steps drastically improves the resulting MN.[202] It is important to note that currently the creation of a relevant MN with GNPS remains a rather long process because it relies on the selection of appropriate parameters, and each parameter modification restarts the MN calculation. Therefore, use of MetGem could ease the selection of appropriate parameters as they can me modified to provide directly the new MN.[191]

Similarly, various tools exist to achieve the dereplication of compounds within a given LC-HRMS/MS chromatogram depending on the strategy that is used (part 3.2). In the case of “Targeted manual dereplication”, researchers rely on a spreadsheet-based technique and manually interrogate a small local DB. Although this strategy is rather manual, it is very efficient. For other strategies, however, various tools remain accessible. The approaches used for “Automated annotation against known compounds” and “Suspect analysis using known spectral features” sometimes remain largely manual due to the use of specific DB interfaces. However, when the DB itself is downloadable under a .MGF format (see part 3.1.6), it can be rapidly analyzed using TREMOLO[153] (a command line interface that is rather difficult to use) or GNPS[44] (a web-based interface that is quite easy to use). It is interesting to note that in GNPS, a large number of real spectral DBs were implemented for the rapid search of spectral patches, and is also easily possible to add a specific DB.[203] In addition, GNPS can also rapidly query *in silico* DBs to expand the MS/MS DB chemical space.[69] Finally, for substructure searches, the MS2LDAviz web application[196] provides the user with basic MS/MS spectral visualizations but also, more importantly, provides the user with visual information about Mass2Motifs that supports their

annotation, such as feature frequency plots and colored fragment peaks based on the Mass2Motifs.

The last dereplication strategy, "Extensive annotation of all compounds with generation of structural hypothesis for unknowns" (part 3.2), requires more tools to perform the five steps described above.

Finally, it is important that the people developing these tools keep in mind that these tools need to be sufficiently easy to learn and use for the NP chemist community. However, such recommendations seem to have been taken into consideration, as we can learn from various reviews of LC-HRMS/MS tools have provided information about their user-friendliness.[171b, 204]

4 The NMR side

4.1 Global overview of metabolic profiling by NMR

Metabolic profiling in the context of NP studies should provide comprehensive information about the identity and quantity of the low-molecular-weight compounds produced by a living organism. If an organism must be kept alive, as in the definition of *in vivo* NMR spectroscopic imaging, only localized spectroscopy or low-speed HR-MAS spectroscopy should be used for this analysis.[205] Even though these techniques have undergone some development, the sensitivity of NMR allows, at best, the characterization of only the most concentrated metabolites.[206] Going beyond this approach implies the killing of the organism, a process that is likely to produce profound changes in the sample's chemical composition. The killing and quenching of metabolic enzymatic reactions, for example by freeze-drying, is generally followed by the cutting, crushing, and milling of the tissues of interest, as well as the extraction of metabolites by organic solvents of different polarities, which may include supercritical or subcritical fluids, ionic liquids or (natural) deep eutectic solvents.[207] Solvents of broad polarity, such as methanol, are generally used to ensure a relatively large metabolite coverage, and in NMR profiling, at this stage, deuterated solvent can be used directly, thus minimizing the sample preparation steps.[208] Although the resulting extracts are supposed to reflect the metabolism of the living organism under consideration, they may have undergone transformations induced by these physical and chemical extraction processes. Such transformations include the disappearance of "fragile" compounds and the creation of new ones, which are referred to as extraction artifacts. The metabolic profiling of an organism by NMR is thus generally performed by liquid-state NMR on complex mixtures obtained after extraction for early identification purposes. In NP research, NMR is still generally applied on pure NPs obtained after isolation and represents, in most cases, the first step toward structure identification.[209]

4.1.1 Dereplication and *de novo* structure elucidation of new NPs

Fully and accurately reporting the physicochemical and spectroscopic properties of new NPs necessitates their isolation in pure form. However, the work required for the isolation of already

known compounds is considered a time-consuming but unavoidable task.[18] Ongoing efforts are currently directed toward the early identification of known compounds to rationalize this approach. Achieving this task implies the capability of analyzing complex mixtures using NMR. In this context, only high-sensitivity liquid-state NMR is considered here. Dealing with mixtures constitutes a challenge in NMR spectroscopy in terms of resolution and sensitivity.[210] The amount of sample in an NMR sample tube is limited, and the presence of numerous compounds in a mixture limits the amount of each mixture component and therefore limits the detection of the many minor mixture components. The amount of extract may also be a limiting factor because a few mg of dried extract are necessary for measurement in a 5-mm outer diameter tube, while only a few µg need to be injected onto the column for LC-MS profiling. The interpretation of the NMR spectra of a single compound requires the labeling of all spectral peaks, even though the signal superimposition in NMR profiles of the mixture can lead to undecipherable spectra; in such cases, only a few signals can be fully assigned.[211] These sensitivity and resolution issues will be addressed in the next paragraphs.

4.1.1.1 Overcoming sensitivity issues

Overcoming sensitivity issues in NMR is a matter of increasing the signal intensity and decreasing noise intensity. Noise in NMR originates from the thermal agitation of electrons in the metal that constitutes the detection "coil" in the NMR probe. This noise is amplified by the electronic detection chain, which itself introduces additional noise. Reducing thermal noise can be achieved by the reduction of temperature within the detection device and by the careful choice of electronic components. Spectrometer manufacturers have provided cryo-probes and cooled preamplifiers with low-noise-component technology for more than 20 years. The signal that arises from a given sample varies as the square of the intensity B_0 of the static magnetic field. A first B_0 factor originates from the influence of the sample magnetization precession speed on the signal intensity, and a second B_0 factor originates from the population difference of nuclear spin states. Magnet manufacturers have provided increasingly stronger magnetic fields, and combining actual and innovative technologies will allow the resonance frequencies of ^{1}H nuclei to reach 1.2 GHz in the near future.[212] Boosting the population difference by Dynamic Nuclear Polarization is also an active research domain for which practical applications are still rare but promising. [213] Finally, the filling factor of the receiver coil is a parameter whose adjustment leads to a noticeable improvement in sensitivity: a sub-milligram amount of a compound dissolved in 30 µL of solvent and placed in a sample tube with a diameter of 1.7 mm, which is itself placed in a dedicated probe head, produces more signal than it would if it were placed in a standard 5-mm sample tube. Small-diameter, cryogenically cooled coil technology currently represents the best practical method for the study of minute amounts of NPs.[214] Basically, increasing the sensitivity in NMR at a given field has always been a matter of increasing recording time through signal

averaging. Pushing to the extreme the limits of sensitivity in NMR by non-conventional experimental devices constitutes an active research field.[215] The relationship between conventional NMR and recording time will be detailed in the following section.

4.1.1.2 Overcoming resolution issues

In 1D ^{1}H NMR, resolution can be understood in two different ways. In spectroscopy, frequency resolution is usually related to the width at the half-height of a resonance peak, which is expressed in Hz. In NMR this resolution depends on the natural resonance linewidth and is related to the relaxation rate of sample transverse magnetization and to the residual static field homogeneity within the active volume. The inhomogeneities that arise from the introduction of the sample in the probe and its magnetization have to be corrected by a process define as "shimming" and a poor shimming prior to NMR acquisition may strongly lower the resolution. The magnetic field intensity has a very low impact on the latter factor, and the "line resolution" depends on the efficiency of the shimming strategy. Drawing a series of spectra of a compound (for example, sucrose) recorded on spectrometers with increasing magnetic field intensity (e.g., 60 MHz up to 1 GHz) with an identical chemical shift scale clearly shows that the multiplets become increasingly narrower, even though individual lines may have comparable widths when expressed in Hz.[216] This increase in "multiplet resolution" yields the possibility of studying more complex molecules or samples at higher fields. Such considerations do not appear when dealing with ^{1}H broadband-decoupled ^{13}C NMR spectroscopy due to the absence of multiplet structures in spectra. The wide range of chemical shifts in ^{13}C nuclei and the extreme narrowing of the resonance lines (which are most often intentionally broadened for spectral noise reduction purposes) would make^{13}C NMR an ideal tool for the study of complex samples if it were not hampered by the low natural abundance of this nucleus.[216] The removal of multiplet structures in ^{1}H NMR spectra or, equivalently, the recording of "pure chemical shift" ("pure-shift" in brief) spectra, which look like ^{1}H-decoupled ^{13}C NMR spectra or ^{1}H NMR spectra at ultrahigh fields, constitutes a quest for the "Holy Grail" of NMR spectroscopists. Beyond the first approaches that relied on 2D *J*-resolved[217] or 2D constant-time COSY[218] spectra, pure-shift spectra recording strategies that are derived from the pioneering work of Zangger and Sterk[219] on the application of static field gradients during adiabatic pulses constitute an active research field.[220] The actual price to pay for the simplification of ^{1}H spectra is however, a heavy decrease in sensitivity, which makes the practical use of this approach limited to the most concentrated compounds within mixtures.

2D NMR spectroscopy is a way of spreading the NMR signals packed on a single chemical shift axis along an indirect dimension (*F1*).[221] In 2D NMR intensity is plotted as a function of two frequencies F1 and F2. Each frequency axis is associated with one of the two time variables from which the recorded signal depends. These variables are the duration of the evolution period (the evolution time t_1; F1

indirect dimension) and the physical detection period (the detection time; F2 direct dimension). 1D slices of 2D spectra represent simplified parts of complex spectra. This second dimension is most often related to chemical shifts in the 2D chemical shift correlation spectroscopy routinely used by NP chemists for structure determination of pure NPs. Since 2D NMR refer to numerous complex acronyms for the pulse sequences applied, the reader can refer to the review of Breton *et al.* for an overview.[209] The 2nd NMR dimension may also be related to coupling patterns in *J*-spectroscopy, which is commonly practiced in metabolomics, since the 2D *J*-resolved spectra of common metabolites are present in public and commercial DBs.[222] Another way of introducing a supplementary discriminating variable consists of modulating signal intensities according to translational diffusion coefficients. This is performed by an experiment defined as DOSY (Diffusion-ordered spectroscopy). The resulting 2D DOSY spectra allow for the extraction of the 1D spectra of mixture components along the F2 axis, in which each component is characterized by its own diffusion coefficient value, readable in the F1 axis.[223] DOSY represents thus a way to obtain pure 1D ^{1}H NMR of individual constituents of a mixture provided that they had sufficient enough difference in their diffusion coefficients.[224] The DOSY principle was also applied to the recording of 3D spectra that contain a diffusion coefficient axis and from which 2D *J*-resolved,[225] COSY[225] or HSQC[225] planes can be extracted. A NOESY spectrum correlates the chemical shifts of pairs of ^{1}H nuclei that can transfer magnetization through cross-relaxation or chemical exchange. Cross-relaxation originates from dipolar coupling between two nuclei and occurs when their distance through space is short (typically less than 0.5 nm). Cross correlation efficiency depends on the molecular tumbling speed in solution and leads easily to multiple-relayed NOESY correlations, in a process called spin diffusion, when tumbling is slow.[226] More recently, the 2D NOESY spectra of mixtures dissolved in viscous solvents have been recorded so that spin diffusion correlates together all the resonances on a compound-by-compound basis within a mixture. Such an approach was exemplified by means of an artificial mixture of small NPs and a mixture of dipeptides.[227]

For sensitivity reasons, recording a 2D NMR spectrum is based on ^{1}H signal detection because the signal intensity depends on the third power of the nucleus magnetogyric ratio. The pure-shift approach may, in principle, enhance the resolution in the direct (F2) dimension of 2D spectra. This idea was successfully applied to HSQC spectra, for which pure-shift signals can be obtained at a minimal sensitivity cost. However, the decoupling artefacts makes this approach more suitable for pure compounds than for complex mixtures[228], though recent advances may change that in the future.[229] Pure-shift along F2 was also involved in the spectral simplification of DOSY spectra.[230]

The resonance resolution in the indirect dimension (F1) of 2D NMR spectra is related to the extent t_1^{max} of the evolution delay t_1. A longer t_1^{max} will increase resolution, but if the indirect spectral width is preserved and a regular t_1 sampling value is used, the resolution enhancement caused by a given factor

will require the recording time to increase by the same factor if the number of recorded transients per t_1 value is left unchanged. Breaching this rule necessitates modifying the way that t_1 values are chosen. The reduction of the spectral width, while maintaining regular sampling, leads to spectral aliasing in F1 for increasing resolution in 2D NMR and has been used in NP studies, including *de novo* structure elucidation.[231] High resolution may be obtained without aliasing in the F1-band-selective spectra, assuming that only a narrow band of resonances must be zoomed along the F1 dimension.[232] More recently, the non-uniform sampling (NUS) of t_1 has become popular because it allows for a long t_1^{max} while keeping the number of used t_1 values to practical values by "skipping" a user-supplied proportion of them, with the missing data being reconstructed by extrapolation from the recorded data using a specific algorithm.[233] Another resolution enhancement technique relies on a special type of NMR data processing, known as covariance NMR, which provides cross peaks completely lacking multiplet structures in the correlation maps of homonuclear ^{1}H 2D spectra such as COSY, NOESY and TOCSY.[234]

4.1.1.3 Overcoming acquisition time issues

Achieving the optimal use of available spectrometer time and the necessity of observing quickly evolving chemical or biochemical systems have triggered the search for innovative concepts in NMR to reduce the time needed for the recording of spectra. A recent article by J. Farjon *et al.* provides a good idea of what can be achieved by combining fast pulsing techniques, NUS, aliasing and pure-shift techniques for the recording of the HSQC spectrum of a mixture of low-molecular-weight metabolites.[235] The term Ultra-Fast (UF) NMR is applied to a wide category of experiments in which nD ($n \geq 2$) NMR data acquisition is performed on the order of one second. Nearly all conventional 2D NMR pulse sequences have UF counterparts available. Of course, some sacrifices in resolution and sensitivity have been made at the price of speed. The sensitivity question can be solved by dynamic nuclear polarization (see section 4.1.1.1). Likewise, UF methods are suited for the monitoring of HPLC effluents by 2D NMR.[236]

Commercial NMR hardware has recently evolved so that two or more free precession signals can be recorded during a time lapse that was previously devoted to the acquisition of a single one. These multi-receiver systems will certainly become increasingly important in the field of NP chemical profiling.[237]

4.1.2 Metabolite profiling by NMR as practiced today

This section intends to expose the common practice of NMR in the field of the discovery of new NPs with significant biological activity. It does not attempt to present a thoroughly representative study but instead relies on 17 recently published articles, which were published from late August to mid-September 2018 in a journal in which the structures of bioactive NPs are often reported, i.e., the

Journal of Natural Products. All presented NPs were isolated and purified before being subjected to structural analysis. The set of used NMR spectra includes 1D ^{1}H, ^{13}C (possibly APT or DEPT), 2D COSY, NOESY or ROESY, ^{1}H-^{13}C HSQC (if not HMQC) and ^{1}H-^{13}C HMBC. In one case, a ^{1}H-^{15}N HMBC spectrum was recorded, and in another case, a ^{1}H-^{13}C 2D HSQC-TOCSY spectrum was recorded. The underlying methodology has remained unchanged for at least 20 years,[209, 238] with the only striking change being the recourse to high-field NMR, with resonance frequencies for ^{1}H nuclei of 600 MHz (6 times), 800 MHz (3 times) and 900 MHz (1 time). In all cases, the structure, data tables and spectrum drawings were only available in PDF format, thus making them unsuitable for subsequent computerized data extraction, an operation that would make the creation or improvement of spectro-structural DBs easier.[34] The spectra drawings in the supplementary information documents attached to these articles still revealed some improper mastering of processing steps, such as phasing and baseline correction, which were overlooked by the reviewers. Some NOESY or ROESY spectra were drawn in poorly usable single-color mode. The selection of a proper resolution setting at the acquisition time was not always adequate. The analysis of this set of recent typical structural identification reports of NPs indicates that there is still significant room for improvement to exploit the full potential of NMR to better support the NP identification process. Stories from the past have shown that when a new method provides a decisive advantage, its use will spread rapidly among the concerned community; in this category, one can find “inverse detection” (HMBC instead of COLOC), pulsed field gradients (which most chemists ignore but benefit daily from), cryoprobes, NUS data acquisition, and ultrahigh field NMR. The resorting to a conservative choice of a limited subset of NMR spectra types may either prove that this subset is sufficient for most uses or that most newly proposed methodologies do not have enough advantages to break the barrier of habit.

What is true for the analysis for pure NPs is also true for mixtures, even in the framework of natural extract profiling concepts such as HMBC barcoding,[239] HMBC networking,[240] 2D NMR differential analysis,[241] or heterocovariance processing.[242] Hopefully, patient efforts from spectrometer manufacturers and the promoters of new methodologies will certainly allow pertinent innovations to diffuse through the community of NMR users, even if they only apply the methodology without being an expert in the field, as it the case for most NP or organic chemists.

4.2 Dereplication strategies by NMR

A rapid bibliometric study shows that LC-MS is generally preferred to NMR for the rapid annotation of known molecules. Although the former is praised for its sensitivity and the richness of its related ecosystem of DBs and software tools, the latter is often chosen for its reproducibility, its quantitative aspects, and the richness of the structural information carried by NMR spectra.[211] This argument is hereafter discussed in more detail.

4.2.1 Identification of pure compounds

The dereplication of pure NPs by NMR is a matter of recording NMR spectra, extracting spectral features, and searching for them in an NP DB containing the experimental or predicted (see below) values of these NMR signals. Considering that a few hundred thousand NP structures have been reported and that their associated descriptions are not all available from the same source, pure NP identification is not always easy but is often possible. ^{13}C NMR has rapidly become the method of choice for NP identification, as proven by the existence of dedicated DB interrogation systems, such as those of NAPROC-13 for NPs only or those of NMRshiftDB, CSEARCH or ACD/Labs for more general DBs that also contain the ^{13}C NMR data of NPs from experiments or *in silico* prediction (see section 2 for a discussion of DBs).

4.2.2 Identification of compounds within mixtures

In the field of metabolomics, metabolite profiling by 1D ^{1}H NMR has been and is still intensively used as a primary source of data mainly for main metabolites monitoring in mixture.[210] This is the most sensitive NMR technique and is therefore adapted to the time constraints of high-throughput analysis. In such a way 1D ^{1}H NMR is mainly used in combination with chemometric data analysis to evidence NMR signals of given biomarker when large sets of samples are compared notably in approaches define as metaboNomics.[243] Such methods are extensively used especially for metabolomics studies of body fluids where NMR robustness, holistic and intrinsically quantitative nature has clear advantages.[244] The proprietary Chenomx suite combines a large in-house reference database with advanced deconvolution methods exposing overlapped and hidden signals in biofluids thereby combining identification and quantification for a number of clinically relevant small molecules.[245] Statistical analysis performed on 1D ^{1}H NMR profiles have been used for example to study the metabolic response of plant or microorganisms[246] to stimuli or differences related to composition modifications related to the origin of the sample.[211] As this is the case for body fluid metabolites,[244] standard protocols NMR metabolite profiling of plant extract exist.[208] These latter are however less generally applied than in the case of body fluid profiling (e.g., urine or plasma) probably because of solubility issues that may arise when specific classes of NPs have to be profiled.

Once the characteristic resonances of biologically relevant compounds have been revealed by multivariate data analysis, the structures of these compounds remain to be unveiled. The identification of metabolites in mixtures using 1D ^{1}H NMR spectra is possible provided that associated reference spectra DB exist obtained at the same magnetic strength and can be queried in solvent conditions that match the profiling experiment made and provided that, ideally, the adequate documentation of the original spectra is given.[247] For human metabolomics studies the HMDB contains about 3'000 1D NMR spectra that can be queried either by 1D ^{1}H NMR or ^{13}C NMR chemical shifts.[46] This DB contains,

however, very few NPs as it is targeted towards human metabolites (300 NPs versus > 110'000 metabolites). A NP DB (CH-NMR-NP) has been released by Jeol that contains 30'500 NP for which data were reported in major journals between 2000 and 2014 this DB is however not exhaustive for ^{1}H NMR data while ^{13}C NMR is completed (Table 1).

It has to be noted that even though statistical processing of ^{1}H NMR data sets is sufficient to reveal the presence of biomarkers in metabolomics studies, their structural identification is most often ensured by the interpretation of 1D and 2D NMR spectra.[246]

Indeed, while less sensitive than ^{1}H NMR, ^{13}C NMR is a very powerful method for the identification of compounds within mixture. The direct identification of low-molecular-weight compounds in mixtures by ^{13}C NMR without separation was reported in the 1980s for fresh plant extracts[248] and for petroleum distillates (presumably originating from extremely old plants).[249] This process comprised the building of a DB, either an experimental or an *in silico* one, and the matching of the ^{13}C NMR chemical shifts of the compounds stored in the DB with those from the spectrum of a mixture. Terpenes were also identified from essential oils using the SISTEMAT DB and associated SISCONST algorithm.[249] The same approach was more recently reported and led to the unexpected identification of minor amounts of monoterpenes in an alkaloid extract of *Peumus boldus*.[250] Another strategy combined extract fractionation by liquid-liquid partition chromatography and 1D ^{13}C NMR spectroscopy for the compound-by-compound grouping of experimental chemical shift values based on their chromatographic emergence profile (Figure 4). This grouping proceeds through the hierarchical clustering of emergence profiles using the freely available PermutMatrix software. The chemical shift groups are used as search targets in a DB that was locally developed using ACD/Labs Workbook and enriched with data from the literature and ACD/CNMR-predicted chemical shifts. This protocol, which is named CARAMEL (CARActérisation de MELanges, in French), has been put into practice more than one hundred times over the last four years,[251] most often for the characterization of plant extracts upon the request of the cosmetic industry, and it was recently routinely used by a start-up company (i.e., NatExplore - http://nat-explore.com/).

Due to the richness of their information, 2D NMR spectra have also been employed for dereplication from mixtures. The correlation cross-peaks in the HMBC spectrum of a pure compound form a network that is a subset of the nodes of a rectangular grid structure. Such networks are superimposed in the HMBC spectrum of a mixture, but they can be disentangled using a community detection algorithm in case a chemical shift in the identity between compounds would erroneously connect two unrelated networks. Each network can then be assigned to a molecular structure by means of an HMBC and HSQC DB built *in silico* from structures and the corresponding 1D NMR data. Candidate structures are further

validated using HSQC data. This method was exemplified by the analysis of a bark extract from *Picea abies*.[240] The pattern recognition (or barcoding) of HMBC spot clusters has also been investigated to identify known compounds from mixtures and even to obtain clues about the structures of unknown compounds by spectral subtraction. This approach was successfully applied to triterpenes from plants of the *Actaea* genus.[239] The use of prior knowledge, such as the concentration or biological activity (which is related to concentration through some non-linear but increasing function) of a single component of a mixture, leads to the identification of the resonances of this component in 1D and 2D spectra (COSY, TOCSY, HSQC, HMBC, DOSY) by means of a completely automatic workflow for data acquisition, processing and interpretation.[241] The term heterocovariance has been proposed to refer to the correlations between biological activities and spectroscopic features.[242]

The instantaneous vision of the metabolite set of a living organism would not be decipherable without a vision of the chemical pathways that govern its transformations. The goal of fluxomics is the unveiling of these metabolic mechanisms.[252] Fluxomics by NMR makes use of the organic compound precursors enriched in low-abundance isotopes, such as ^{13}C and ^{15}N.[253] The identification of the metabolites that are produced by feeding an organism with such compounds can be achieved using dedicated methods. The presence of neighboring ^{13}C nuclei within a molecule allows us to relate their chemical shifts by a TOCSY-type experiment, even within complex metabolite mixtures. This concept was illustrated by an *E. coli* cell lysate, which was fully enriched in ^{13}C, and its compounds were identified by searching a dedicated *in silico* DB.[254]

4.2.3 Chemical shift prediction

Chemical shift prediction is the cornerstone for building NMR *in silico* DBs. The abstract of a publication by the Merck company in 1995 indicates that "Using spectra estimated from structures circumvents problems of inconsistent, incomplete, missing or irrelevant data. It also enables rapid generation of reasonably sized DBs that are unavailable from commercial sources."[82] A commercial company has integrated 22 million compounds from the public ChemSpider DB,[75] along with the predicted chemical shifts for ^{1}H, ^{13}C, ^{15}N, ^{19}F and ^{31}P nuclei. Although this DB contains only approximately 0.2% of NPs, their approach was validated by the identification of unknown NPs in pure form based on their ^{1}H chemical shifts and the ^{13}C chemical shifts deduced from a multiplicity-edited HSQC spectrum.[255] The key point of this approach is the availability of reliable chemical shift predictors. Relating ^{1}H chemical shift values to chemical group substitution was first published by J. H. Shoolery in 1959 for CH_2 groups.[256] This first approach evolved to the finding of additivity rules that can be applied to a wide variety of structural contexts and was extended to ^{13}C NMR. Other approaches rely on structural descriptors that resume the environment of a nucleus for which a prediction is searched. An environment coding scheme, named Hierarchically Ordered Spherical description of Environment, or

HOSE, was proposed by Bremser in 1978 and is still in use.[257] A third approach aimed at relating environment descriptors and chemical shift values by means of an artificial neural network was reported in 2002, when no one spoke about Deep Learning and Artificial Intelligence was a forgotten concept.[258] More accurate predictions may be obtained by combining the results of different methods. It is, however, rather difficult to know what happens behind the scenes in commercial prediction software, such as those developed by ACD/Labs or NMRpredict by Modgraph. Nevertheless, the NMRpredict website provides a free service for ^{13}C NMR-based Spectral Similarity Search with Ranking based on more than 64 million compounds from the PubChem DB[76] and for which ^{13}C chemical shifts have been predicted (http://nmrpredict.orc.univie.ac.at/). Noncommercial predictors such as NMRshiftDB2[259] or Spinus[260] are available for ^{13}C and ^{1}H chemical shift prediction through web interfaces. The abovementioned software relies on a corpus of reference molecular structures and associated chemical shift values from which calculation models were elaborated. Most often, the reference data are hidden from the end-user, with the exception of NMRshiftdb2, for which the corresponding DB is available in SDF format. These prediction methods are extremely fast, and the goal of substituting experimental data with calculated data for the constitution of *in silico* DBs is not an overwhelming task. The level of ^{13}C NMR prediction has reached such a level of accuracy that some NP journals recommend researchers to check their structure assignment by preforming *in silico* ^{13}C signal prediction.

The prediction of chemical shifts and coupling constants by *ab initio* methods is of interest for molecules presenting rare chemical functional groups[261] and has allowed for the revision of structures or spectral assignments erroneously assigned to NPs.[32] A recent article proposed optimized prediction conditions, namely, a wave function basis, a density function, a solvation model and a calculation method; this shows how this methodology can be adequately applied to the determination of the relative configuration of asymmetric centers in NPs and how the necessary conformation analysis step may lead to absolute configurations when used in conjunction with chiroptical methods, such as ECD or VCD.[262]

4.3 Automated interpretation of NMR spectra

As a general rule, full *de novo* structure elucidation follows unsuccessful dereplication. For this the manual structural assignment from NMR data can be efficiently assisted by software. The goal of computer-aided structure elucidation (CASE) software is to find solutions to structural problems by placing bonds between atoms based on the connectivity relationships inferred from NMR.[263] The nature and number of these atoms are defined by the MF formula obtained by HRMS (see section 3.2). Solution structures fulfill constraints defined by the rules of organic chemistry, NMR data, and any information about the compound origin derived from another spectroscopic method or phylogenic

considerations. A chemical shift value places a constraint on the environment of the atom it concerns, and a correlation in a 2D spectrum or a coupling constant value imposes a constraint on the distance, measured in bonds, between two atoms. It may be noted at this point that the structure elucidation processes performed by humans and computers are not very different from each other, although computers never tire of finding solutions when one has been found and are less prone to be guided by false preconceptions about solution structures.

The computer-aided structure elucidation of small organic compounds was one of the first application fields of Artificial Intelligence (AI) during the 1960s, when emerging computer technology started to reach the necessary efficiency level. Then, AI was then the science of automatic deduction, but it had very few possibilities for autonomous learning; since then, this paradigm has been inverted.

Only a few companies presently produce CASE software, namely, ACD/Labs (Structure Elucidator), Bruker (CMC-se), and Mestrelab (Mnova v.12). Academic software such as SENECA[264] and LSD[265] (see Figure 4 for the operating principles of LSD) are available as free software, whereas COCON[266] can be accessed through both through the web and as part of Mnova v.12.

Figure 4

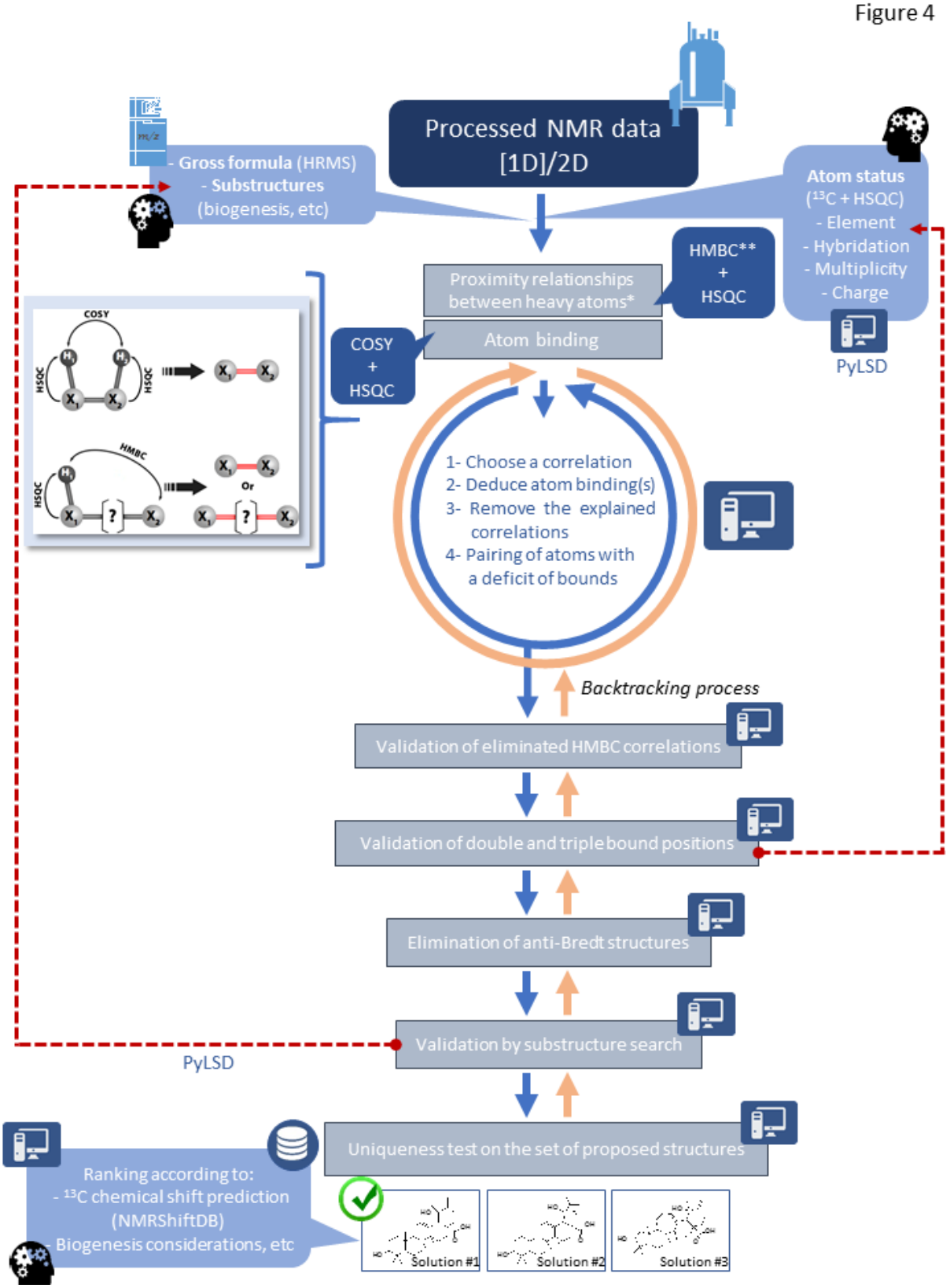


Figure 4. Principle of structure generation by the LSD CASE software. The goal of the LSD software is to draw bonds between initially non-bonded atoms to find the possible planar solutions of a de novo structure elucidation problem. The number and nature of the atoms of the solution structure is supposed to be known from HRMS data. The set of atom statuses must be fully determined before the beginning of the problem resolution process. In the case this could not be reliably achieved, a software layer written above LSD, named pyLSD267,306 (a), can be invoked to resolve status ambiguities. LSD is not aware of the relationships between chemical shifts and structural features and places bonds between atoms solely based on 2D NMR

correlation data. The combination of COSY and HSQC data yields bonds between heavy (i.e., non-hydrogen) atoms from which the resolution process starts. The combination of HSQC and HMBC data yields proximity relationships between heavy atoms expressed as distances measured in number of bonds (1, 2, or more). HMBC correlations of close 13C resonances may be declared ambiguous, and all possible interpretations will be systematically considered. The resolution process starts by the recursive use of proximity relationships for the formation of bonds and removes those that become explained by the newly formed bonds. The atoms for which not all of their bonds are present, as inferred from their status, are thensystematically paired in a recursive process to build complete structures. Recursive processes are needed to explore all possibilities and reconsider choices (backtracking) opened by data interpretation so that the exhaustivity of the solution search is ensured. Each structure then passes through a series of validation steps. The distances between atoms are checked to address the HMBC correlations through 4 bonds or more. Double and triple bonds are placed between atoms to reach the needed coherence with their hybridization state. Anti-Bredt structures are eliminated. (b) The user may impose substructural elements to be present or absent in the solution structures (any combination of such constraints is allowed) according to external information sources, possibly spectroscopic or biogenetic. Ambiguous HMBC correlations may lead to duplicated solutions that need to be removed. The pyLSD software layer sorts the solutions according to the similarity of the 13C NMR chemical shifts with those predicted by NMRshiftDB.

The publication entitled "Exploiting the Complementarity between Dereplication and Computer-Assisted Structure Elucidation for the Chemical Profiling of Natural Cosmetic Ingredients: *Tephrosia purpurea* as a Case Study" provides a good example of how a non-conventional workflow can be applied to speed up the chemical analysis of a plant extract.[268] This study relied on the CARAMEL dereplication workflow and on the LSD CASE software. Three compounds declared as unknowns after dereplication were identified as known compounds by structure elucidation but were never reported in the *Tephrosia* genus, likely due to the difficulty of their isolation by conventional chromatographic methods. In this example, as in many others, the dereplication process made it possible to quickly identify approximately 80% of the dry *mass* of an extract. This examples also highlight the usefulness of the inherent quantitative aspect of NMR since the metabolite coverage can be expressed in term of percentage of composition and not only in number of metabolites as is the case in most of the MS-based metabolomic output results demonstrating again the complementarity of the MS/NMR approaches for natural extract composition assignments.

SENECA operates in a stochastic way, which means that it starts from an initial guess of a solution structure and refines it by atom permutation to reduce the number of constraint violations. Many of these processes may be run in parallel, increasing the odds of rapidly finding a solution that presents a minimum number of constraint violations. The other CASE systems are deterministic, meaning that they systematically explore all possible interpretations of data.[263]

The structural constraints provided by NMR are ambiguous by nature. A chemical shift value is generally associated with a multitude of possible molecular fragments, and even a COSY correlation may originate from a ^{2}J coupling (which is easy to detect from an HSQC spectrum), a ^{3}J coupling (as expected) or an ^{n}J coupling with $n \geq 4$. A ^{1}H-^{1}H COSY correlation from an ^{n}J coupling indicates n-2 bonds between the non-H atoms that bear the concerned ^{1}H nuclei. Consequently, the ambiguity in the value of n leads to structural ambiguity because the NMR signal intensity is sensitive to coupling constant values but not to bond counts. In the same way, a ^{1}H-^{13}C HMBC correlation through n bonds (where n

may be equal to 2, 3, 4 or even more) indicates an *n*-1 bond distance between two non-H atoms. A given CASE software program must therefore be able to address coupling paths of nonstandard length.

Resorting to complementary 2D NMR spectra is one possible way to reduce the size of the chemical space explored by a CASE algorithm.[269] A ^{1}H-^{13}C H2BC (Heteronuclear 2 Bond Correlation) spectrum directly indicates the existence of a bond between two non-H atoms through a $^{1}J(^{1}H\text{-}^{13}C)$ coupling and a $^{3}J(^{1}H\text{-}^{1}H)$ coupling.[270] Considering both a HMBC and a H2BC spectrum, the latter allows one to classify HMBC correlations as arising from either a ^{2}J or a ^{3}J correlation, with the limitation that a $^{4}J(^{1}H\text{-}^{1}H)$ coupling may have a significantly non-zero value and that some $^{2}J(^{1}H\text{-}^{1}H)$ couplings may not be detected, thus leading to the incorrect interpretation of the H2BC spectra. Ultimately, non-ambiguous 1-bond connectivity between non-H atoms is revealed by a 2D ^{13}C-^{13}C INADEQUATE spectrum. The well-known limited sensitivity of INADEQUATE reduces its scope because the mass of the required substance is rarely, but possibly, available for an unknown compound.[271] A 1,1-ADEQUATE spectrum, which is a ^{1}H-detected version of INADEQUATE, benefits from better sensitivity and can be used as a reliable source of 1-bond connectivity in CASE software.[266] The structure elucidation of H-poor molecules may incite one to search for data with very long distance connectivity, a goal that can be achieved using spectra such as 1,n-ADEQUATE[272] or LR-HSQMBC,[273] with a possible cross validation of the solution structures by the theoretical calculation of the small coupling constants by quantum mechanical methods.

Another type of ambiguity arises from resolution issues in 2D NMR spectra. This point has already been addressed in section 4.1.1.2. The resolution in the indirect dimension of ^{1}H-detected 2D ^{1}H-^{13}C correlation spectra is of great importance, and the impact of its improvement through spectral aliasing along F1 (see section 4.1.1.2) has been demonstrated by the reduction of the number of proposed solution structures.[231] Band-selective HSQC and HMBC experiments have proven to be useful for recovering pertinent connectivity data from particularly crowded spectral regions.[232] The 2D correlation spectra of other nuclei, mainly ^{15}N, constitute useful sources of complementary constraints for proposed structures.[232]

Unresolved contradictions between experimental structural constraints and the ways they are interpreted lead to a lack of solutions to a problem and necessitate the reexamination of software input data. A loosely constrained set produces a high number of solutions for which validation and ranking must be carried out.

CASE software generally does not directly propose 3D structures for solution structures but instead proceeds by first proposing 2D structures. Elaborating a 3D structure from a 2D structure in an automatic way has recently been proposed; this process begins with the automatic generation of

diastereoisomers from a planar structure. Stereoisomer selection relies on *ab initio* chemical shift calculations and on the measurements of residual dipolar couplings and residual chemical shift anisotropies in anisotropic media. Molecular flexibility is handled through automatic conformer set generation and the averaging of NMR parameter values over the conformer population. This approach was exemplified by two well-known NPs and two recently reported ones for which structural revisions were proposed.[274]

4.4 Toward user-friendly computer interfaces

Currently, computers intervene during nearly every stage of metabolite profiling by NMR. Spectrometer manufacturers provide hardware and associated software tools that grant access to regularly updated libraries of pulse sequences for spectrum acquisition. The automated acquisition of spectra sets using sample changers has become standard in most NMR facilities and can be set up with minimal human intervention through specific user interfaces. Manufacturers also provide data processing tools, even though processing may be carried out using proprietary software, for which the ease of use is put forward as a major commercial argument. The capability of this software may extend toward automated spectrum interpretation, providing access to structure verification and structure elucidation tools.

In contrast to metabolomics, NP research does not yet benefit from freely accessible data handling workflows leading from the collection of NMR raw data sets to statistical data analysis and possibly to the identification of biomarkers. One possible reason for this difference is the size and structural diversity of the chemical spaces related to these domains, which comprise a few tens of thousands of molecules of a reasonable number of compound classes for metabolomics and a few hundred thousand structurally very diverse molecules for NP chemistry. In fact, automated annotation and quantification of (very) small molecules in matrices like urine is starting to work;[210] for NPs such automation in complex mixtures is currently still out of being reached. However, with current progress it is not unlikely that tools will emerge that could do a similar job as for (very) small molecules. For example, MetIDB[54] is publicly available database containing ^{1}H-NMR spectra predicted by PERCH NMR software[275] for specific NP compound classes such as flavonoids and aurones that could in theory be linked within an automatic annotation framework. Another possible reason may lie in a difference between the organizational stages of the MS and NMR scientific communities.

In the future, it can be expected that the structures of all known NPs will be freely accessible to the scientific community. These structures should be part of a global knowledge base that could possibly be designed as follows. Structures must be linked to other types of data, such as the original bibliographic reference, the storage location of spectroscopic raw data,[33a] the values of the extracted

spectroscopic parameters, and the biological origins of the compounds, thus altogether resulting in structure-spectra-origin triplets. Here, "origin" represents the full taxonomic data of the organism and, if relevant, the part of the organism that was studied. The DBs of simpler structure-spectra pairs, such as NAPROC-13, can be used to solve many problems, but using a given compound's origin information can greatly speed up the search for a structure from spectral data, as compound classes are related to biogenesis, which is itself related to organism classes.[276]

The possibility of combining the results of searches for structural, spectroscopic and taxonomic specifications is of great importance for dereplication and structure elucidation (see section 3.2). Such an integrated approach was put into practice in the 1980s with the creation of the "knowledge base" SISTEMAT, a project that was driven in Brazil by Pr. Emerenciano. These underlying concepts would certainly be worth being used as a source of inspiration for the design of future DBs and software dedicated to NP chemistry.[277]

5 Quality and Reporting of Metabolite IDs

Traditionally, in NP research, new purified compounds from extracts are reported with a complete interpretation of their NMR (^{1}H, ^{13}C and 2D) spectra and MS data as well as additional spectroscopic data (UV, IR, ECD), which can be used to obtain a *full de novo* structural determination. The NP structures reported in the literature are thus considered high-quality standards. Even under such ideal conditions, however, the misinterpretation of spectral data may lead to the misreporting of structures.[278]

The accurate annotation of NP structures based on LC-HRMS/MS or NMR profiling in complex extracts remains a challenge. Therefore, it is important to keep in mind that such identifications may be (and often are) putative; thus, such annotations must be reported with care. Several very important questions remain: "are you sure of your annotation?" and "how do your peers gauge your confidence?"[24, 279] It is crucial to avoid having a putative annotation later become incorrect knowledge that is considered true. While it is important to share the results of even approximate annotations throughout the NP chemistry community, reporting such results in a correct manner is still an issue. In dynamic systems, such as the GNPS platform, annotations can be refined by iteration cycles through various studies; for example, by sharing data about a given organism, annotations can be revised.[44] To publish tables of annotations, reliable metabolite profiling reporting standards must be defined.

In the field of metabolomics, this question has been a matter of debate since 2005. Thus, reporting standards are regularly discussed, and a common consensus was reached for reporting the annotation (or identification) of a compound, which is based on 4 different levels of identification that need to be indicated in addition to the proposed annotation along with accurate information about the analytical

method.[84a] These annotation levels correspond to the following:

- **Level I - Identified compounds:** indicates that annotation is based on comparison with authentic standards and that two types of orthogonal information were used for confirmation. Thus, in principle, this level can only concern unambiguous identifications made by comparisons against reported NPs and cannot reflect the full characterization of unknowns. In the case of LC-HRMS/MS, accurate MS and RT matches are necessary for compound identification and usually require standards (fully identified by NMR).
- **Level II - Putatively annotated compounds:** indicates that annotation was achieved without chemical standards but based on spectral similarities in comparison with previously reported data (*in silico* spectral data may be considered if no real data are available). In this context, there is a reasonable chance that the compound annotation is correct. When dealing with LC-HRMS/MS data, this means that the mass accuracy of MS and MS/MS must reach level II, sometime complemented by biological source consistency. Similarity of all available NMR spectral data is warrantied.
- **Level III - Putatively characterized compound classes:** indicates that partial structural annotation is possible based only on spectral information. This usually corresponds to partial spectral matches and thus yields a limited chance to achieve accurate compound annotation. In the case of LC-HRMS/MS, this corresponds to the accurate mass accuracy of MS completed with a partial MS/MS match and sometimes complemented by phylogenetic consistency. Characteristic NMR signals orient compound identification toward a particular class of compounds.
- **Level IV - Unknown compounds:** only spectral (and chromatographic data, in the case of LC-HRMS/MS) are available, but with no correct or partial matches. Thus, information is available for the latter observations under similar acquisition conditions, and identification can be performed later if structural information about this compound becomes available.

Very recently, a **Level 0** annotation of "Unambiguous 3D structure" was proposed by Blaženović *et al.*,[171b] corresponding to a "fully identified 3D structure with absolute identification", as traditionally achieved by NP chemists and thus fully integrating 1D and 2D NMR interpretation as described in section 4.2. These 5 levels of identification have the advantage of being rather simple to determine; however, it remains difficult to compare different annotations within a level using such an approach as some terminology used within the level definitions remains arbitrary as are the requirements for each level and their implementation for the different analytical techniques and (hyphenated) combinations thereof. Therefore, some researchers are proposing a numeric quantification of the annotation quality, but they have not yet reached a consensus.[24, 280] However, such a strategy is

difficult to generalize and ultimately depends on the analytical method, annotation strategy and tools used. Therefore, evaluating the knowledge and experience about the apparatus used is a good starting point with which to evaluate annotation confidence.

In the specific case of NMR, a scoring process was proposed by J. R. Everett.[281] This evaluation method was established in the frame of metabolomics but can be a source of inspiration in any other field of chemical analysis by NMR or even beyond. The publication by Everett defines two numerical criteria that indicate whether or not a compound can be safely identified by NMR by dividing the number of reported descriptive bits of spectroscopic information by either the number of carbon atoms or the number of heavy atoms in the considered molecule. The criteria values are obviously higher when all available spectra types are used than they are when using a 1D ^{1}H NMR spectrum only. A molecule with a single ^{1}H resonance cannot thus be identified by means of only a 1D ^{1}H NMR spectrum due to its associated low value criteria. Once a set of spectra is defined, the experimental values of their ^{1}H and ^{13}C chemical shifts and ^{1}H-^{1}H coupling constants are compared to the expected ones. The typical deviation values for these parameters are provided as guidelines to determine whether a proposed identification is valid or not. In this way, a compound for which no reference sample is available and whose identification is precluded to reach the confidence level of 1 according to Sumner *et al.*[84a] can nevertheless be safely identified. The publication by Everett also nicely demonstrates the importance of leaving raw NMR data freely accessible: he was able, by reprocessing an FID from the HMDB DB,[46] to measure a small coupling constant that gave rise to a COSY correlation, a constant that was not detectable at the time the spectrum was first interpreted. Another important aspect in this paper is that it refers to only 75 metabolites for which NMR spectral features are available from a public DB. The lack of so many accurately known experimental spectral features for most specialized metabolites forces us to rely on *in silico* predicted (or evaluated) features, including ^{1}H and ^{13}C NMR chemical shifts and those that permit the simulation of COSY and HMBC spectra, namely, the values of the ^{1}H-^{1}H and ^{1}H-^{13}C coupling constants.

An intrinsic limitation of MS-based structural assignments compared to NMR ones is the limited possibilities of stereoisomeric differentiation based on MS/MS interpretations. Therefore, such annotation strategies usually yield only flat chemical structures. This represents a strong limitation of NP assignments because a majority of them exhibit complex 3D structures. To address this issue, orthogonal detection that can distinguish stereoisomers could be implemented in profiling studies. One method of choice would be detection by an ECD detector,[282] which is easy to couple to liquid chromatography. LC-ECD-HRMS/MS could potentially enhance structural elucidation by additionally defining the stereochemistry of the flat structure by analyzing ECD spectra[283] and therefore approaching a Level 0 annotation.

In this context, a new trend currently emerging in metabolomics is to improve the structural annotation confidence by combining multiple analytical platforms to take advantage of the benefits of each of them (see also section 7.1).

To evaluate the quality of annotation strategies and their improvements over time, an international initiative was launched in 2012, which is called the Critical Assessment of Small Molecule Identification (CASMI, http://www.casmi-contest.org).[284] CASMI was proposed to blindly solve annotation challenges based on MS/MS information (and sometimes additional metadata). Participants propose annotation(s) for every challenge on an approximately yearly basis, and the correct solution is later released. It remains important to note that in the CASMI contest, only flat chemical structures are considered due to the limitations of MS/MS annotations. In fact, only the first 16 characters of InChIKey (a hashed version of the full InChI)[74a] are used to highlight correct annotations. The results obtained in CASMI demonstrate that since 2014, MF determination is no longer an important issue for metabolites below 500 Da based on MS and MS/MS data.[285] Since 2016, these challenges have clearly focused on the annotation of NPs; these results provide good examples for assessing the efficiency of the current trends in annotation strategies as they are typically larger than 500 Da and structurally very diverse. It is interesting to note that when dealing with a low number of challenges compatible with manual annotation, as was the case in 2016, semiautomatic annotation with human interpretation was able to achieve the correct annotation almost every time.[286] However, this manual strategy was determined to be too slow to use for a large number of annotations. In 2017, a much larger number of challenges was proposed, and it was observed that strategies based on *in silico* spectral simulations performed "satisfactorily" for the large-scale annotation of LC-HRMS/MS data. Among all of the evaluated interpretation tools from CASMI 2017, CSI:FingerID[111, 143, 164] and MS-Finder[174, 287] were the most promising.[169] However, very few proposed workflows were able to perform annotations automatically from raw LC-MS data.[149, 169] The global feedback from CASMI 2016 and 2017 revealed that (1) annotation efficiency is directly related to the availability of the compounds within DBs; (2) efficient tools are currently being developed for compound annotations directly from LC-MS/MS data, even though accurate annotation is still yet to be achieved, as highlighted by the fact that at best, only half of the challenges were successfully solved; and (3) exhaustive spectral DBs are still necessary. These results also showed that the careful manual interpretation of the results obtained by *in silico* tools is still very much needed. Furthermore, many publications still do not report their metabolite annotation and identification strategies nor report metabolite identification levels hampering proper reuse of their findings.

6 NMR/MS Combination

6.1 Recent advances in LC-NMR / LC-SPE-NMR

One theoretically ideal way to link LC-MS information to NMR would be to perform LC-NMR either on-line with MS or using similar LC conditions for separate NMR and MS detection to obtain matching and complementary MS and NMR structural information for all peaks detected. Such a solution is, however, not practical because of the inherent very low sensitivity of NMR compared to that of MS detection and the compatibility of the solvents needed for both the NMR and MS sides. In typical on-flow LC-NMR experiments, deuterated water must be used, while the ^{1}H-NMR organic solvent modifiers can be suppressed by dedicated NMR pulse sequences. The use of D_2O, however, will cause the exchange of protons due to deuterium, which will then complicate the interpretation of MS spectra when recorded on-line.[288]

The coupling of high-performance liquid chromatography with NMR spectroscopy (LC-NMR) represents one of the most powerful methods for the separation and structural elucidation of NPs in mixtures.[289] The on-line coupling of both techniques is made feasible by the use of dedicated flow probes (30-60 µL) and has permitted the acquisition of the ^{1}H-NMR spectra of the main NPs in various natural extracts. The sensitivity of the flow cell can be enhanced by using cryogenic flow probes.[290] However, on-flow LC-NMR is limited because of the need for solvent suppression, which compromises the quality of the spectra obtained.[288]

An effective way to overcome this problem has been the introduction of the LC-Solid Phase Extraction (SPE)-NMR technique, which enables the efficient preconcentration of the sample prior to NMR detection and enables its measurement in fully deuterated solvent while HPLC separation is performed in standard HPLC solvents.[291] The latest developments correspond to the fully automated integration of MS hyphenation for trapping in LC-MS-SPE-NMR setups.[292] In brief, in such systems, LC peaks are automatically trapped on SPE cartridges and released in either an LC-NMR flow cell or 1.7-mm ID microtubes for the analysis of 1D and/or 2D-NMR spectra. Sampling is performed with multiple injections of extracts on the column, resulting in the efficient enrichment of given analytes on the SPE columns. An alternative consists of a single injection of extracts (typically a tenth of a mg) on semipreparative columns after the geometric chromatographic transfer[293] of the LC-MS conditions used for the metabolite profiling for micro-fractionation and drying and subsequent NMR analysis.[294] Both approaches enable the full *de novo* characterization of NPs by NMR using only 1-5 µg of analytes with cryogenic probes[295] or microprobes fitted to 1.7-mm microtubes (30-µL volume) on a high-field magnet (600 MHz).[296] For a recent review of the applications of LC-SPE-NMR, see Sumner *et al.* and Gomes *et al*.[296-297]

For example, such an approach was successfully applied for the unambiguous identification of 22

compounds in *Pueraria lobata* with the injection of ca. 700 µg of extract on the column[298], and more than 20 coumarins, including several regioisomers that were difficult to separate, were identified in *Coleonema album.*[299]

LC-MS-SPE-NMR is thus efficient for providing complete sets of NMR data about the main constituents of an extract. This method can be used for the selected full *de novo* identification of given LC peaks that cannot be dereplicated or for the identification of unknowns. It can also be used as a method for the identification of all major extract constituents (typically a few tenths of a plant extract) that can be used to unambiguously identify the main LC peak in deep metabolome studies for the further improved annotation of minor constituents by the propagation of annotation in MN, as discussed in section 3.3.

However, such analyses are rather time-consuming, require complex automated procedures and SPE-NMR setups and as a result are still relatively seldom applied in the NP community. To combine MS and NMR information, other approaches that make use of classical NMR profiling and MS data exist. Moreover, new cheminformatics and computational methods have recently been developed (see section 6.2).[300]

6.2 MS/NMR data combination strategies

Classically, NMR and MS metabolomics workflows are performed independently, and annotated metabolites are then compared based on their occurrence in a table of annotations.[301] More recently, complementary structural information from MS and NMR was efficiently combined by obtaining the main substructures from MS/MS data and linking these building blocks using specific regions of the ^{1}H-NMR data only to solve the complete structures of conjugated phenylvalerolactones from human urine and large (>1000 Da) NPs in the form of glycosylated flavonoids at µgram levels.[295, 302] This very low amount of material needed was achieved by effectively combining MS and NMR analyses, thus allowing to zoom in on selected regions of the NMR spectra thereby ignoring regions with interfering impurities from the samples and column materials used during LC-MS-SPE-NMR.

An interesting more systematic way to combine MS and NMR information without compromising both techniques consists of using HRMS for metabolite profiling data on one side to extract the MF of all metabolites and to generate all possible associated structures. Using this information, the NMR spectra of each member of a structural manifold are predicted and compared with the experimental NMR spectra to identify the molecular structures that match the information obtained from both the MS and NMR techniques. This approach is termed SUMMIT MS/NMR.[303] It was applied to profile *Escherichia coli* extracts, and a wide range of different types of metabolites, including amino acids, nucleic acids, polyamines, nucleosides, and carbohydrate conjugates, were successfully identified. The integration of such information has the potential to link MS and NMR data, provided that the structural

annotations based on HRMS data are sufficiently well filtered, as in section 3.1.

The approach can also be performed starting with NMR data by querying in the NMR profiles against an NMR DB of standards followed by the automated prediction of the masses (*m/z*) of all likely ions and adducts of metabolite candidates with their characteristic isotope distributions made available; this approach has been termed "NMR/MS Translator".[304] The expected *m/z* ratios are then compared with the experimental MS spectrum for the direct assignment of the signals of the MS that correspond to the generated metabolites. For example, this approach was used to identify 88 metabolites in human urine by combining 2D ^{13}C-^{1}H HSQC with direct infusion ESI-MS spectra that have consensus signals in both NMR and MS spectra.

Both the SUMMIT and "NMR/MS Translator" approaches can be nicely integrated.[301] Because chemical shift and accurate mass data are co-analyzed, such a method of combining MS and NMR was found to significantly increase the accuracy of metabolite identification compared with approaches where samples were independently profiled and processed in both MS and NMR. Having MS and NMR data at hand acquired in mixture on the same samples also makes it possible to search for unknowns by in depth investigation of 2D NMR spectra obtained on mixtures.[305] This technique thus has the potential to overcome the need for experimental MS and NMR metabolite DBs since the approached rely on predicted spectra.[303]

An intermediate approach between the correlation of MS and NMR data from crude extracts and those of fully separated metabolites, such as LC-SPE-NMR, could be the application of MS/NMR correlation approaches to the ^{13}C-NMR spectra acquired from the rapid coarse chromatographic fractionation of NP extracts, as is the case for the **CARAMEL approach** (see section 4.2.2 and Figure 4).[306] Here, the structures annotated from the LC-MS/MS profiles of the enriched fractions could be correlated to NMR by the ^{13}C-NMR simulation of a candidate structure, similar to the SUMMIT approach. A similar approach can also be achieved by coupling the data obtained by semi-preparative micro-fractionation with their analysis by both NMR and LC/MS. This was recently demonstrated on a standard compound mixture where ^{1}H-NMR signals were correlated with MS signals among all micro-fractions.[307] Then, analyzing the most abundant signals in the NMR spectra may benefit from additional semi-quantitative detection methods, such as ELSD or CAD (see section 3.1.4), during the profiling of fractions. Such an approach may, in the future, represent an interesting alternative to tackling the complexity of NP structures in natural extracts and unambiguously annotating the main metabolites as these semi-quantitative detection methods allow for better linking between NMR signals and their corresponding MS peaks based on quantitative relationships that do not exist in MS due to ionization and matrix effects.

7 Toward improved confidence in metabolite ID

7.1 Additional orthogonal methods (RT/CCS) and weighting meta-scores

As discussed in sections 3 and 4, LC-HRMS/MS profiling provides in-depth metabolite profiling over a large dynamic range with a high sensitivity but with a relatively restricted capacity for the unambiguous annotation of NPs when used in a generic approach. On the other hand, NMR provides detailed structural elements for the identification and semi-quantitative estimation of the amount of metabolites present, but its inherent sensitivity is one or more orders of magnitude lower.

Today, the challenge for improving confidence in annotating LC-HRMS/MS data relies, in our view, on combining different information sources that can further filter the structural hypothesis assignment generated by similarity scores in MS/MS for candidate structures with a given MF. Indeed, an annotation is only valuable if it can be scored within an established reference system to rate its confidence. The ideal reference system would be a score of good-quality HRMS/MS spectra acquired with optimal fragmentation energy compared to a DB of experimentally acquired spectra recorded with identical parameters.[81] As discussed above, such a DB does not exist for all NPs and would very difficult to create because fragmentation conditions will evolve with the continuous development of MS platforms, unless efficient ways to precisely standardize CID conditions are developed (see section 3.1.3).

Thus, using only the ranking of possible NP structures based on MF determinations and partial matches against *in silico* DBs, or experimental DBs acquired under conditions that do not perfectly match those under which reference spectra were recorded, we cannot obtain perfect scores.

Thus, presently, most NP annotations obtained in this way provide only hints about their structures and yield MSI annotation levels of 2 or 3.[24] To further improve on the currently existing identification levels and make them compatible with NP research, we argue that the scoring system should be multifactorial and take into account additional factors leading to higher confidence in annotation, such as the retention behavior of the analyte or the consistency of the annotated structure with its metabolic context.[73] Ideally, a meta-score integrating these various aspects should be established.[81] Such a meta-score may be based on various information sources as shown in Figure 5.

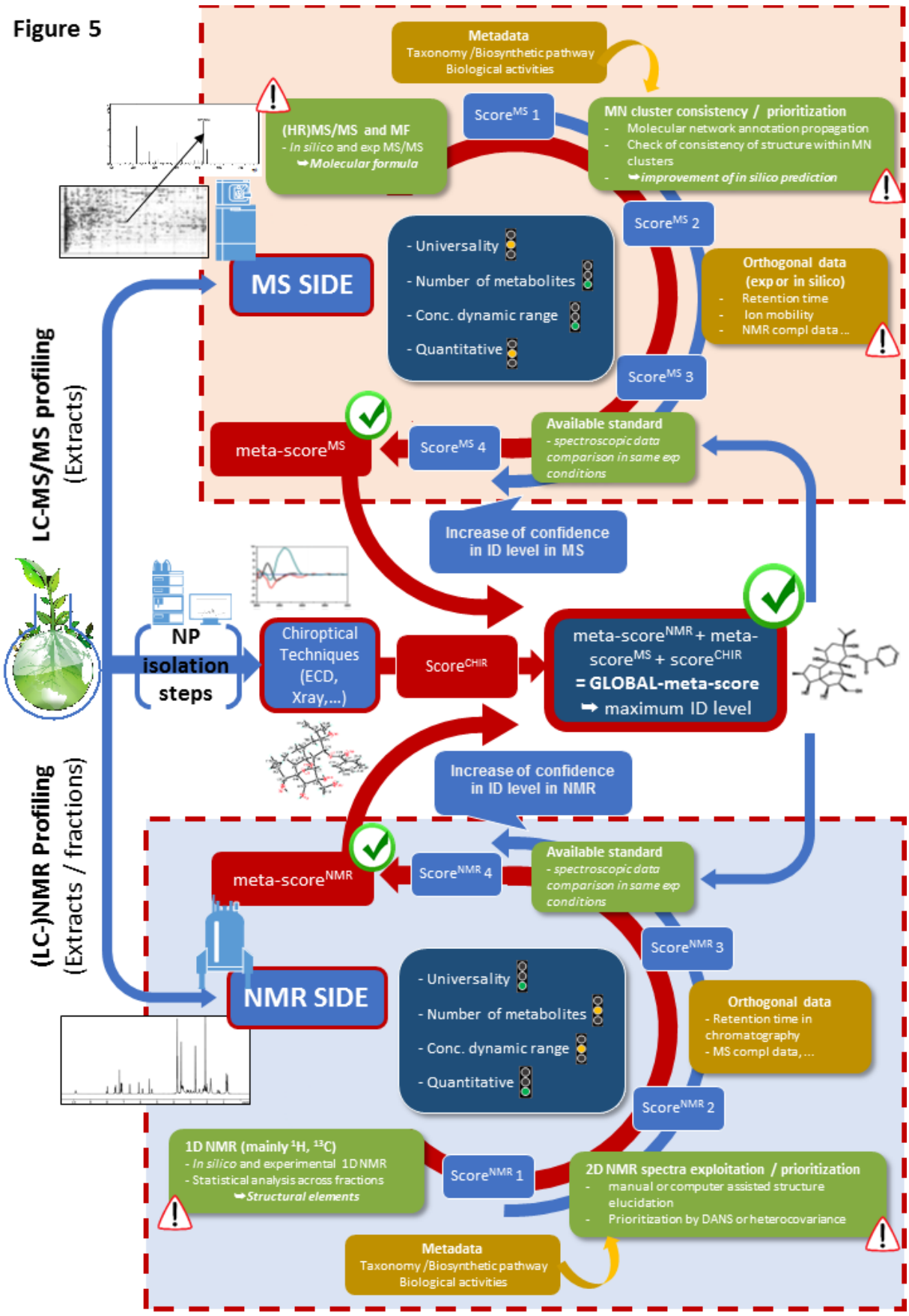


Figure 5. Schematic view of possibilities to improve the annotation scoring by weighting different partial and imperfect scores given by all methods on both LC–MS and NMR for maximum confidence in the compound identification (ID) process. NMR and

MS are the main spectroscopic techniques for the identification of a chemical structure. Whether on the NMR side or the MS side, orthogonal techniques are used to increase the ID level of confidence (e.g., liquid chromatography for both MS and NMR often offline or ionic mobility on the MS side). Moreover, some extra-data (brown boxes) are combined with spectroscopic data, leading to workflows that can be organized in the same way for MS and NMR approaches. They start from the lowest ID level (scoreMS or NMR 1) when only (HR)MS or 1D NMR are recorded and compared with calculated or experimental spectra, for instance, by means of in silico or experimental databases, and they reach the highest score (scoreMS or NMR 4) after complementary data acquisition and processing (e.g., MSn, 2D NMR) and usage of extra-data (e.g., taxonomy, biosynthetic pathways, NMR data for MS-based workflows and MS data for NMR ones). Thus, ideally a meta-score (meta-scoreNMR or meta-scoreMS) resulting from the addition of the scores validated for the NMR side ($\Sigma x = 14$ scoreNMRx) or the MS ($\Sigma x = 14$ scoreMSx) side can be thus calculated. For 3D sutural aspects relative configuration is generally proposed by combining coupling constant analysis and the Overhauser effect in NMR, sometimes together with conformational analysis and molecular modeling. Absolute configuration determination must resort from chiroptical techniques such as ECD, VCD, or X-ray crystallography (scoreCHIR). If all aspects can be taken into consideration with a correct weighting of all scores, in the future, a GLOBAL-meta-score could ideally be defined as a combination of the three individual meta-scores. The exclamation marks highlight critical steps in metabolite ID.

In NP research, a key aspect is to contextualize an annotation based on taxonomy. As mentioned in section 3.2, such information is important to reduce the number of possible candidate structures to those previously reported for a given species, genus or family. From a scoring perspective, hits that structurally relate an NP to a metabolite previously reported in a phylogenetically related species should thus be ranked better than an NP that has been reported to occur in unrelated organisms. Such deductions are usually made manually when integrating the metadata associated with a given sample (e.g., in the literature or in a specialized structural DB for NP research, such as the DNP[7]), but ideally, such information should be weighted in quantitative terms automatically in the future and taken into consideration for scoring.

When performing LC-MS metabolite profiling, another source of structural information that is orthogonal to MS and is linked to the physicochemical properties of the analyte is its **retention time** (RT). Being able to score calculated RTs versus experimental ones would also be important for improving metabolite annotation. In GC-MS, RT prediction tools are efficient[308], and the relative Kovats retention index is used to convert RTs into system-independent constants. In contrast to GC, in HPLC, the standardization and prediction of RT is a much more challenging task to undertake. This is due to the wide variety of chromatographic conditions available (e.g., solvents, buffers, column chemistries) and the difficulty of having access to a large set of experimental RTs for the NPs in a given condition when building a model.[171b] Some of these difficulties can be partly overcome because the large majority of NP separation is usually performed under RP-C_{18} conditions using a generic acetonitrile/water gradient with formic acid as a modifier, and large libraries of NP standards, such as the WEIZMASS spectral library, contain experimental RT values that have been reported for more than 3500 plant metabolites.[152] Furthermore, when initiatives, such as the PredRet DB[309] that collects RTs from standards across LC platforms, gain momentum the underlying models used to predict RTs across different platforms will become more accurate.[309]

To predict RTs based on structure annotation, some attempts have been made to build models that

use calculated physicochemical parameters (e.g., hydrogen bond acidity/basicity, polarity/polarizability, volume, MW, lipophilicity coefficient, topological polar surface, number of rotatable bonds) and can establish relationships with the RTs measured under generic RP metabolite profiling conditions. Quantitative structure retention relationship (QSRR) models were proposed on a representative training set of 260 non-alkaloid NPs that allowed RT predictions with an accuracy of approximately 3.3 min over a 30-min UHPLC gradient for 90% of the tested metabolites analyzed.[310] A similar type of RT prediction was calculated using a much more focused group of 91 steroid standards under any gradient mode condition. This model was able to predict RTs with an average error of 4.4%, which allowed the significant reduction of the list of steroid candidate structures associated with identical monoisotopic masses.[311] These examples show that good-quality predictions are partially achievable but mainly on a series of structurally related analytes. In biological fluid metabolomics, a set of 1955 synthetic compounds was recently analyzed and used to predict the RTs of an independent set of 202 human metabolites; the results showed reasonable predictions of this very chemodiverse group of metabolites, as well as some limitations.[312] In the current state of NP research, such tools should permit the elimination of hits that clearly cannot match the observed chromatographic behavior of the analytes of interest, but these models need to be fed with much more experimental data, which may have to be divided into chemical classes or subclasses to improve their accuracy. In addition, ways to compare RT behaviors between various LC conditions must be further investigated.

Another orthogonal dimension to MS that can be retrieved from state-of-the-art LC-MS platforms that can acquire IMS is **the collisional cross section** (CCS). CCS values are not only dependent on the molecular weight but also influenced by conformational parameters and by the molecule size of the analytes. They have a very high reproducibility (RSD <1-2%). As mentioned in section 3.1.1, IMS enables us to separate isomeric compounds based on shape; thus, associated CCS values have the potential to be used for such discrimination during annotation. Very interestingly, CCS values can be predicted based on computational and quantum chemical models, as well as machine learning predictions, such as artificial neural networks.[171b] Comparted to RT predictions, very low errors of <3% have been reported for CCS models performed using a very large dataset.[313] In this study, training was performed on 400 metabolites, and the model used 14 common molecular descriptors. A large-scale predicted CCS DB was built for 35000 metabolites from the Human Metabolome Database (HMDB).[46] Having such data at hand should effectively improve the accuracy and efficiency of identification in untargeted metabolomics, and similar efforts must be made for more accurate dereplication in NP research. Moreover, IMS could also reveal novel structural isomers that could so far not be resolved using chromatography. To effectively use this orthogonal structural information, an IMS-MS needs to be available in the lab which is currently one of the limiting factors in its more widespread use.

In addition to parameters that can be measured in metabolite profiles, such as RT and CCS, ways to quantitatively estimate the consistency of an annotation in clusters by MN approaches should be used for scoring. This task is usually performed by the visual inspection of structures, but calculating structural comparison scores, e.g., the Tanimoto index,[278] should be possible to automate.[81] In this regard, methods using the MN topology and structural similarities to improve *in silico* annotations have been implemented by creating a network consensus using re-ranked structural candidates, most notably by using Tanimoto scoring.[84b] This recently developed tool, termed "Network Annotation Propagation" (NAP), is accessible through the GNPS platform discussed in section 3.3. Such approaches allow us to rerank candidate structures that do not match the majority of scored candidates within a Molecular Family (see section 3.3) after the MN analysis of MS/MS spectra, which has been defined as "**MN consistency"**.

Finally, a significant improvement in the confidence of an annotation is obtained when **standards** are analyzed and compared to experimental data. As mentioned above, MN allows the efficient linking of structural analogues to standards. The co-occurrence of identified standards in clusters showing relationships with a metabolite of interest should also have a weight that should be considered during annotation. Similarly, the co-occurrence of substructures (Mass2Motifs) could also weight the annotation score.

To our knowledge, there are no tools that are capable of weighting various scores, as shown in Figure 5, and automatically converting them into a meta-score for a higher-confidence annotation. However, many individual methods exist, as discussed in this review, and their algorithms should be further developed and combined to enable such scoring for NP dereplication. We are convinced that this would boost the quality of the annotation of NPs in the future. Likewise, we expect that effective merging of experimental and *in silico* data would provide a significant breakthrough to the field as well.

7.2 Linking structural information between the genome and metabolome

With the increase in available whole genome sequences and so-called "paired data sets", i.e., samples for which both genome and metabolome information is available, another interesting route to obtain complementary structural information for NPs is now possible. Indeed, the biosynthesis of many NPs is encoded by a physically close set of biosynthesis genes termed biosynthesis gene clusters (BGCs). The presence of such gene clusters has been well established in microbial research and represents a growing field of investigation in plants.[314] Based on recognizable gene functions, genome mining tools such as AntiSmash[315] or PRISM[316] are able to predict such BGCs in bacterial genomes. When those BGCs can be linked to the MS/MS spectral data of the molecules they encode, valuable complementary structural information can be retrieved that is not easy to obtain using only MS.[317] For example, in

some cases, their regioselectivity and stereochemistry could be inferred based on their genome annotations. This is particularly helpful for peptidic NPs in which bacteria can alternate between L and D amino acid variants. In addition, modifications such as methylations and glycosylations could take place at different sites; in some cases, the transferase enzymes have clear preferences. If such structural information can be efficiently transferred from the genome to the metabolome, it would then be possible to further narrow down possible NP structural candidates. Vice versa, the structural information obtained from the MS/MS data could then be transferred to the genome to functionally annotate current genes with hypothetical functions. This field of research is expected to grow rapidly in the future and will very likely improve the quality of mining metabolomics and genomics data.

8 Full metabolome coverage

How far are we today from achieving full NP metabolome coverage in terms of both detection and annotation by NMR and MS?

The first point to consider is that extraction is generally necessary prior to LC-MS and NMR profiling, and as a result the chemical profiles obtained may provide a biased view of the metabolite content of an organism in its living form. Furthermore, the compartmentalization of given metabolites, such as those in plants, can generate profiles that may greater differ than those if one were to consider specialized cell types compared to the mean metabolome profile of an entire leaf.[318] Thus, the metabolite profile of a crude extract, even if it is perfect, will only represent what has been extracted,[207] and multiple extracts of different cell types would probably be required to obtain a cumulative view of a given metabolome.[319]

The NMR spectrum of an extract will theoretically provide the universal and unbiased metabolite coverage of an extract and the quantitative relationship between metabolites. A ^{1}H 1D NMR spectrum can easily be recorded under conditions that allow for concentration measurements, even in the absence of a reference compound in the sample. Such measurements proceed through direct signal integration or signal modeling.[320] This means that a compound can be detected and quantified by ^{1}H NMR in an extract, provided that the sensitivity and dynamic range of the spectrometer is sufficiently high, which represents the current main limitation of NMR toward the detection of minor metabolites. A clear advantage of NMR quantitation is that it does not need a standard because at least one of its ^{1}H NMR signals is clearly distinguishable from overlapping signals, but this condition is rarely fulfilled in complex extracts. The occurrence of intra- or inter-molecular chemical exchange constitutes a common reason for the invisibility of a present ^{1}H NMR signal. The overlapping of ^{1}H NMR signals is also a source of trouble for the identification of compounds in mixtures. Compound detection and identification by ^{13}C NMR, either directly (by 1D NMR) or indirectly (by 2D NMR), is the ultimate

technique applicable to hydrogen-poor or hydrogen-lacking organic molecules, which are rare in the field of NP chemistry. The detection of nitrogen-containing compounds by ^{15}N NMR is relevant for many bioactive NP classes, such as alkaloids and peptides. The low magnetogyric ratio of the ^{15}N nucleus (approximately one tenth of that of ^{1}H) and its low natural abundance (0.37%) prevent its direct observation. The indirectly detected ^{15}N resonances in ^{1}H-^{15}N HSQC spectra, if the ^{1}H nuclei are not chemically exchanged, and in the ^{1}H-^{15}N HMBC spectra provide complementary clues to compound identification. Similar considerations apply to the numerically smaller but biologically important class of phosphorus-containing compounds, even inorganic ones, but they have low constraints on detection sensitivity due to the 100% natural abundance of the ^{31}P nucleus and its high magnetogyric ratio (approximately 40% of that of ^{1}H). The exhaustiveness of metabolome analysis by NMR is thus fundamentally a problem of sensitivity, for which the best practical solutions are currently an increase in the static magnetic field intensity and a decrease in the probe head active volume. Possible future developments may include sensitivity enhancement by dynamic nuclear polarization and subsequent spectra recording on liquid-state samples combined with ultrafast data acquisition methods.

In contrast, LC-MS/MS metabolite detection is selective because MS requires the ionization of the analytes *prior* to detection but, compared to NMR, MS is orders of magnitude more sensitive than NMR and has a higher dynamic range. Thus, MS can provide a very large metabolite coverage provided that the necessary conditions for the ionization of the analytes are met. To be “exhaustive”, the reaction should be performed in PI and NI modes, and this reaction should be performed on LC-MS interfaces that provide complementary ionization methods to ESI. Such ionization can be obtained by APCI and/or APPI, which generates molecular ion species with different mechanisms than ESI and are better suited for less polar metabolites.[321] Ideally, state-of-the-art MS platforms should switch between all these ionization modes and ion polarities over a single run and at an acquisition frequency that is high enough to maintain a sufficiently high chromatographic resolution integrity (enough data points per LC peak over each ionization and polarity mode). Some instrument manufacturers provide ion sources with dual ionizations, but MS platforms that provide ideal coverage for all possible ionization modes in a single run do not yet exist. Instead, natural extracts are usually profiled in separate analytical runs to maintain good data quality. On the other hand, ideally, all detected methods should have their associated MS/MS spectra recorded at different CID energies. Here, as discussed in section 3.1.5, more extensive coverage will result from the usage of DIA MS/MS, provided that the associated deconvolution methods continue to improve. All of these points can be sorted out by the continuous improvement of MS platforms, mainly in terms of an increase in the frequency of acquisition and the development of sufficiently stable electronics to alternate over all acquisition modes.

In LC-MS metabolite profiling, full metabolome coverage also necessitates the use of complementary column phase chemistries for the separations of the analytes, as discussed in section 3.1.1. Currently, in NP research, mostly single profiles are acquired on RP C_{18} columns, but dual RP and HILIC separation may be envisaged, as in the case of body fluid metabolomics.[322] Furthermore, for the better coverage of the lipophilic constituent, SFC-MS must also be considered. Finally, it should be kept in mind that full metabolome coverage also encompasses volatile compound profiling, for which GC-MS must be used, and such methods is also very effective for profiling primary metabolites in natural extracts after derivatisation;[323] however, this was not the topic of the present review.

LC-MS is thus capable of providing very extensive metabolite coverage but still requires multiple injections of the same sample of different-phase chemistries and ionization modes. In the future, the improvement of methods for the data integration of multiple profiles would be beneficial to simplify this aspect. A limitation of MS compared to NMR is that no generic quantification of the detected metabolite is possible by MS alone, and complementary universal detection methods such as CAD or ELSD should be used and integrated to at least differentiate the major and minor constituents in extracts. In LC-MS, from the acquisition side, full metabolome coverage can thus probably be achieved with good spectral quality over the vast majority of NPs, but, as discussed here, the annotation workflows should still be considerably improved to annotate all the detected analytes with confidence.

9 Conclusions

Our review highlights the enormous progress that has been made in NP research over the last decade, in particular, by starting to embrace the developments made in the field of metabolomics. Moreover, it also shows that the unambiguous annotation of NPs, possibly in “all” detectable NP mixtures, is still far from reality, and interrogating what is called “dark matter” remains an important challenge. Dark matter essentially consists of chemical signatures that remain uncharacterized[23b] and encompasses compounds that remain invisible because they are not covered or detected by conventional isolation, bioactivity-based screening or other analytical techniques.[324] Furthermore, it has generally been stated that in metabolomics, less than 2% of the spectra in untargeted metabolomics can be annotated.[23b] As discussed in this review, many powerful tools are emerging approaches that may have a large impact on improving this situation, but at present, rapid unambiguous annotation can only be performed using cross searches in experimental DBs with the same acquisition conditions, while generic *in silico* DBs will mainly help with confirming the structural elucidation processes in the near future. The latter methods and workflows are not yet advanced enough to ascertain, for example, the position of a given hydroxyl group on an NP scaffold with precision. Furthermore, without NMR, determining the configuration of a carbon-bearing hydroxyl group is not possible and remains hypothetical without

considering the biosynthetic information known for a specific organism under study. One interesting complementary method to NMR, is the use of X-ray diffraction by means of the "crystalline sponge" methods, which allow the crystallographic analysis of even oily compounds in dilute quantities. It is thus well suitable for the absolute structure determination of NPs.[325] For unambiguous metabolite identification the emergence of micro electron diffraction (ED) methods such as the CryoEM method MicroED for the analysis of nanocrystals are worth mentioning. Indeed a very recent paper demonstrates that simple powders, even solids isolated *via* silica gel chromatography and dried could be directly used in MicroED studies, rapidly leading to high quality molecular structures often at atomic resolutions (better than 1Å).[326] This was even possible on heterogeneous mixture of natural products mixed together.[326]

Thus, many challenges remain during the metabolite annotation and identification process that need to be addressed in the future to assist with the structural elucidation of complex mixtures of NPs. In our view, the following set of recommendations forms the foundations upon which future solutions can be based.

Novel NP structures are discovered every day, but in practice, we cannot use all this novel information in a straightforward manner. To do so for future novel findings, we need to improve our reporting of known and novel chemical structures and their accompanying spectral data in papers using the following guidelines: i) always include a computer-extractable list of computer-readable structures with links to the spectral data; ii) share the relevant raw and spectral data in public repositories; iii) report metabolite identification confidences for all identifications and annotations; and iv) add the annotated and identified NPs to public MS/MS and/or NMR libraries. This will ease the future dereplication of identified molecules and provide the input data needed to train future annotation tools.

A wide variety of tools and pipelines are currently being developed in different groups around the globe. To further develop the field, we therefore recommend that groups working in the NP research area, to ensure the compatibility of novel tools with other processing and annotation tools, follow a modular plan to allow others to mix and match different tools into novel pipelines. Additionally, as we foresee dedicated pipelines of integrated mining and annotation tools for diverse aims, it is essential that developers clearly indicate the purpose of their tools and highlight the key areas where they can be used. It is hereby important to keep the end result user-friendly and provide typical examples of use cases. Finally, the number of adjustable parameters should be kept to a minimum, as it can have a massive impact on the outcome, which is often not communicated very clearly with the end users.

An increasing number of databases has appeared, each of which covers different aspects of metabolite

metadata. Ensuring structural consistency between those databases is essential to allow the linking of the collected metadata. Moreover, the creation of freely accessible metabolite resources with extensive and community-curated metadata will further assist in the selection of candidates from the lists of possible structures produced by annotation tools, as described above.

In summary, most of these recommendations require the NP field to think like a large community rather than as individual islands of knowledge: if we all keep adding findable, accessible, interoperable, and reproducible (i.e., following the FAIR principles) NP structures to the existing large pool of known NPs, we can all more effectively exploit the fruits of our labor as a community to illuminate the "dark matter" in the field of NP chemistry.

NP chemists have been and are still involved in the unambiguous *full de novo* identification of unknown compounds from complex natural biological matrices. Their involvement in metabolomics can help with the development of methods to improve metabolite ID, a key point that represents a major bottleneck in the metabolomics field. With the tremendous development of bioinformatics tools, analytical platforms and the access to increasing amounts of genome data, it is a safe bet to say that NP metabolomics will evolve rapidly toward full metabolome annotations of organisms and change the paradigm on how NP research is performed, which will open new gateways to discoveries.

10 Author information

Corresponding Author

*E-mail: jean-luc.wolfender@unige.ch. Phone: 41-22 379-3385.

ORCID

Jean-luc Wolfender: 0000-0002-0125-952

Notes

The authors declare no competing financial interest.

11 brief biographies

11.1 Jean-Luc Wolfender

Jean-Luc Wolfender is a chemist, heading a group in Natural Product research at the School of Pharmaceutical Sciences of the University of Geneva (Switzerland). He was strongly involved in the 90s in the introduction of LC-MS and LC-NMR for the profiling of crude plants extracts for dereplication purposes. He is currently developing innovative MS- and NMR- based metabolomics strategies in

natural product based drug discovery and chemical ecology projects. His main research interests focus on the search of novel inducible bioactive natural products in response to various biotic and abiotic stimuli as well on the study of the mode of action of phytopharmaceuticals from a systems biology perspective. He his involved in promotion of metabolomics within the natural product community.

11.2 Jean-Marc Nuzillard

Jean-Marc Nuzillard is Research Director at the French National Council for Scientific Research, in the Reims Institute of Molecular Chemistry. After a PhD in asymmetric hydrogenation catalyzed by rhodium complexes, he started a research program for the synthesis of indole alkaloids in an environment of natural product chemists. The need for an accurate structure determination of complex natural compounds incited him to acquire some knowledge of the physics of NMR and to write his own data acquisition sequences. In parallel, he developed computer codes for automatic structure verification and elucidation from 2D NMR data, based on the concepts of artificial intelligence. Presently, his main research topic concerns the analysis of complexes mixtures by NMR.

11.3 Justin J. J. van der Hooft

Justin JJ van der Hooft is an analytical biochemist by training and currently a postdoctoral researcher in metabolome and genome mining on an eScienceCenter/NWO Accelerating Scientific Discoveries research grant in the Bioinformatics Group at Wageningen University, The Netherlands. He obtained his BSc and Masters in Molecular Sciences, and a PhD in systematic metabolite identification at Wageningen University, NL. His research is focusing on closing the structural annotation gap in metabolomics. He is equipped with hands-on experience in setting up mass spectrometry fragmentation and NMR workflows dedicated to systematic metabolite annotation and identification and spectral analysis of the resulting data. Furthermore, Justin has knowhow of existing state-of-the-art metabolome mining and annotation tools and he has developed MS2LDA for automated substructure discovery from metabolomics data. He is now working on integrating genome mining and metabolome mining workflows to enhance natural product discovery workflows. Justin is actively involved in the metabolomics community currently as Director of the Metabolomics Society. He is a strong advocate of metabolomics and computational metabolomics metabolite annotation workflows in particular.

11.4 Jean-Hugues Renault

Jean-Hugues Renault is Professor of Pharmacognosy at the University of Reims Champagne Ardenne (URCA), and he heads the Institute of Molecular Chemistry of Reims (UMR 7312), a research unit affiliated with the University of Reims Champagne Ardenne and the French National Centre for Scientific Research. He was the coordinator of research activities of the Health sector at the University

of Reims Champagne Ardenne (2012-2016). He also headed a master degree on “chemistry, natural products and drugs” during 17 years. Since the end of the 90’s, his research is focusing - at the interface of Natural Product Chemistry and Process Engineering – on the development, the modelling and the intensification of original purification processes by Centrifugal Partition Chromatography. Since 8 years, his research interests focus also - at the interface of Natural Product Chemistry and Chemoinformatics – on the development of original tools for the chemical profiling and the dereplication of complex mixture, involving NMR and Centrifugal partition Chromatography.

11.5 Samuel Bertrand

Samuel Bertrand is a Chemist, currently Associate Professor at “Laboratoire Mer Molécules Santé – EA2160” (See, Molecules, Health laboratory), of the UFR des Sciences Pharmaceutiques et Biologiques (School of pharmaceutical and biological sciences), University of Nantes, France. He obtains his PhD in Fungal Natural Product Chemistry in the University of Angers (France), and moved to a post-doctoral position in Natural Product research at the School of Pharmaceutical Sciences of the University of Geneva (Switzerland). His main field of interest is the use of LC-MS Metabolomics and Lipidomics to rationalize natural product drug discovery, with a focus on annotation workflows. He mainly applies such approaches on fungal co-cultures in drug discovery and chemical ecology context.

12 Acknowledgments

JLW is grateful to the Swiss National Science Foundation (SNF) for supporting its natural product metabolomics projects (grants nos. 310030E-164289, 31003A_163424 and 316030_164095). JJJvdH acknowledges an ASDI eScience grant (ASDI.2017.030) from the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Dr Pierre-Marie Allard (Wolfender’s Lab) for constructive discussion during the preparation of this review and for the creation of Figure 2.

13 Figure Captions

Figure 1. Schematic workflow of structure elucidation/dereplication in natural product chemistry. The principal task consists of connecting the space of samples such as extracts, chemically simplified fractions (FRs) or isolated compounds (NPs) (left panel) and the space of molecular structures (right panel). Extracts are obtained by different extractions process that lead to complex mixtures of NPs with given physicochemical properties according to the nature of the solvent used. Fractions and pure NPs are obtained after single or multiple preparative chromatographic steps. This task is achieved by a combination of physico-chemical spectroscopic methods, mainly MS and NMR (central panel). ‘Others’ indicate additional method eg. X-ray diffractions for pure NPs, LC-ECD for fractions or extracts. When a mixture of NPs is submitted to spectroscopic analysis, often an orthogonal analytical

separation method is used prior spectral acquisition (liquid chromatography, ion mobility, etc.). The space of physico-chemical spectroscopic data is divided into two subspaces; i) acquired "raw data" (e.g. FID time domain data in NMR, LC-MS raw data files) ii) "processed data" (e.g. NMR spectra expressed in HZ/ppm calculated by the Fourier transformation FIDs, Peak picking of MS features and combination of related MS and MS/MS spectra in LC-MS). Molecular structure determination results from data interpretation through two different strategies: *de novo* structure elucidation or dereplication. The latter is generally computer-assisted for the database search step, whereas the *de novo* approach resorts from manual or computer-assisted strategies interpretation of the spectroscopic data. Connection between NP mixtures and the space of molecular structures may involve the use of chemometrics for deconvolution purposes for finally generating composition information on extracts or fractions. Consistency of the structural data generated are cheek based on taxonomy and the known biosynthetic pathways of the organisms studied.

Figure 2. Example of data acquisition in a typical UHPLC-HRMS/MS metabolite profiling of plant extracts in both data dependent (DDA) and data independent (DIA) MS/MS modes. The analysis of a mixture of five plants extracts presenting a broad chemodiversity is shown.[119] The HRMS and MS/MS spectra are displayed for a specific feature *m/z* 567.17@3.66 min corresponding to the $[M+H]^+$ of the biflavonoid isoginkgetin (Accurate Mass: 566.1213; Formula: $C_{32}H_{22}O_{10}$) present in the extract of *Gingko biloba*, one the extracts of the mix. A) UHPLC-ESI-HRMS metabolite profile acquired in the in PI mode (m/z 150-1200) on an Orbitrap mass spectrometer on a C_{18} column (50x2.0mm i.d.; 1.7 µm) with a generic acetonitrile gradient 5-95% in 8 min for a broad profiling over a large NP polarity range. B) Visualization in the form of an ion map (*m/z* vs RT) of all features having an associated MS/MS with the MS-Dial software[120] (2631 features with MS^2). On the attached panel isotopic clustering of MS-Dial of all features corresponding to a given analyte. C) Same plot as (B) for data acquired in the DIA mode (18420 features with MS^2) showing an increased coverage of fragmentation data compared to DDA. D) HRMS spectrum recorded at the apex of the LC-peak at RT 3.66 displaying the features *m/z* 567.12 $[M+H]^+$ of isoginkgetin as well as its dimer *m/z* 1133.25 $[2M+H]^+$ obtained with ESI in positive ionisation mode. E) Zoom on the feature $[M+H]^+$ of isoginkgetin showing the accurate mass and isotopic pattern of $[M+H]^+$. This information is necessary to ascertain the corresponding molecular formula (MF): $C_{32}H_{22}O_{10}$ (calculated for $C_{32}H_{22}O_{10}$ 566.1213, Δppm = 0.5). F) DDA MS/MS spectrum of the precursor ion *m/z* 567.12 automatically selected during profiling and fragmented with HCD at 3 different normalized collision energies (NCE 15, 30 and 45) on the Orbitrap MS analyser. G) Raw DIA M/MS spectrum at RT 3.66 min. All ions are fragmented at 3 different NCE. H) Superposition of all ion traces of the fragment ions for selection of ions coeluting in LC at the apex of RT 3.66 min for the deconvolution of the raw DIA MS/MS spectrum (F). I) Deconvoluted DIA spectrum at RT 3.66 min

associated to the feature *m/z* 567.12 eluting at the same retention time in the full scan spectra (D). Comparison of the spectra (G) and (I) allows a comparison of the DIA deconvolution, the DDA MS/MS spectrum (F) does not requires deconvolution since a specific precursor ion is selectively selected. J) *In silico* simulated spectrum of isoginkgetin obtained by CFM-ID[121] by input of its SMILES structural string. To be noted in the selected example a few fragments are common between the different MS/MS modes and the *in silico* spectra generated. The richness of fragment information generated is dependent on the type of NP scaffolds analysed. Such information in addition to the MF formula assignment already allows a significant reduction of structural candidates. In this case, the isoginkgetin standard was also analysed under the same conditions and its identity was confirmed.

Figure 3. Top left to right – Key Molecular Networking concepts: all MS/MS spectra within a sample or across different samples are compared based on the similarity of mass fragments. The parent mass difference (indicated by blue dotted arrow) is considered by shifting both spectra with this *m/z* value and checking for improved or additional matches that will add to the final similarity score (left panel, bottom). Using all these comparison scores, Molecular Families can be formed (middle panel), where fragmented molecules are the nodes and connections (edges) are present when the similarity score is above a user-defined threshold. Moreover, different layers of information can be displayed on the nodes and edges, such as where molecules result in a library match from reference MS/MS spectra (node in blue). Applying this to all the fragmented molecules typically results in a Molecular Network (MN) consisting of larger and smaller Molecular Families (right panel) as well as unconnected molecules (singletons). Bottom left to right – conceptualized MS2LDA substructure search: MS2LDA starts with a large set of MS/MS spectra from one or multiple samples (left panel) and then searches for Mass2Motifs (middle panel). These Mass2Motifs consist of commonly occurring mass fragments and/or neutral losses that can then be annotated by substructures. In the middle panel, one mass fragment-based (purple) and one neutral loss-based (green) Mass2Motif are exemplified, where each Mass2Motif is present in a spectrum where its corresponding substructure is indicated in the structure on the left. Both Mass2Motifs are present in one MS/MS spectrum and in the corresponding structure on the right, and both corresponding substructures show how in this case the complete structure can be built from its substructures. Finally, by collecting all of the connections between fragmented molecules and Mass2Motifs, a substructure network can be formed that connects Mass2Motifs/substructures (orange circles) with fragmented molecules (blue squares), as displayed on the right end.

Figure 4. Principle of structure generation by the LSD CASE software. The goal of the LSD software is to draw bonds between initially non-bonded atoms to find the possible planar solutions of a *de novo* structure elucidation problem. The number and nature of the atoms of the solution structure is

supposed to be known from HRMS data. The set of atom statuses must be fully determined before the beginning of the problem resolution process. In the case this could not be reliably achieved, a software layer written above LSD, named pyLSD[267, 306] (a), can be invoked to resolve status ambiguities. LSD is not aware of the relationships between chemical shifts and structural features and places bonds between atoms solely based on 2D NMR correlation data. The combination of COSY and HSQC data yields bonds between heavy (i.e., non-hydrogen) atoms from which the resolution process starts. The combination of HSQC and HMBC data yields proximity relationships between heavy atoms expressed as distances measured in number of bonds (1, 2 or more). HMBC correlations of close ^{13}C resonances may be declared ambiguous, and all possible interpretations will be systematically considered. The resolution process starts by the recursive use of proximity relationships for the formation of bonds and removes those that become explained by the newly formed bonds. The atoms for which not all of their bonds are present, as inferred from their status, are then systematically paired in a recursive process to build complete structures. Recursive processes are needed to explore all possibilities and reconsider choices (backtracking) opened by data interpretation so that the exhaustivity of the solution search is ensured. Each structure then passes through a series of validation steps. The distances between atoms are checked to address the HMBC correlations through 4 bonds or more. Double and triple bonds are placed between atoms to reach the needed coherence with their hybridization state. Anti-Bredt structures are eliminated. (b) The user may impose sub-structural elements to be present or absent in the solution structures (any combination of such constraints is allowed) according to external information sources, possibly spectroscopic or biogenetic. Ambiguous HMBC correlations may lead to duplicated solutions that need to be removed. The pyLSD software layer sorts the solutions according to the similarity of the ^{13}C NMR chemical shifts with those predicted by NMRshiftDB.

Figure 5. Schematic view of possibilities to improve the annotation scoring by weighting different partial and imperfect scores given by all methods on both LC-MS and NMR for maximum confidence in the compound identification (ID) process. NMR and MS are the main spectroscopic techniques for the identification of a chemical structure. Whether on the NMR side or the MS side, orthogonal techniques are used to increase the ID level of confidence (e.g., liquid chromatography for both MS and NMR often offline, or ionic mobility on the MS side). Moreover, some extra-data (brown boxes) are combined with spectroscopic data, leading to workflows that can be organized in the same way for MS and NMR approaches. They start from the lowest ID level ($score^{MS\ or\ NMR}$ 1) when only (HR)MS or 1D NMR are recorded and compared with calculated or experimental spectra, for instance, by means of *in silico* or experimental databases, and they reach the highest score ($score^{MS\ or\ NMR}$ 4) after complementary data acquisition and processing (e.g., MS^n, 2D NMR) and usage of extra-data (e.g., taxonomy, biosynthetic pathways, NMR data for MS-based workflows and MS data for NMR ones).

Thus ideally a meta-score (meta-score^{NMR} or meta-score^{MS}) resulting from the addition of the scores validated for the NMR side ($\sum_{x=1}^{4} \text{score}^{NMR}x$) or the MS ($\sum_{x=1}^{4} \text{score}^{MS}x$) side can be thus calculated. For 3D sutural aspects relative configuration is generally proposed by combining coupling constant analysis and the Overhauser effect in NMR, sometimes together with conformational analysis and molecular modeling. Absolute configuration determination must resort from chiroptical techniques such as ECD, VCD or X-ray crystallography (score^{CHIR}). If all aspects can be taken into consideration with a correct weighting of all scores, in future, a GLOBAL-meta-score could ideally be defined as a combination of the three individual meta-scores. Exclamation marks (⚠) highlight critical steps in metabolite ID.

14 References

1. Grotewold, E., Plant metabolic diversity: a regulatory perspective. *Trends Plant Sci.* **2005,** *10* (2), 57-62.
2. Pichersky, E.; Lewinsohn, E., Convergent Evolution in Plant Specialized Metabolism. *Ann rev Plant Biol* **2011,** *62* (1), 549-566.
3. David, B.; Wolfender, J.-L.; Dias, D. A., The pharmaceutical industry and natural products: historical status and new trends. *Phytochem. Rev.* **2014,** *14* (2), 299-315.
4. Bernardini, S.; Tiezzi, A.; Laghezza Masci, V.; Ovidi, E., Natural products for human health: an historical overview of the drug discovery approaches. *Nat. Prod. Res.* **2018,** *32* (16), 1926-1950.
5. (a) Ngo, L. T.; Okogun, J. I.; Folk, W. R., 21st Century natural product research and drug development and traditional medicines. *Nat. Prod. Rep.* **2013,** *30* (4), 584-592; (b) Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016,** *79* (3), 629-661.
6. Hartmann, T., From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* **2007,** *68* (22), 2831-2846.
7. Chapman; Hall Dictionary of Natural Products. http://dnp.chemnetbase.com/.
8. (a) Dobson, C. M., Chemical space and biology. *Nature* **2004,** *432*, 824-828; (b) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S., Progress in visual representations of chemical space. *Expert Opin. Drug Discov.* **2015,** *10* (9), 959-973.
9. Feher, M.; Schmidt, J. M., Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **2003,** *43* (1), 218-227.
10. Harvey, A. L.; Clark, R. L.; Mackay, S. P.; Johnston, B. F., Current strategies for drug discovery through natural products. *Expert Opin. Drug Discov.* **2010,** *5* (6), 559-568.
11. Rosén, J.; Lövgren, A.; Kogej, T.; Muresan, S.; Gottfries, J.; Backlund, A., ChemGPS-NPWeb: chemical space navigation online. *J. Comput. Aided Mol. Des.* **2009,** *23* (4), 253-259.
12. Koehn, F. E.; Carter, G. T., The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discovery* **2005,** *4* (3), 206-220.
13. Shrestha, G.; St. Clair, L. L.; O'Neill, K. L., The Immunostimulating Role of Lichen Polysaccharides: A Review. *Phytother. Res.* **2014,** *29* (3), 317-322.
14. Chiurchiù, V.; Maccarrone, M., Bioactive lipids as modulators of immunity, inflammation and emotions. *Curr. Opin. Pharmacol.* **2016,** *29*, 54-62.
15. Hooft, J. J.; Vos, R. H.; Ridder, L.; Vervoort, J.; Bino, R., Structural elucidation of low abundant metabolites in complex sample matrices. *Metabolomics* **2013,** *9* (5), 1009-1018.
16. Hostettmann, K.; Wolfender, J.-L.; Terreaux, C., Modern Screening Techniques for Plant Extracts. *Pharm. Biol.* **2001,** *39* (s1), 18-32.

17. Van Voorhis, W. C.; Hooft van Huijsduijnen, R.; Wells, T. N. C., Profile of William C. Campbell, Satoshi Ōmura, and Youyou Tu, 2015 Nobel Laureates in Physiology or Medicine. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (52), 15773-15776.
18. Hubert, J.; Nuzillard, J.-M.; Renault, J.-H., Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochem. Rev.* **2015,** *16* (1), 55-95.
19. Wolfender, J.-L.; Marti, G.; Thomas, A.; Bertrand, S., Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* **2015,** *1382*, 136–164.
20. Guy, C.; Kopka, J.; Moritz, T., Plant metabolomics coming of age. *Physiol. Plant.* **2008,** *132* (2), 113-116.
21. Sumner, L. W.; Mendes, P.; Dixon, R. A., Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **2003,** *62* (6), 817-836.
22. Nguyen, D. H.; Nguyen, C. H.; Mamitsuka, H., Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinformatics* **2018**, bby066-bby066.
23. (a) Dark matter. *Nature* **2008,** *455* (7213), 698; (b) da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A., Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (41), 12549-50; (c) Peisl, B. Y. L.; Schymanski, E. L.; Wilmes, P., Dark matter in host-microbiome metabolomics: Tackling the unknowns–A review. *Anal. Chim. Acta* **2018,** *1037*, 13-27.
24. Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.; Neumann, S.; Schymanski, E. L.; Sumner, L. W.; Trengove, R.; Wolfender, J.-L., Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics* **2014,** *10* (3), 350-353.
25. Dias, D.; Jones, O.; Beale, D.; Boughton, B.; Benheim, D.; Kouremenos, K.; Wolfender, J.-L.; Wishart, D., Current and Future Perspectives on the Structural Identification of Small Molecules in Biological Systems. *Metabolites* **2016,** *6* (4), 46.
26. Clardy, J.; Walsh, C., Lessons from natural molecules. *Nature* **2004,** *432*, 829.
27. Verpoorte, R.; Choi, Y. H.; Kim, H. K., Ethnopharmacology and systems biology: A perfect holistic match. *J. Ethnopharmacol.* **2005,** *100* (1), 53-56.
28. Cao, H.; Zhang, A.; Zhang, H.; Sun, H.; Wang, X., The application of metabolomics in traditional Chinese medicine opens up a dialogue between Chinese and Western medicine. *Phytother. Res.* **2015,** *29* (2), 159-166.
29. Lee, K. M.; Jeon, J. Y.; Lee, B. J.; Lee, H.; Choi, H. K., Application of Metabolomics to Quality Control of Natural Product Derived Medicines. *Biomol. Ther.* **2017,** *25* (6), 559-568.
30. Hu, C.; Xu, G., Metabolomics and traditional Chinese medicine. *Trends Anal. Chem.* **2014,** *61*, 207-214.
31. Schymanski, E. L.; Williams, A. J., Open Science for Identifying "Known Unknown" Chemicals. *Environ. Sci. Technol.* **2017,** *51* (10), 5357-5359.
32. Saielli, G.; Bagno, A., Can two molecules have the same NMR spectrum? Hexacyclinol revisited. *Org. Lett.* **2009,** *11* (6), 1409-12.
33. (a) McAlpine, J. B.; Chen, S.-N.; Kutateladze, A.; MacMillan, J. B.; Appendino, G.; Barison, A.; Beniddir, M. A.; Biavatti, M. W.; Bluml, S.; Boufridi, A.; Butler, M. S.; Capon, R. J.; Choi, Y. H.; Coppage, D.; Crews, P.; Crimmins, M. T.; Csete, M.; Dewapriya, P.; Egan, J. M.; Garson, M. J.; Genta-Jouve, G.; Gerwick, W. H.; Gross, H.; Harper, M. K.; Hermanto, P.; Hook, J. M.; Hunter, L.; Jeannerat, D.; Ji, N.-Y.; Johnson, T. A.; Kingston, D. G. I.; Koshino, H.; Lee, H.-W.; Lewin, G.; Li, J.; Linington, R. G.; Liu, M.; McPhail, K. L.; Molinski, T. F.; Moore, B. S.; Nam, J.-W.; Neupane, R. P.; Niemitz, M.; Nuzillard, J.-M.; Oberlies, N. H.; Ocampos, F. M. M.; Pan, G.; Quinn, R. J.; Reddy, D. S.; Renault, J.-H.; Rivera-Chávez, J.; Robien, W.; Saunders, C. M.; Schmidt, T. J.; Seger, C.; Shen, B.; Steinbeck, C.; Stuppner, H.; Sturm, S.; Taglialatela-Scafati, O.; Tantillo, D. J.; Verpoorte, R.; Wang, B.-G.; Williams, C. M.; Williams, P. G.; Wist, J.; Yue, J.-M.; Zhang, C.; Xu, Z.; Simmler, C.; Lankin, D. C.; Bisson, J.; Pauli, G. F., The value of universally available raw NMR data for transparency, reproducibility, and integrity in natural product research. *Nat. Prod. Rep.* **2018**, DOI: 10.1039/c7np00064b; (b) Bisson, J.; Simmler, C.; Chen, S.-N.; Friesen, J. B.; Lankin, D. C.; McAlpine, J. B.; Pauli, G. F., Dissemination of original NMR data enhances reproducibility and integrity in chemical research. *Nat. Prod. Rep.* **2016,** *33* (9), 1028-33.

34. Pupier, M.; Nuzillard, J. M.; Wist, J.; Schlorer, N. E.; Kuhn, S.; Erdelyi, M.; Steinbeck, C.; Williams, A. J.; Butts, C.; Claridge, T. D. W.; Mikhova, B.; Robien, W.; Dashti, H.; Eghbalnia, H. R.; Fares, C.; Adam, C.; Kessler, P.; Moriaud, F.; Elyashberg, M.; Argyropoulos, D.; Perez, M.; Giraudeau, P.; Gil, R. R.; Trevorrow, P.; Jeannerat, D., NMReDATA, a standard to report the NMR assignment and parameters of organic compounds. *Magn. Reson. Chem.* **2018,** *56* (8), 703-715.
35. (a) Viant, M. R.; Kurland, I. J.; Jones, M. R.; Dunn, W. B., How close are we to complete annotation of metabolomes? *Curr. Opin. Chem. Biol.* **2017,** *36*, 64-69; (b) Aksenov, A. A.; da Silva, R.; Knight, R.; Lopes, N. P.; Dorrestein, P. C., Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **2017,** *1*, 0054; (c) Xie, T.; Song, S.; Li, S.; Ouyang, L.; Xia, L.; Huang, J., Review of natural product databases. *Cell Prolif.* **2015,** *48* (4), 398-404.
36. Laatsch, H., *AntiBase 2013: The Natural Compound Identifier, Upgrade*. Wiley-Vch: 2013.
37. Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L., BioMagResBank. *Nucleic Acids Res.* **2008,** *36* (suppl_1), D402-D408, URL: http://www.bmrb.wisc.edu/.
38. Asakura, K., A NMR Spectral Database of Natural Products "CH-NMR-NP". *J. Synth. Org. Chem. Japan* **2015,** *73* (12), 1247-1252, URL: https://www.j-resonance.com/en/nmrdb/.
39. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C., ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016,** *44* (D1), D1214-1219, URL: https://www.ebi.ac.uk/chebi/.
40. Zani, C. L.; Carroll, A. R., Database for Rapid Dereplication of Known Natural Products Using Data from MS and Fast NMR Experiments. *J. Nat. Prod.* **2017,** *80* (6), 1758-1766.
41. Blunt, J. W.; Munro, M. H. G., Dictionary of Marine Natural Products. Blunt, J. W.; Munro, M. H. G., Eds. CRC Press, URL: http://dmnp.chemnetbase.com/.
42. Guo, A. C.; Jewison, T.; Wilson, M.; Liu, Y.; Knox, C.; Djoumbou, Y.; Lo, P.; Mandal, R.; Krishnamurthy, R.; Wishart, D. S., ECMDB: The *E. coli* Metabolome Database. *Nucleic Acids Res.* **2013,** *41* (Database issue), D625-D630, URL: http://www.ecmdb.ca.
43. FooDB food component database. URL: http://foodb.ca/.
44. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crusemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderon, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodriguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Muller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. O.; Pogliano, K.; Linington, R. G.; Gutierrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016,** *34* (8), 828-837, URL: https://gnps.ucsd.edu/.
45. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D., GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2005,** *21* (8), 1635-1638, URL http://gmd.mpimp-golm.mpg.de.
46. Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.;

Johnson, D.; Li, C.; Karu, N.; Sayeeda, Z.; Lo, E.; Assempour, N.; Berjanskii, M.; Singhal, S.; Arndt, D.; Liang, Y.; Badran, H.; Grant, J.; Serra-Cayuela, A.; Liu, Y.; Mandal, R.; Neveu, V.; Pon, A.; Knox, C.; Wilson, M.; Manach, C.; Scalbert, A., HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **2018,** *46* (D1), D608-D617.

47. Shinbo, Y.; Nakamura, Y.; Altaf-Ul-Amin, M.; Asahi, H.; Kurokawa, K.; Arita, M.; Saito, K.; Ohta, D.; Shibata, D.; Kanaya, S., KNApSAcK: A comprehensive species-metabolite relationship database. In *Plant Metabolomics*, Saito, K.; Dixon, R.; Willmitzer, L., Eds. Springer Berlin Heidelberg: 2006; Vol. 57, pp 165-181, URL: http://kanaya.naist.jp/KNApSAcK/KNApSAcK.php.

48. Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L., Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008,** *26* (2), 162-164, URl: http://mmcd.nmrfam.wisc.edu/.

49. MarinLit, a database of the marine natural products literature. Dabb, S.; Potter, H., Eds. Royal Society of Chemistry 2014.

50. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T., MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010,** *45* (7), 703-714, URL: http://www.massbank.jp/.

51. NORMAN Association European MassBank (NORMAN MassBank). https://massbank.eu/MassBank/.

52. (a) Steinbeck, C.; Conesa, P.; Haug, K.; Mahendraker, T.; Williams, M.; Maguire, E.; Rocca-Serra, P.; Sansone, S.-A.; Salek, R.; Griffin, J., MetaboLights: towards a new COSMOS of metabolomics data management. *Metabolomics* **2012,** *8* (5), 757-760, URL: http://www.ebi.ac.uk/metabolights; (b) Haug, K.; Salek, R. M.; Conesa, P.; Hastings, J.; de Matos, P.; Rijnbeek, M.; Mahendraker, T.; Williams, M.; Neumann, S.; Rocca-Serra, P.; Maguire, E.; González-Beltrán, A.; Sansone, S.-A.; Griffin, J. L.; Steinbeck, C., MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013,** *41* (Database issue), D781-D786.

53. Karp, P. D.; Riley, M.; Paley, S. M.; Pellegrini-Toole, A., The MetaCyc Database. *Nucleic Acids Res.* **2002,** *30* (1), 59-61, URL: http://www.metacyc.org.

54. Mihaleva, V. V.; te Beek, T. A.; van Zimmeren, F.; Moco, S.; Laatikainen, R.; Niemitz, M.; Korhonen, S. P.; van Driel, M. A.; Vervoort, J., MetIDB: a publicly accessible database of predicted and experimental 1H NMR spectra of flavonoids. *Anal. Chem.* **2013,** *85* (18), 8700-8707, URL: www.metidb.org/.

55. Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G., METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* **2005,** *27* (6), 747-751, URL: http://metlin.scripps.edu/index.php.

56. Sakurai, N.; Ara, T.; Kanaya, S.; Nakamura, Y.; Iijima, Y.; Enomoto, M.; Motegi, T.; Aoki, K.; Suzuki, H.; Shibata, D., An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values. *Bioinformatics* **2013,** *29* (2), 290-291, URL: http://webs2.kazusa.or.jp/mfsearcher/.

57. MassBank of North America (MoNA). http://mona.fiehnlab.ucdavis.edu/.

58. Draper, J.; Enot, D.; Parker, D.; Beckmann, M.; Snowdon, S.; Lin, W.; Zubair, H., Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour 'rules'. *BMC Bioinf.* **2009,** *10* (1), 227, URL: http://maltese.dbs.aber.ac.uk:8888/hrmet/index.html.

59. Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; A. Moumbock, A. F.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S., NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, DOI: 10.1021/acs.jnatprod.7b00283, URL: http://african-compounds.org/nanpdb/.

60. National Institute of Standards and Technology NIST Mass Spectral Library.

https://chemdata.nist.gov/.
61. The Natural Products Atlas. http://www.npatlas.org/joomla/.
62. Pupin, M.; Esmaeel, Q.; Flissi, A.; Dufresne, Y.; Jacques, P.; Leclère, V., Norine : a powerful resource for novel nonribosomal peptide discovery. *Synth. Syst. Biotechnol.* **2016,** *1* (2), 89-94, URL: http://bioinfo.lifl.fr/NRP/.
63. Choi, H.; Cho, S. Y.; Pak, H. J.; Kim, Y.; Choi, J.-y.; Lee, Y. J.; Gong, B. H.; Kang, Y. S.; Han, T.; Choi, G.; Cho, Y.; Lee, S.; Ryoo, D.; Park, H., NPCARE: database of natural products and fractional extracts for cancer regulation. *J. Cheminformatics* **2017,** *9* (1), 2, URL: http://silver.sejong.ac.kr/npcare.
64. Schläpfer, P.; Zhang, P.; Wang, C.; Kim, T.; Banf, M.; Chae, L.; Dreher, K.; Chavali, A. K.; Nilo-Poyanco, R.; Bernard, T.; Kahn, D.; Rhee, S. Y., Genome-Wide Prediction of Metabolic Enzymes, Pathways, and Gene Clusters in Plants. *Plant Physiol* **2017,** *173* (4), 2041, URL: https://www.plantcyc.org.
65. Sawada, Y.; Nakabayashi, R.; Yamada, Y.; Suzuki, M.; Sato, M.; Sakata, A.; Akiyama, K.; Sakurai, T.; Matsuda, F.; Aoki, T.; Hirai, M. Y.; Saito, K., RIKEN tandem mass spectral database (ReSpect) for phytochemicals: A plant-specific MS/MS-based data resource and database. *Phytochemistry* **2012,** *82*, 38-45, URL: http://spectra.psc.riken.jp/.
66. Davis, G. D. J.; Vasanthi, A. H. R., Seaweed metabolite database (SWMD): A database of natural compounds from marine algae. *Bioinformation* **2011,** *5* (8), 361-364, URL: http://www.swmd.co.in/home.html.
67. Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K. K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O. S.; Bechthold, A.; Günther, S., StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res.* **2016,** *44* (D1), D509-D514, URL: http://132.230.56.4/streptomedb2/.
68. Preissner, R.; Dunkel, M.; Banerjee, P.; Erehman, J.; Gohlke, B. O., Super Natural II - a database of natural products URL: http://bioinf-applied.charite.de/supernatural_new/.
69. Allard, P.-M.; Péresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V. C.; Roussi, F.; Litaudon, M.; Wolfender, J.-L., Integration of molecular networking and *in-silico* MS/MS fragmentation for natural products dereplication. *Anal. Chem.* **2016,** *88* (6), 3317-3323, URL: http://oolonek.github.io/ISDB/.
70. Jewison, T.; Knox, C.; Neveu, V.; Djoumbou, Y.; Guo, A. C.; Lee, J.; Liu, P.; Mandal, R.; Krishnamurthy, R.; Sinelnikov, I.; Wilson, M.; Wishart, D. S., YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.* **2012,** *40* (D1), D815-D820, URL: http://www.ymdb.ca.
71. Frainay, C.; Schymanski, E.; Neumann, S.; Merlet, B.; Salek, R.; Jourdan, F.; Yanes, O., Mind the Gap: Mapping Mass Spectral Databases in Genome-Scale Metabolic Networks Reveals Poorly Covered Areas. *Metabolites* **2018,** *8* (3), 51.
72. Prigent, S.; Nielsen, J. C.; Frisvad, J. C.; Nielsen, J., Reconstruction of 24 Penicillium genome-scale metabolic models shows diversity based on their secondary metabolism. *Biotechnol. Bioeng.* **2018,** *115* (10), 2604-2612.
73. Allard, P.-M.; Bisson, J.; Azzollini, A.; Pauli, G. F.; Cordell, G. A.; Wolfender, J.-L., Pharmacognosy in the digital era: shifting to contextualized metabolomics. *Curr. Opin. Biotechnol.* **2018,** *54*, 57-64.
74. (a) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D., InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **2015,** *7* (1), 23; (b) Akhondi, S. A.; Kors, J. A.; Muresan, S., Consistency of systematic chemical identifiers within and between small-molecule databases. *J. Cheminformatics* **2012,** *4* (1), 35.
75. Pence, H. E.; Williams, A., ChemSpider: An online chemical information resource. *J. Chem. Educ.* **2010,** *87* (11), 1123-1124, URL: http://www.chemspider.com/.
76. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016,** *44* (D1), D1202-D1213, URL: https://pubchem.ncbi.nlm.nih.gov.
77. Chemical American Society SciFinder. https://www.cas.org/products/scifinder.
78. Bertrand, S.; Roullier, C.; Guitton, Y., Successes and Pitfalls in Automated Dereplication Strategy using Mass Spectrometry Data: a CASMI Experience. *Curr. Metabolomics* **2017,** *5* (1), 25-34.

79. Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H. Z.; Xu, X., Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* **2013,** *8* (4), e62839.
80. Moco, S.; Bino, R. J.; Vorst, O.; Verhoeven, H. A.; de Groot, J.; van Beek, T. A.; Vervoort, J.; de Vos, C. H. R., A Liquid Chromatography-Mass Spectrometry-Based Metabolome Database for Tomato. *Plant Physiol* **2006,** *141* (4), 1205.
81. Allard, P.-M.; Genta-Jouve, G.; Wolfender, J.-L., Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Curr. Opin. Chem. Biol.* **2017,** *36*, 40-49.
82. Tsipouras, A.; Ondeyka, J.; Dufresne, C.; Lee, S.; Salituro, G.; Tsou, N.; Goetz, M.; Singh, S. B.; Kearsley, S. K., Using similarity searches over databases of estimated 13C NMR spectra for structure identification of natural product compounds. *Anal. Chim. Acta* **1995,** *316* (2), 161-171.
83. (a) Heinonen, M.; Rantanen, A.; Mielikäinen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R. A.; Rousu, J., FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Commun. Mass Spectrom.* **2008,** *22* (19), 3043-3052; (b) Yesiltepe, Y.; Nuñez, J. R.; Colby, S. M.; Thomas, D. G.; Borkum, M. I.; Reardon, P. N.; Washton, N. M.; Metz, T. O.; Teeguarden, J. G.; Govind, N.; Renslow, R. S., An automated framework for NMR chemical shift calculations of small organic molecules. *J. Cheminformatics* **2018,** *10* (1), 52.
84. (a) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M.; Beger, R.; Daykin, C.; Fan, T. M.; Fiehn, O.; Goodacre, R.; Griffin, J.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A.; Lindon, J.; Marriott, P.; Nicholls, A.; Reily, M.; Thaden, J.; Viant, M., Proposed minimum reporting standards for chemical analysis. *Metabolomics* **2007,** *3* (3), 211-221; (b) da Silva, R. R.; Wang, M.; Nothias, L.-F.; van der Hooft, J. J. J.; Caraballo-Rodríguez, A. M.; Fox, E.; Balunas, M. J.; Klassen, J. L.; Lopes, N. P.; Dorrestein, P. C., Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **2018,** *14* (4), e1006089.
85. Draper, J.; Lloyd, A.; Goodacre, R.; Beckmann, M., Flow infusion electrospray ionisation mass spectrometry for high throughput, non-targeted metabolite fingerprinting: a review. *Metabolomics* **2012,** *9* (S1), 4-29.
86. Black, C.; Chevallier, O. P.; Elliott, C. T., The current and potential applications of Ambient Mass Spectrometry in detecting food fraud. *Trends Anal. Chem.* **2016,** *82*, 268-278.
87. Ho, Y. N.; Shu, L. J.; Yang, Y. L., Imaging mass spectrometry for metabolites: technical progress, multimodal imaging, and biological interactions. *WIREs Syst. Biol. Med.* **2017,** *9* (5), e1387.
88. Faccin, H.; Viana, C.; do Nascimento, P. C.; Bohrer, D.; de Carvalho, L. M., Study of ion suppression for phenolic compounds in medicinal plant extracts using liquid chromatography-electrospray tandem mass spectrometry. *J. Chromatogr. A* **2016,** *1427*, 111-24.
89. Ganzera, M.; Sturm, S., Recent advances on HPLC/MS in medicinal plant analysis-An update covering 2011-2016. *J. Pharmaceut. Biomed. Anal.* **2018,** *147*, 211-233.
90. Bicchi, C.; Cagliero, C.; Rubiolo, P., New trends in the analysis of the volatile fraction of matrices of vegetable origin: a short overview. A review. *Flavour Frag. J.* **2011,** *26* (5), 321-325.
91. Fiehn, O., Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr. Protoc. Mol. Biol.* **2016,** *114*, 30.4.1-30.4.32.
92. Tubaon, R. M. S.; Rabanes, H.; Haddad, P. R.; Quirino, J. P., Capillary electrophoresis of natural products: 2011–2012. *Electrophoresis* **2014,** *35* (1), 190-204.
93. (a) Maier, T. V.; Schmitt-Kopplin, P., Capillary Electrophoresis in Metabolomics. *Meth. Mol. Biol.* **2016,** *1483*, 437-70; (b) Lisec, J.; Schauer, N.; Kopka, J.; Willmitzer, L.; Fernie, A. R., Gas chromatography mass spectrometry–based metabolite profiling in plants. *Nat. Protoc.* **2006,** *1* (1), 387 - 396; (c) Shuman, J.; Cortes, D.; Armenta, J.; Pokrzywa, R.; Mendes, P.; Shulaev, V., Plant Metabolomics by GC-MS and Differential Analysis. In *Plant Reverse Genetics*, Pereira, A., Ed. Humana Press: 2011; Vol. 678, pp 229-246; (d) Cacciola, F.; Farnetti, S.; Dugo, P.; Marriott, P. J.; Mondello, L., Comprehensive two-dimensional liquid chromatography for polyphenol analysis in foodstuffs. *J. Sep. Sci.* **2017,** *40* (1), 7-24.
94. Fekete, S.; Schappler, J.; Veuthey, J.-L.; Guillarme, D., Current and future trends in UHPLC. *Trends Anal. Chem.* **2014,** *63*, 2-13.

95. (a) Guillarme, D.; Grata, E.; Glauser, G.; Wolfender, J.-L.; Veuthey, J.-L.; Rudaz, S., Some solutions to obtain very efficient separations in isocratic and gradient modes using small particles size and ultra-high pressure. *J. Chromatogr. A* **2009,** *1216* (15), 3232-3243; (b) Grata, E.; Boccard, J.; Guillarme, D.; Glauser, G.; Carrupt, P.-A.; Farmer, E. E.; Wolfender, J.-L.; Rudaz, S., UPLC–TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *J. Chromatogr. B* **2008,** *871* (2), 261-270.
96. Periat, A.; Guillarme, D.; Veuthey, J.-L.; Boccard, J.; Moco, S.; Barron, D.; Grand-Guillaume Perrenoud, A., Optimized selection of liquid chromatography conditions for wide range analysis of natural compounds. *J. Chromatogr. A* **2017,** *1504*, 91-104.
97. Shulaev, V.; Isaac, G., Supercritical fluid chromatography coupled to mass spectrometry - A metabolomics perspective. *J. Chromatogr. B* **2018,** *1092*, 499-505.
98. Perrenoud, A. G.-G.; Guillarme, D.; Boccard, J.; Veuthey, J.-L.; Barron, D.; Moco, S., Ultra-high Performance Supercritical Fluid Chromatography coupled with quadrupole-time-of-flight mass spectrometry as a performing tool for bioactive analysis. *J. Chromatogr. A* **2016,** *1450*, 101-11.
99. Bonaccorsi, I.; Cacciola, F.; Utczas, M.; Inferrera, V.; Giuffrida, D.; Donato, P.; Dugo, P.; Mondello, L., Characterization of the pigment fraction in sweet bell peppers (Capsicum annuum L.) harvested at green and overripe yellow and red stages by offline multidimensional convergence chromatography/liquid chromatography-mass spectrometry. *J. Sep. Sci.* **2016,** *39* (17), 3281-3291.
100. Tranchida, P. Q.; Aloisi, I.; Giocastro, B.; Mondello, L., Current state of comprehensive two-dimensional gas chromatography-mass spectrometry with focus on processes of ionization. *Trends Anal. Chem.* **2018,** *105*, 360-366.
101. Pirok, B. W. J.; Schoenmakers, P. J., Practical Approaches to Overcome the Challenges of Comprehensive Two-Dimensional Liquid Chromatography. *Lc Gc Europe* **2018,** *31* (5), 242-249.
102. Lee, J. W., Basics of Ion Mobility Mass Spectrometry. *Mass Spectrom. Lett.* **2017,** *8* (4), 79-89.
103. Keelor, J. D.; Zambrzycki, S.; Li, A.; Clowers, B. H.; Fernandez, F. M., Atmospheric Pressure Drift Tube Ion Mobility-Orbitrap Mass Spectrometry: Initial Performance Characterization. *Anal. Chem.* **2017,** *89* (21), 11301-11309.
104. Wang, Z.; Kang, D.; Jia, X.; Zhang, H.; Guo, J.; Liu, C.; Meng, Q.; Liu, W., Analysis of alkaloids from Peganum harmala L. sequential extracts by liquid chromatography coupled to ion mobility spectrometry. *J. Chromatogr. B* **2018,** *1096*, 73-79.
105. Chalet, C.; Hollebrands, B.; Janssen, H.-G.; Augustijns, P.; Duchateau, G., Identification of phase-II metabolites of flavonoids by liquid chromatography-ion-mobility spectrometry-mass spectrometry. *Anal. Bioanal. Chem.* **2018,** *410* (2), 471-482.
106. Benton, H. P.; Ivanisevic, J.; Mahieu, N. G.; Kurczy, M. E.; Johnson, C. H.; Franco, L.; Rinehart, D.; Valentine, E.; Gowda, H.; Ubhi, B. K.; Tautenhahn, R.; Gieschen, A.; Fields, M. W.; Patti, G. J.; Siuzdak, G., Autonomous Metabolomics for Rapid Metabolite Identification in Global Profiling. *Anal. Chem.* **2015,** *87* (2), 884-891.
107. De Vijlder, T.; Valkenborg, D.; Lemiere, F.; Romijn, E. P.; Laukens, K.; Cuyckens, F., A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrom. Rev.* **2018,** *37* (5), 607-629.
108. Henke, M. T.; Kelleher, N. L., Modern mass spectrometry for synthetic biology and structure-based discovery of natural products. *Nat. Prod. Rep.* **2016,** *33* (8), 942-950.
109. Kind, T.; Fiehn, O., Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinf.* **2006,** *7* (1), 234.
110. Kind, T.; Fiehn, O., Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinf.* **2007,** *8*, 105.
111. Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A., SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009,** *25* (2), 218-224.
112. Broeckling, C. D.; Hoyes, E.; Richardson, K.; Brown, J. M.; Prenni, J. E., Comprehensive Tandem-Mass-Spectrometry Coverage of Complex Samples Enabled by Data-Set-Dependent Acquisition. *Anal. Chem.* **2018,** *90* (13), 8020-8027.
113. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.;

Showalter, M. R.; Arita, M.; Fiehn, O., Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **2018,** *37* (4), 513-532.
114. Olivon, F.; Allard, P. M.; Koval, A.; Righi, D.; Genta-Jouve, G.; Neyts, J.; Apel, C.; Pannecouque, C.; Nothias, L. F.; Cachet, X.; Marcourt, L.; Roussi, F.; Katanaev, V. L.; Touboul, D.; Wolfender, J. L.; Litaudon, M., Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. *ACS Chem. Biol.* **2017,** *12* (10), 2644-2651.
115. Quinn, R. A.; Nothias, L.-F.; Vining, O.; Meehan, M.; Esquenazi, E.; Dorrestein, P. C., Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. *Trends Pharmacol. Sci.* **2017,** *38* (2), 143-154.
116. Egertson, J. D.; Kuehn, A.; Merrihew, G. E.; Bateman, N. W.; MacLean, B. X.; Ting, Y. S.; Canterbury, J. D.; Marsh, D. M.; Kellmann, M.; Zabrouskov, V.; Wu, C. C.; MacCoss, M. J., Multiplexed MS/MS for improved data-independent acquisition. *Nat. Meth.* **2013,** *10* (8), 744-746.
117. Chapman, J. D.; Goodlett, D. R.; Masselon, C. D., Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrom. Rev.* **2014,** *33* (6), 452-470.
118. Kirk, J.; Sanig, R.; Mortishire-Smith, R.; Naughton, S.; Alelyunas, Y.; Wrona, M., Using ion mobility lLC-MS and LC-MS/MS to resolve and identify isobaric glucuronide metabolites. *Drug Metab. Pharmacokinet.* **2018,** *33* (1), S77-S78.
119. Donoue-Kubo, M.; Allard, P. M.; Wolfender, J.-L., Establishment of a quality control mixture for benchmarking LC-MS based dereplication protocols in natural product research. *Planta Med. Int. Open* **2017,** *4* (S1), DOI: 10.1055/s-0037-1608332.
120. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M., MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Meth.* **2015,** *12* (6), 523-6.
121. Allen, F.; Pon, A.; Wilson, M.; Greiner, R.; Wishart, D., CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.* **2014,** *42* (Web Server issue), W94-9.
122. Vidova, V.; Spacil, Z., A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal. Chim. Acta* **2017,** *964*, 7-23.
123. Le, P. M.; McCooeye, M.; Windust, A., Application of UPLC-QTOF-MS in MSE mode for the rapid and precise identification of alkaloids in goldenseal (Hydrastis canadensis). *Anal. Bioanal. Chem.* **2014,** *406* (6), 1739-1749.
124. Viacava, G. E.; Roura, S. I.; Berrueta, L. A.; Iriondo, C.; Gallo, B.; Alonso-Salces, R. M., Characterization of phenolic compounds in green and red oak-leaf lettuce cultivars by UHPLC-DAD-ESI-QToF/MS using MSE scan mode. *J. Mass Spectrom.* **2017,** *52* (12), 873-902.
125. Vazquez, P. P.; Lozano, A.; Ferrer, C.; Bueno, M. J. M.; Fernandez-Alba, A. R., Improvements in identification and quantitation of pesticide residues in food by LC-QTOF using sequential mass window acquisition (SWATH (R)). *Analytical Methods* **2018,** *10* (24), 2821-2833.
126. Bonner, R.; Hopfgartner, G., SWATH acquisition mode for drug metabolism and metabolomics investigations. *Bioanalysis* **2016,** *8* (16), 1735-1750.
127. Bouslimani, A.; Sanchez, L. M.; Garg, N.; Dorrestein, P. C., Mass spectrometry of natural products: current, emerging and future technologies. *Nat. Prod. Rep.* **2014,** *31* (6), 718-729.
128. Yang, X. Y.; Neta, P.; Stein, S. E., Quality Control for Building Libraries from Electrospray Ionization Tandem Mass Spectra. *Anal. Chem.* **2014,** *86* (13), 6393-6400.
129. Brendle, K.; Kordel, M.; Schneider, E.; Wagner, D.; Braese, S.; Weis, P.; Kappes, M. M., Collision Induced Dissociation of Benzylpyridinium-Substituted Porphyrins: Towards a Thermometer Scale for Multiply Charged Ions? *J. Am. Soc. Mass Spectrom.* **2018,** *29* (2), 382-392.
130. Revesz, A.; Rokob, T. A.; Fouque, D. J. D.; Turiak, L.; Membouf, A.; Vekey, K.; Drahos, L., Selection of Collision Energies in Proteomics Mass Spectrometry Experiments for Best Peptide Identification: Study of Mascot Score Energy Dependence Reveals Double Optimum. *J. Proteome Res.* **2018,** *17* (5), 1898-1906.
131. Waridel, P.; Wolfender, J. L.; Ndjoko, K.; Hobby, K. R.; Major, H. J.; Hostettmann, K., Evaluation of quadrupole time-of-flight tandem mass spectrometry and ion-trap multiple-stage mass

spectrometry for the differentiation of C-glycosidic flavonoid isomers. *J. Chromatogr. A* **2001,** *926* (1), 29-41.
132. van der Hooft, J. J. J.; Padmanabhan, S.; Burgess, K. E. V.; Barrett, M. P., Urinary antihypertensive drug metabolite screening using molecular networking coupled to high-resolution mass spectrometry fragmentation. *Metabolomics* **2016,** *12* (7), 125.
133. Chen, G. B.; Walmsley, S.; Cheung, G. C. M.; Chen, L. Y.; Cheng, C. Y.; Beuerman, R. W.; Wong, T. Y.; Zhou, L.; Choi, H. W., Customized Consensus Spectral Library Building for Untargeted Quantitative Metabolomics Analysis with Data Independent Acquisition Mass Spectrometry and MetaboDIA Workflow. *Anal. Chem.* **2017,** *89* (9), 4897-4906.
134. Henriksen, T.; Juhler, R. K.; Svensmark, B.; Cech, N. B., The relative influences of acidity and polarity on responsiveness of small organic molecules to analysis with negative ion electrospray ionization mass spectrometry (ESI-MS). *J. Am. Soc. Mass Spectrom.* **2005,** *16* (4), 446-455.
135. Kiontke, A.; Oliveira-Birkmeier, A.; Opitz, A.; Birkemeyer, C., Electrospray Ionization Efficiency Is Dependent on Different Molecular Descriptors with Respect to Solvent pH and Instrumental Configuration. *PloS ONE* **2016,** *11* (12), e0167502.
136. Wolfender, J.-L., HPLC in natural product analysis: the detection issue. *Planta Med.* **2009,** *75*, 719-734.
137. Ligor, M.; Studzinska, S.; Horna, A.; Buszewski, B., Corona-Charged Aerosol Detection: An Analytical Approach. *Crit. Rev. Anal. Chem.* **2013,** *43* (2), 64-78.
138. Vervoort, N.; Daemen, D.; Torok, G., Performance evaluation of evaporative light scattering detection and charged aerosol detection in reversed phase liquid chromatography. *J. Chromatogr. A* **2008,** *1189* (1-2), 92-100.
139. (a) Yang, J.; Liang, Q.; Wang, M.; Jeffries, C.; Smithson, D.; Tu, Y.; Boulos, N.; Jacob, M. R.; Shelat, A. A.; Wu, Y.; Ravu, R. R.; Gilbertson, R.; Avery, M. A.; Khan, I. A.; Walker, L. A.; Guy, R. K.; Li, X.-C., UPLC-MS-ELSD-PDA as a Powerful Dereplication Tool to Facilitate Compound Identification from Small-Molecule Natural Product Libraries. *J. Nat. Prod.* **2014,** *77* (4), 902-9; (b) Granica, S., Quantitative and qualitative investigations of pharmacopoeial plant material polygoni avicularis herba by UHPLC-CAD and UHPLC-ESI-MS methods. *Phytochem. Anal.* **2015,** *26* (5), 374-382.
140. da Rocha, C. Q.; de-Faria, F. M.; Marcourt, L.; Ebrahimi, S. N.; Kitano, B. T.; Ghilardi, A. F.; Luiz Ferreira, A.; de Almeida, A. C. A.; Dunder, R. J.; Souza-Brito, A. R. M.; Hamburger, M.; Vilegas, W.; Queiroz, E. F.; Wolfender, J.-L., Gastroprotective effects of hydroethanolic root extract of Arrabidaea brachypoda: Evidences of cytoprotection and isolation of unusual glycosylated polyphenols. *Phytochemistry* **2017,** *135*, 93-105.
141. Leitner, A.; Emmert, J.; Boerner, K.; Lindner, W., Influence of solvent additive composition on chromatographic separation and sodium adduct formation of peptides in HPLC-ESI MS. *Chromatographia* **2007,** *65* (11-12), 649-653.
142. (a) van der Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; Beekwilder, J.; de Vos, R. C. H., Polyphenol Identification Based on Systematic and Robust High-Resolution Accurate Mass Spectrometry Fragmentation. *Anal. Chem.* **2011,** *83* (1), 409-416; (b) van der Hooft, J. J. J.; Vervoort, J.; Bino, R. J.; de Vos, R. C. H., Spectral trees as a robust annotation tool in LC-MS based metabolomics. *Metabolomics* **2011,** *8* (4), 691-703; (c) Vaniya, A.; Fiehn, O., Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Anal. Chem.* **2015,** *69*, 52-61.
143. Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S., Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U.S.A.* **2015,** *112* (41), 12580-5.
144. Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; Bino, R. J.; Vervoort, J., Automatic Chemical Structure Annotation of an LC-MSn Based Metabolic Profile from Green Tea. *Anal. Chem.* **2013,** *85* (12), 6033-6040.
145. Spicer, R.; Salek, R. M.; Moreno, P.; Cañueto, D.; Steinbeck, C., Navigating freely-available software tools for metabolomics analysis. *Metabolomics* **2017,** *13* (9), 106.
146. Wohlgemuth, G.; Mehta, S. S.; Mejia, R. F.; Neumann, S.; Pedrosa, D.; Pluskal, T.; Schymanski, E. L.; Willighagen, E. L.; Wilson, M.; Wishart, D. S.; Arita, M.; Dorrestein, P. C.; Bandeira, N.; Wang, M.;

Schulze, T.; Salek, R. M.; Steinbeck, C.; Nainala, V. C.; Mistrik, R.; Nishioka, T.; Fiehn, O., SPLASH, a hashed identifier for mass spectra. *Nat. Biotechnol.* **2016,** *34* (11), 1099-1101.
147. Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; Montecchi-Palazzi, L.; Tasman, N.; Coleman, M.; Reisinger, F.; Souda, P.; Hermjakob, H.; Binz, P.-A.; Deutsch, E. W., mzML—a Community Standard for Mass Spectrometry Data. *Mol. Cell. Proteomics* **2011,** *10* (1), R110 000133.
148. Deutsch, E. W., File Formats Commonly Used in Mass Spectrometry Proteomics. *Mol. Cell. Proteomics* **2012,** *11* (12), 1612-1621.
149. Bertrand, S.; Guitton, Y.; Roullier, C., Successes and pitfalls in automated dereplication strategy using liquid chromatography coupled to mass spectrometry data: A CASMI 2016 experience. *Phytochem. Lett.* **2017,** *21*, 297-305.
150. Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; Hoff, K.; Kessner, D.; Tasman, N.; Shulman, N.; Frewen, B.; Baker, T. A.; Brusniak, M.-Y.; Paulse, C.; Creasy, D.; Flashner, L.; Kani, K.; Moulding, C.; Seymour, S. L.; Nuwaysir, L. M.; Lefebvre, B.; Kuhlmann, F.; Roark, J.; Rainer, P.; Detlev, S.; Hemenway, T.; Huhmer, A.; Langridge, J.; Connolly, B.; Chadick, T.; Holly, K.; Eckels, J.; Deutsch, E. W.; Moritz, R. L.; Katz, J. E.; Agus, D. B.; MacCoss, M.; Tabb, D. L.; Mallick, P., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012,** *30* (10), 918-920.
151. Lopez-Lopez, A.; Lopez-Gonzalvez, A.; Barker-Tejeda, T. C.; Barbas, C., A review of validated biomarkers obtained through metabolomics. *Expert Rev. Mol. Diag.* **2018,** *18* (6), 557-575.
152. Shahaf, N.; Rogachev, I.; Heinig, U.; Meir, S.; Malitsky, S.; Battat, M.; Wyner, H.; Zheng, S.; Wehrens, R.; Aharoni, A., The WEIZMASS spectral library for high-confidence metabolite identification. *Nat. Comm.* **2016,** *7*, 12423.
153. Wang, M.; Bandeira, N., Spectral library generating function for assessing spectrum-spectrum match significance. *J. Proteome Res.* **2013,** *12* (9), 3944-51.
154. Allen, F.; Greiner, R.; Wishart, D., Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* **2014,** *11* (1), 98-110.
155. Schymanski, E. L.; Singer, H. P.; Slobodnik, J.; Ipolyi, I. M.; Oswald, P.; Krauss, M.; Schulze, T.; Haglund, P.; Letzel, T.; Grosse, S.; Thomaidis, N. S.; Bletsou, A.; Zwiener, C.; Ibanez, M.; Portoles, T.; de Boer, R.; Reid, M. J.; Onghena, M.; Kunkel, U.; Schulz, W.; Guillon, A.; Noyon, N.; Leroy, G.; Bados, P.; Bogialli, S.; Stipanicev, D.; Rostkowski, P.; Hollender, J., Non-target screening with high-resolution mass spectrometry: critical review using a collaborative trial on water analysis. *Anal. Bioanal. Chem.* **2015,** *407* (21), 6237-55.
156. Cabral, R. S. A.; Allard, P.-M.; Marcourt, L.; Young, M. C. M.; Queiroz, E. F.; Wolfender, J.-L., Targeted Isolation of Indolopyridoquinazoline Alkaloids from Conchocarpus fontanesianus Based on Molecular Networks. *J. Nat. Prod.* **2016,** *79* (9), 2270-8.
157. Ma, Y.; Kind, T.; Yang, D.; Leon, C.; Fiehn, O., MS2Analyzer - a software for small molecule substructure annotations from accurate mass MS/MS spectra. *Anal. Chem.* **2014,** *86* (21), 10724-31.
158. Nielsen, K. F.; Månsson, M.; Rank, C.; Frisvad, J. C.; Larsen, T. O., Dereplication of microbial natural products by LC-DAD-TOFMS. *J. Nat. Prod.* **2011,** *74* (11), 2338-2348.
159. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S., CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **2012,** *84* (1), 283-289.
160. Mahieu, N. G.; Spalding, J. L.; Gelman, S. J.; Patti, G. J., Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. *Anal. Chem.* **2016,** *88* (18), 9037-9046.
161. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf.* **2010,** *11* (1), 395.
162. (a) Scheubert, K.; Hufsky, F.; Böcker, S., Computational mass spectrometry for small molecules. *J. Cheminformatics* **2013,** *5* (1), 12; (b) Kind, T.; Fiehn, O., Advances in structure elucidation of small molecules using mass spectrometry. *Bioanal. Rev.* **2010,** *2* (1-4), 23-60; (c) Krug, D.; Muller, R., Secondary metabolomics: the impact of mass spectrometry-based approaches on the discovery and

characterization of microbial natural products. *Nat. Prod. Rep.* **2014,** *31* (6), 768-83.
163. Konishi, Y.; Kiyota, T.; Draghici, C.; Gao, J.-M.; Yeboah, F.; Acoca, S.; Jarussophon, S.; Purisima, E., Molecular Formula Analysis by an MS/MS/MS Technique To Expedite Dereplication of Natural Products. *Anal. Chem.* **2006,** *79* (3), 1187-1197.
164. Dührkop, K.; Böcker, S., Fragmentation Trees Reloaded. In *Research in Computational Molecular Biology: 19th Annual International Conference, RECOMB 2015, Warsaw, Poland, April 12-15, 2015, Proceedings*, Przytycka, M. T., Ed. Springer International Publishing: Cham, 2015; pp 65-79.
165. Vanderplanck, M.; Glauser, G., Integration of non-targeted metabolomics and automated determination of elemental compositions for comprehensive alkaloid profiling in plants. *Phytochemistry* **2018,** *154*, DOI: 10.1016/j.phytochem.2018.06.011.
166. Meusel, M.; Hufsky, F.; Panter, F.; Krug, D.; Müller, R.; Böcker, S., Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal. Chem.* **2016,** *88* (15), 7556-66.
167. Roullier, C.; Guitton, Y.; Valery, M.; Amand, S.; Prado, S.; Robiou du Pont, T.; Grovel, O.; Pouchus, Y. F., Automated detection of natural halogenated compounds from LC-MS profiles – Application to the isolation of bioactive chlorinated compounds from marine-derived fungi. *Anal. Chem.* **2016,** *88* (18), 9143-50.
168. (a) Bertrand, S.; Azzollini, A.; Schumpp, O.; Bohni, N.; Schrenzel, J.; Monod, M.; Gindro, K.; Wolfender, J.-L., Multi-well fungal co-culture for *de novo* metabolite-induction in time series studies based on untargeted metabolomics. *Mol. BioSyst.* **2014,** *10* (9), 2289-2298; (b) Funari, C. S.; Eugster, P. J.; Martel, S.; Carrupt, P.-A.; Wolfender, J.-L.; Silva, D. H. S., High resolution ultra high pressure liquid chromatography–time-of-flight mass spectrometry dereplication strategy for the metabolite profiling of Brazilian Lippia species. *J. Chromatogr. A* **2012,** *1259*, 167-78.
169. Neumann, S.; Nikolic, D.; Schymanski, E.; Shahaf, N., Critical Assessment of Small Molecule Identification: Looking at the 5th Edition of CASMI. *MetaboNews* **2018,** *8* (3), 5-8.
170. Fuzzati, N.; Gabetta, B.; Strepponi, I.; Villa, F., High-performance liquid chromatography–electrospray ionization mass spectrometry and multiple mass spectrometry studies of hyperforin degradation products. *J. Chromatogr. A* **2001,** *926* (1), 187-198.
171. (a) Hufsky, F.; Böcker, S., Mining molecular structure databases: Identification of small molecules based on fragmentation mass spectrometry data. *Mass Spectrom. Rev.* **2017,** *36* (5), 624-633; (b) Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O., Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018,** *8* (2), 31.
172. Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S., Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom.* **2014,** *3* (Special_Issue_2), S0033-S0033.
173. Ruttkies, C.; Schymanski, E.; Wolf, S.; Hollender, J.; Neumann, S., MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminformatics* **2016,** *8* (1), 3.
174. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; Kind, T.; Beal, P.; Arita, M.; Fiehn, O., Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Meth.* **2017,** *15* (1), 53-56.
175. Laponogov, I.; Sadawi, N.; Galea, D.; Mirnezami, R.; Veselkov, K. A., ChemDistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics* **2018,** *34* (12), 2096-2102.
176. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G., Retrospective analysis of natural products provides insights for future discovery trends. *Proc. Natl. Acad. Sci. U. S. A.* **2017,** *114* (22), 5601-5606.
177. Kanehisa, M.; Goto, S., KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000,** *28* (1), 27-30, URL: http://www.genome.jp/kegg/.
178. Jeffryes, J. G.; Colastani, R. L.; Elbadawi-Sidhu, M.; Kind, T.; Niehaus, T. D.; Broadbelt, L. J.; Hanson, A. D.; Fiehn, O.; Tyo, K. E.; Henry, C. S., MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J. Cheminformatics* **2015,** *7*, 44.
179. Duigou, T.; du Lac, M.; Carbonell, P.; Faulon, J.-L., RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Res.* **2018**, gky940, doi: 10.1093/nar/gky940.

180. Beauxis, Y.; Genta-Jouve, G., Metwork: a web server for natural products anticipation. *Bioinformatics* **2018**, bty864, DOI: 10.1093/bioinformatics/bty864.
181. Kind, T.; Liu, K.-H.; Lee, D. Y.; DeFelice, B.; Meissen, J. K.; Fiehn, O., LipidBlast *in silico* tandem mass spectrometry database for lipid identification. *Nat. Meth.* **2013,** *10* (8), 755-758.
182. Herzog, R.; Schuhmann, K.; Schwudke, D.; Sampaio, J. L.; Bornstein, S. R.; Schroeder, M.; Shevchenko, A., LipidXplorer: A Software for Consensual Cross-Platform Lipidomics. *PLoS ONE* **2012,** *7* (1), e29851.
183. (a) Demuth, W.; Karlovits, M.; Varmuza, K., Spectral similarity versus structural similarity: mass spectrometry. *Anal. Chim. Acta* **2004,** *516* (1), 75-85; (b) Schollee, J. E.; Schymanski, E. L.; Stravs, M. A.; Gulde, R.; Thomaidis, N. S.; Hollender, J., Similarity of High-Resolution Tandem Mass Spectrometry Spectra of Structurally Related Micropollutants and Transformation Products. *J. Am. Soc. Mass Spectrom.* **2017,** *28* (12), 2692-2704.
184. (a) Borges Ricardo, M.; Taujale, R.; Souza Juliana, S.; Andrade Bezerra, T.; Silva Eder Lana, e.; Herzog, R.; Ponce Francesca, V.; Wolfender, J. L.; Edison Arthur, S., Dereplication of plant phenolics using a mass-spectrometry database independent method. *Phytochem. Anal.* **2018,** *29* (6), 601-612; (b) Soares, V.; Taujale, R.; Garrett, R.; Silva, A. J. R.; Borges, R. M., Extending compound identification for molecular network using the LipidXplorer database independent method: A proof of concept using glycoalkaloids from Solanum pseudoquina A. St.-Hil. *Phytochem. Anal.* **2018**, DOI: 10.1002/pca.2798.
185. (a) Klitgaard, A.; Iversen, A.; Andersen, M. R.; Larsen, T. O.; Frisvad, J. C.; Nielsen, K. F., Aggressive dereplication using UHPLC–DAD–QTOF: screening extracts for up to 3000 fungal secondary metabolites. *Anal. Bioanal. Chem.* **2014,** *406* (7), 1933-43; (b) Kildgaard, S.; Mansson, M.; Dosen, I.; Klitgaard, A.; Frisvad, J.; Larsen, T.; Nielsen, K., Accurate Dereplication of Bioactive Secondary Metabolites from Marine-Derived Fungi by UHPLC-DAD-QTOFMS and a MS/HRMS Library. *Mar. Drugs* **2014,** *12* (6), 3681-3705.
186. (a) Cuyckens, F.; Claeys, M., Mass spectrometry in the structural analysis of flavonoids. *J. Mass Spectrom.* **2004,** *39* (1), 1-15; (b) Kerzaon, I.; Pouchus, Y. F.; Monteau, F.; Le Bizec, B.; Nourrisson, M.-R.; Biard, J.-F.; Grovel, O., Structural investigation and elucidation of new communesins from a marine-derived Penicillium expansum Link by liquid chromatography/electrospray ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **2009,** *23* (24), 3928-3938.
187. (a) Watrous, J.; Roach, P.; Alexandrov, T.; Heath, B. S.; Yang, J. Y.; Kersten, R. D.; van der Voort, M.; Pogliano, K.; Gross, H.; Raaijmakers, J. M.; Moore, B. S.; Laskin, J.; Bandeira, N.; Dorrestein, P. C., Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **2012,** *109* (26), E1743–E1752; (b) Naake, T.; Gaquerel, E., MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics* **2017,** *33* (15), 2419-2420.
188. (a) Depke, T.; Franke, R.; Brönstrup, M., Clustering of MS2 spectra using unsupervised methods to aid the identification of secondary metabolites from Pseudomonas aeruginosa. *J. Chromatogr. B* **2017,** *1071*, 19-28; (b) Dayringer, H. E.; McLafferty, F. W.; Venkataraghavan, R., Computer-aided interpretation of mass spectra. Increased information from neutral loss data. *Organic Mass Spectrometry* **1976,** *11* (8), 895-900.
189. (a) Treutler, H.; Tsugawa, H.; Porzel, A.; Gorzolka, K.; Tissier, A.; Neumann, S.; Balcke, G. U., Discovering Regulated Metabolite Families in Untargeted Metabolomics Studies. *Anal. Chem.* **2016,** *88* (16), 8082-90; (b) van der Hooft, J. J. J.; Wandy, J.; Barrett, M. P.; Burgess, K. E. V.; Rogers, S., Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci. U.S.A.* **2016,** *113* (48), 13738-13743.
190. Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; de Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C., Molecular Networking as a Dereplication Strategy. *J. Nat. Prod.* **2013,** *76* (9), 1686-99.
191. Olivon, F.; Elie, N.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D., MetGem software for the generation of molecular networks based on t-SNE algorithm. *Anal. Chem.* **2018**, DOI: 10.1021/acs.analchem.8b03099.
192. (a) Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada,

K.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of peptidic natural products through database search of mass spectra. *Nat. Chem. Biol.* **2017,** *13* (1), 30-37; (b) Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P. A., Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat. Microbiol.* **2018,** *3* (3), 319-327; (c) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of microbial metabolites through database search of mass spectra. *Nat. Com.* **2018,** *9* (1), 4035.

193. (a) Longnecker, K.; Kujawinski, E. B., Mining mass spectrometry data: Using new computational tools to find novel organic compounds in complex environmental mixtures. *Org. Geochem.* **2017,** *110*, 92-99; (b) Naman, C. B.; Rattan, R.; Nikoulina, S. E.; Lee, J.; Miller, B. W.; Moss, N. A.; Armstrong, L.; Boudreau, P. D.; Debonsi, H. M.; Valeriote, F. A.; Dorrestein, P. C.; Gerwick, W. H., Integrating Molecular Networking and Biological Assays To Target the Isolation of a Cytotoxic Cyclic Octapeptide, Samoamide A, from an American Samoan Marine Cyanobacterium. *J. Nat. Prod.* **2017,** *80* (3), 625-633; (c) Crüsemann, M.; O'Neill, E. C.; Larson, C. B.; Melnik, A. V.; Floros, D. J.; da Silva, R. R.; Jensen, P. R.; Dorrestein, P. C.; Moore, B. S., Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. *J. Nat. Prod.* **2017,** *80* (3), 588-597; (d) Luzzatto-Knaan, T.; Garg, N.; Wang, M.; Glukhov, E.; Peng, Y.; Ackermann, G.; Amir, A.; Duggan, B. M.; Ryazanov, S.; Gerwick, L.; Knight, R.; Alexandrov, T.; Bandeira, N.; Gerwick, W. H.; Dorrestein, P. C., Digitizing mass spectrometry data to explore the chemical diversity and distribution of marine cyanobacteria and algae. *eLife* **2017,** *6*, e24214; (e) Nguyen, D. D.; Melnik, A. V.; Koyama, N.; Lu, X.; Schorn, M.; Fang, J.; Aguinaldo, K.; Lincecum, T. L., Jr.; Ghequire, M. G.; Carrion, V. J.; Cheng, T. L.; Duggan, B. M.; Malone, J. G.; Mauchline, T. H.; Sanchez, L. M.; Kilpatrick, A. M.; Raaijmakers, J. M.; De Mot, R.; Moore, B. S.; Medema, M. H.; Dorrestein, P. C., Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.* **2016,** *2*, 16197.

194. Dhurjad, P. S.; Marothu, V. K.; Rathod, R., Post-acquisition data mining techniques for LC-MS/MS-acquired data in drug metabolite identification. *Bioanalysis* **2017,** *9* (16), 1265-1278.

195. Blei, D. M.; Ng, A. Y.; Jordan, M. I., Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003,** *3*, 993-1022.

196. Wandy, J.; Zhu, Y.; van der Hooft, J. J. J.; Daly, R.; Barrett, M. P.; Rogers, S., Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. *Bioinformatics* **2018,** *34* (2), 317-318.

197. van der Hooft, J. J. J.; Wandy, J.; Young, F.; Padmanabhan, S.; Gerasimidis, K.; Burgess, K. E. V.; Barrett, M. P.; Rogers, S., Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Anal. Chem.* **2017,** *89* (14), 7569-7577.

198. Ernst, M.; Nothias-Scaglia, L.-F.; van der Hooft, J.; Silva, R. R.; Saslis-Lagoudakis, C. H.; Grace, O. M.; Martinez-Swatson, K.; Hassemer, G.; Funez, L.; T. Simonsen, H.; Medema, M. H.; Staerk, D.; Nilsson, N.; Lovato, P.; Dorrestein, P.; Ronsted, N., Did a plant-herbivore arms race drive chemical diversity in Euphorbia? *bioRxiv* **2018**, DOI: 10.1101/323014.

199. Tautenhahn, R.; Bottcher, C.; Neumann, S., Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinf.* **2008,** *9*, 504.

200. Gowda, H.; Ivanisevic, J.; Johnson, C. H.; Kurczy, M. E.; Benton, H. P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; Westenskow, P. D.; Wang, J.; Arkin, A. P.; Deutshbauer, A. M.; Patti, G. J.; Siuzdak, G. E., Interactive XCMS Online: Simplifying Advanced Metabolomic Processing and Subsequent Statistical Analyses. *Anal. Chem.* **2014,** *86* (14), 6931-9.

201. Pluskal, T.; Uehara, T.; Yanagida, M., Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal. Chem.* **2012,** *84* (10), 4396-403.

202. Olivon, F.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D., MZmine 2 data-preprocessing to enhance Molecular Networking reliability. *Anal. Chem.* **2017,** *89* (15), 7836-7840.

203. Fox Ramos, A. E.; Alcover, C.; Evanno, L.; Maciuk, A.; Litaudon, M.; Duplais, C.; Bernadat, G.; Gallard, J.-F.; Jullian, J.-C.; Mouray, E.; Grellier, P.; Loiseau, P. M.; Pomel, S.; Poupon, E.; Champy, P.; Beniddir, M. A., Revisiting Previously Investigated Plants: A Molecular Networking-Based Study of

Geissospermum laeve. *J. Nat. Prod.* **2017,** *80* (4), 1007-1014.
204. Misra, B. B., New tools and resources in metabolomics: 2016-2017. *Electrophoresis* **2018,** *39* (7), 909-923.
205. Pontoizeau, C. m.; Mouchiroud, L.; Molin, L.; Mergoud-dit-Lamarche, A.; Dallière, N.; Toulhoat, P.; Elena-Herrmann, B. n. d.; Solari, F., Metabolomics analysis uncovers that dietary restriction buffers metabolic changes associated with aging in Caenorhabditis elegans. *J. Proteome Res.* **2014,** *13* (6), 2910-2919.
206. Munz, E.; Jakob, P. M.; Borisjuk, L., The potential of nuclear magnetic resonance to track lipids in planta. *Biochimie* **2016,** *130*, 97-108.
207. Choi, Y. H.; Verpoorte, R., Metabolomics: What You See is What You Extract. *Phytochem. Anal.* **2014,** *25* (4), 289-290.
208. Kim, H. K.; Choi, Y. H.; Verpoorte, R., NMR-based metabolomic analysis of plants. *Nat. Protoc.* **2010,** *5* (3), 536-549.
209. Breton, R. C.; Reynolds, W. F., Using NMR to identify and characterize natural products. *Nat. Prod. Rep.* **2013,** *30* (4), 501-24.
210. van der Hooft, J. J. J.; Rankin, N., Metabolite Identification in Complex Mixtures Using Nuclear Magnetic Resonance Spectroscopy. In *Modern Magnetic Resonance*, Webb, G. A., Ed. Springer International Publishing: Cham, 2017; pp 1-33.
211. Wolfender, J.-L.; Rudaz, S.; Hae Choi, Y.; Kyong Kim, H., Plant metabolomics: from holistic data to relevant biomarkers. *Curr. Med. Chem.* **2013,** *20* (8), 1056-1090.
212. Schwalbe, H., New 1.2 GHz NMR Spectrometers—New horizons? *Angew. Chem. Int. Ed.* **2017,** *56* (35), 10252-10253.
213. (a) Bornet, A.; Maucourt, M. l.; Deborde, C.; Jacob, D.; Milani, J.; Vuichoud, B.; Ji, X.; Dumez, J.-N.; Moing, A.; Bodenhausen, G., Highly repeatable dissolution dynamic nuclear polarization for heteronuclear NMR metabolomics. *Anal. Chem.* **2016,** *88* (12), 6179-6183; (b) Liu, G.; Levien, M.; Karschin, N.; Parigi, G.; Luchinat, C.; Bennati, M., One-thousand-fold enhancement of high field liquid nuclear magnetic resonance signals at room temperature. *Nat. Chem.* **2017,** *9* (7), 676.
214. Molinski, T. F., NMR of natural products at the 'nanomole-scale'. *Nat. Prod. Rep.* **2010,** *27* (3), 321-329.
215. Aslam, N.; Pfender, M.; Neumann, P.; Reuter, R.; Zappe, A.; Favaro de Oliveira, F.; Denisenko, A.; Sumiya, H.; Onoda, S.; Isoya, J.; Wrachtrup, J., Nanoscale nuclear magnetic resonance with chemical resolution. *Science* **2017,** *357* (6346), 67-71.
216. Dayrit, F. M.; de Dios, A. C., 1H and 13C NMR for the Profiling of Natural Product Extracts: Theory and Applications. In *Spectroscopic Analyses-Developments and Applications*, InTech: 2017.
217. Aue, W.; Karhan, J.; Ernst, R., Homonuclear broad band decoupling and two-dimensional J-resolved NMR spectroscopy. *J. Chem. Phys.* **1976,** *64* (10), 4226-4227.
218. Bax, A.; Freeman, R., Investigation of complex networks of spin-spin coupling by two-dimensional NMR. *J. Magn. Res.* **1981,** *44* (3), 542-561.
219. Zangger, K.; Sterk, H., Homonuclear Broadband-Decoupled NMR Spectra. *J. Magn. Res.* **1997,** *124* (2), 486-489.
220. Foroozandeh, M.; Morris, G.; Nilsson, M., PSYCHE Pure Shift NMR Spectroscopy. *Chem. Eur. J.* **2018,** *24* (53), 13988-14000.
221. Mahrous, E. A.; Farag, M. A., Two dimensional NMR spectroscopic approaches for exploring plant metabolome: A review. *J. Adv. Res.* **2015,** *6* (1), 3-15.
222. (a) Ludwig, C.; Easton, J. M.; Lodi, A.; Tiziani, S.; Manzoor, S. E.; Southam, A. D.; Byrne, J. J.; Bishop, L. M.; He, S.; Arvanitis, T. N., Birmingham Metabolite Library: a publicly accessible database of 1-D 1H and 2-D 1H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics* **2012,** *8* (1), 8-18; (b) BBIOREFCODE 2 - Metabolite Reference Database. https://www.bruker.com/products/mr/nmr/nmr-software/nmr-software/bbiorefcode/overview.html.
223. Johnson Jr, C. S., Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999,** *34* (3-4), 203-256.

224. Cassani, J.; Nilsson, M.; Morris, G. A., Flavonoid Mixture Analysis by Matrix-Assisted Diffusion-Ordered Spectroscopy. *J. Nat. Prod.* **2012,** *75* (2), 131-134.
225. Lucas, L. H.; Otto, W. H.; Larive, C. K., The 2D-J-DOSY experiment: resolving diffusion coefficients in mixtures. *J. Magn. Res.* **2002,** *156* (1), 138-145.
226. Neuhaus, D.; Williamson, M. P., *The Nuclear Overhauser Effect in Structural and Conformational Analysis, 2nd Edition*. 2000.
227. Lameiras, P.; Nuzillard, J.-M., Highly Viscous Binary Solvents: DMSO-d6/Glycerol and DMSO-d6/Glycerol-d8 for Polar and Apolar Mixture Analysis by NMR. *Anal. Chem.* **2016,** *88* (8), 4508-4515.
228. Kiraly, P.; Nilsson, M.; Morris, G. A., Practical aspects of real-time pure shift HSQC experiments. *Magn. Reson. Chem.* **2018,** *56* (10), 993-1005.
229. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H., Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Scientific Rep.* **2017,** *7* (1), 14243.
230. Nilsson, M.; Morris, G. A., Pure shift proton DOSY: diffusion-ordered 1 H spectra without multiplet structure. *Chem. Commun.* **2007,** (9), 933-935.
231. Njock, G. B. B.; Pegnyemb, D. E.; Bartholomeusz, T. A.; Christen, P.; Vitorge, B.; Nuzillard, J.-M.; Shivapurkar, R.; Foroozandeh, M.; Jeannerat, D., Spectral aliasing: A super zoom for 2D-NMR spectra. Principles and applications. *CHIMIA* **2010,** *64* (4), 235-240.
232. Gaillet, C.; Lequart, C.; Debeire, P.; Nuzillard, J.-M., Band-selective HSQC and HMBC experiments using excitation sculpting and PFGSE. *J. Magn. Res.* **1999,** *139* (2), 454-459.
233. Delaglio, F.; Walker, G. S.; Farley, K. A.; Sharma, R.; Hoch, J. C.; Arbogast, L. W.; Brinson, R. G.; Marino, J. P., Non-Uniform Sampling for All: More NMR Spectral Quality, Less Measurement Time. *Am. Pharm. Rev.* **2017,** *20* (4), 339681.
234. Foroozandeh, M.; Adams, R. W.; Nilsson, M.; Morris, G. A., Ultrahigh-resolution total correlation NMR spectroscopy. *J. Am. Chem. Soc.* **2014,** *136* (34), 11867-11869.
235. Farjon, J.; Milande, C. m.; Martineau, E.; Akoka, S.; Giraudeau, P., The FAQUIRE Approach: FAst, QUantitative, hIghly Resolved and sEnsitivity Enhanced 1H, 13C Data. *Anal. Chem.* **2018,** *90* (3), 1845-1851.
236. Giraudeau, P.; Frydman, L., Ultrafast 2D NMR: an emerging tool in analytical spectroscopy. *Annu. Rev. Anal. Chem.* **2014,** *7*, 129-161.
237. Kupče, Ē.; Claridge, T. D., NOAH: NMR supersequences for small molecule analysis and structure elucidation. *Angew. Chem. Int. Ed.* **2017,** *56* (39), 11779-11783.
238. Reynolds, W. F.; Enríquez, R. G., Choosing the Best Pulse Sequences, Acquisition Parameters, Postacquisition Processing Strategies, and Probes for Natural Product Structure Elucidation by NMR Spectroscopy. *J. Nat. Prod.* **2002,** *65* (2), 221-244.
239. Qiu, F.; McAlpine, J. B.; Lankin, D. C.; Burton, I.; Karakach, T.; Chen, S.-N.; Pauli, G. F., 2D NMR barcoding and differential analysis of complex mixtures for chemical identification: the Actaea triterpenes. *Anal. Chem.* **2014,** *86* (8), 3964-3972.
240. Bakiri, A.; Hubert, J.; Reynaud, R.; Lambert, C.; Martinez, A.; Renault, J.-H.; Nuzillard, J.-M., Reconstruction of HMBC Correlation Networks: A Novel NMR-based Contribution to Metabolite Mixture Analysis. *J. Chem. Inf. Model.* **2018,** *58* (2), 262-270.
241. Margueritte, L.; Markov, P.; Chiron, L.; Starck, J. P.; Vonthron-Senecheau, C.; Bourjot, M.; Delsuc, M. A., Automatic differential analysis of NMR experiments in complex samples. *Magn. Reson. Chem.* **2018,** *56* (6), 469-479.
242. Aligiannis, N.; Halabalaki, M.; Chaita, E.; Kouloura, E.; Argyropoulou, A.; Benaki, D.; Kalpoutzakis, E.; Angelis, A.; Stathopoulou, K.; Antoniou, S.; Sani, M.; Krauth, V.; Werz, O.; Schütz, B.; Schäfer, H.; Spraul, M.; Mikros, E.; Skaltsounis, L. A., Heterocovariance Based Metabolomics as a Powerful Tool Accelerating Bioactive Natural Product Identification. *ChemistrySelect* **2016,** *1* (10), 2531-2535.
243. Barton, R. H., A decade of advances in metabonomics. *Expert Opin. Drug Metab. Toxicol.* **2011,** *7* (2), 129-136.
244. Beckonert, O.; Keun, H. C.; Ebbels, T. M. D.; Bundy, J.; Holmes, E.; Lindon, J. C.; Nicholson, J. K.,

Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2007,** *2*, 2692.
245. Weljie, A. M.; Newton, J.; Jirik, F. R.; Vogel, H. J., Evaluating Low-Intensity Unknown Signals in Quantitative Proton NMR Mixture Analysis. *Anal. Chem.* **2008,** *80* (23), 8956-8965.
246. Leiss, K. A.; Choi, Y. H.; Verpoorte, R.; Klinkhamer, P. G. L., An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance. *Phytochem. Rev.* **2011,** *10* (2), 205-216.
247. Pauli, G. F.; Chen, S.-N.; Lankin, D. C.; Bisson, J.; Case, R. J.; Chadwick, L. R.; Gödecke, T.; Inui, T.; Krunic, A.; Jaki, B. U.; McAlpine, J. B.; Mo, S.; Napolitano, J. G.; Orjala, J.; Lehtivarjo, J.; Korhonen, S.-P.; Niemitz, M., Essential Parameters for Structural Analysis and Dereplication by 1H NMR Spectroscopy. *J. Nat. Prod.* **2014,** *77* (6), 1473-1487.
248. Rezzi, S.; Bighelli, A.; Castola, V.; Casanova, J., Direct identification and quantitative determination of acidic and neutral diterpenes using 13C-NMR spectroscopy. Application to the analysis of oleoresin of Pinus nigra. *Appl. Spectrosc.* **2002,** *56* (3), 312-317.
249. Laude, D. A.; Wilkins, C. L., Identification of organic mixture components without separation: Quantitative and edited carbon-13 nuclear magnetic resonance spectrometry data for analysis of petroleum distillates. *Anal. Chem.* **1986,** *58* (13), 2820-2824.
250. Ferreira, M. J.; Costantin, M. B.; Sartorelli, P. c.; Rodrigues, G. V.; Limberger, R.; Henriques, A. T.; Kato, M. J.; Emerenciano, V. P., Computer-aided method for identification of components in essential oils by 13C NMR spectroscopy. *Anal. Chim. Acta* **2001,** *447* (1-2), 125-134.
251. Hubert, J.; Nuzillard, J.-M.; Purson, S.; Hamzaoui, M.; Borie, N.; Reynaud, R.; Renault, J.-H., Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on 13C NMR. *Anal. Chem.* **2014,** *86* (6), 2955-2962.
252. Nargund, S.; Joffe, M. E.; Tran, D.; Tugarinov, V.; Sriram, G., Nuclear magnetic resonance methods for metabolic fluxomics. In *Systems Metabolic Engineering*, Springer: 2013; pp 335-351.
253. Millard, P.; Cahoreau, E.; Heuillet, M.; Portais, J.-C.; Lippens, G., 15N-NMR-Based Approach for Amino Acids-Based 13C-Metabolic Flux Analysis of Metabolism. *Anal. Chem.* **2017,** *89* (3), 2101-2106.
254. Bingol, K.; Zhang, F.; Bruschweiler-Li, L.; Brüschweiler, R., TOCCATA: a customized carbon total correlation spectroscopy NMR metabolomics database. *Anal. Chem.* **2012,** *84* (21), 9395-9401.
255. Williams, R. B.; O'Neil-Johnson, M.; Williams, A. J.; Wheeler, P.; Pol, R.; Moser, A., Dereplication of natural products using minimal NMR data inputs. *Org. Biomol. Chem.* **2015,** *13* (39), 9957-9962.
256. Shoolery, J. N., *Varian Associates Technical Information Bulletin* **1959,** *2*.
257. Bremser, W., HOSE—a novel substructure code. *Anal. Chim. Acta* **1978,** *103* (4), 355-365.
258. Meiler, J.; Maier, W.; Will, M.; Meusinger, R., Using neural networks for 13C NMR chemical shift prediction–comparison with traditional methods. *J. Magn. Res.* **2002,** *157* (2), 242-252.
259. Steinbeck, C.; Kuhn, S., NMRShiftDB–compound identification and structure elucidation support through a free community-built web database. *Phytochemistry* **2004,** *65* (19), 2711-2717.
260. Binev, Y.; Aires-de-Sousa, J., Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *J. Chem. Inf. Comput. Sci.* **2004,** *44* (3), 940-945.
261. Toribio, A.; Bonfils, A.; Delannay, E.; Prost, E.; Harakat, D.; Henon, E.; Richard, B.; Litaudon, M.; Nuzillard, J.-M.; Renault, J.-H., Novel s eco-Dibenzopyrrocoline Alkaloid from Cryptocarya o ubatchensis. *Org. Lett.* **2006,** *8* (17), 3825-3828.
262. Iron, M. A., Evaluation of the Factors Impacting the Accuracy of 13C NMR Chemical Shift Predictions using Density Functional Theory· The Advantage of Long-Range Corrected Functionals. *J. Chem. Theory Comput.* **2017,** *13* (11), 5798-5819.
263. Elyashberg, M. E.; Blinov, K. A.; Williams, A. J.; Molodtsov, S. G.; Martin, G. E., Are Deterministic Expert Systems for Computer-Assisted Structure Elucidation Obsolete? *J. Chem. Inf. Model.* **2006,** *46* (4), 1643-1656.
264. Steinbeck, C., SENECA: A platform-independent, distributed, and parallel system for computer-assisted structure elucidation in organic chemistry. *J. Chem. Inf. Comput. Sci.* **2001,** *41* (6), 1500-1507.
265. Nuzillard, J. M.; Plainchont, B., Tutorial for the structure elucidation of small molecules by

means of the LSD software. *Magn. Reson. Chem.* **2018,** *56* (6), 458-468.
266. Lindel, T.; Junker, J.; Köck, M., COCON: From NMR correlation data to molecular constitutions. *Mol. Model. Ann.* **1997,** *3* (8), 364-368.
267. Nuzillard, J. M. PyLSD. http://eos.univ-reims.fr/LSD/JmnSoft/PyLSD/.
268. Hubert, J.; Chollet, S. b.; Purson, S.; Reynaud, R.; Harakat, D.; Martinez, A.; Nuzillard, J.-M.; Renault, J.-H., Exploiting the Complementarity between Dereplication and computer-assisted structure elucidation for the chemical profiling of natural cosmetic ingredients: Tephrosia purpurea as a case study. *J. Nat. Prod.* **2015,** *78* (7), 1609-1617.
269. Januar, H. I.; Zamani, N. P.; Soedharma, D.; Chasanah, E., Logic Structure Determination (LSD) as a Computer Assisted Structure Elucidation (CASE) for Molecular Structure Determination of Cytotoxic Cembranoids from Soft Coral. *Squalen Bull. Mar. Fish. Postharvest Biotechnol.* **2016,** *11* (1), 1-6.
270. Nyberg, N. T.; Duus, J. Ø.; Sørensen, O. W., Heteronuclear two-bond correlation: suppressing heteronuclear three-bond or higher NMR correlations while enhancing two-bond correlations even for vanishing 2 J CH. *J. Am. Chem. Soc.* **2005,** *127* (17), 6154-6155.
271. Boudesocque-Delaye, L.; Agostinho, D.; Bodet, C.; Thery-Kone, I.; Allouchi, H.; Gueiffier, A.; Nuzillard, J.-M.; Enguehard-Gueiffier, C. c., Antibacterial Polyketide Heterodimers from Pyrenacantha kaurabassana Tubers. *J. Nat. Prod.* **2015,** *78* (4), 597-603.
272. Senior, M. M.; Williamson, R. T.; Martin, G. E., Using HMBC and ADEQUATE NMR data to define and differentiate long-range coupling pathways: is the Crews rule obsolete? *J. Nat. Prod.* **2013,** *76* (11), 2088-2093.
273. Blinov, K.; Buevich, A.; Williamson, R.; Martin, G., The impact of LR-HSQMBC very long-range heteronuclear correlation data on computer-assisted structure elucidation. *Org. Biomol. Chem.* **2014,** *12* (47), 9505-9509.
274. Troche-Pesqueira, E.; Anklin, C.; Gil, R. R.; Navarro-Vázquez, A., Computer-Assisted 3D Structure Elucidation of Natural Products using Residual Dipolar Couplings. *Angew. Chem. Int. Ed.* **2017,** *56* (13), 3660-3664.
275. Ibrahim, N.; Allart-Simon, I.; De Nicola, G. R.; Iori, R.; Renault, J. H.; Rollin, P.; Nuzillard, J. M., Advanced NMR-Based Structural Investigation of Glucosinolates and Desulfoglucosinolates. *J. Nat. Prod.* **2018,** *81* (2), 323-334, URL: http://www.perchsolutions.com.
276. Plainchont, B.; Nuzillard, J.-M.; Rodrigues, G. V.; Ferreira, M.; Scotti, M. T.; Emerenciano, V. P., New improvements in automatic structure elucidation using the LSD (logic for structure determination) and the sistemat expert systems. *Nat. Prod. Commun.* **2010,** *5*, 763-770.
277. Costantin, M. B.; Ferreira, M. J.; Rodrigues, G. V.; Emerenciano, V. P., Computer-Assisted Approach to Structural Elucidation of Lignans. *Spectroscopy Lett.* **2008,** *41* (8), 405-421.
278. (a) Chhetri, B. K.; Lavoie, S.; Sweeney-Jones, A. M.; Kubanek, J., Recent trends in the structural revision of natural products. *Nat. Prod. Rep.* **2018,** *35* (6), 514-531; (b) Nicolaou, K. C.; Snyder, S. A., Chasing Molecules That Were Never There: Misassigned Natural Products and the Role of Chemical Synthesis in Modern Structure Elucidation. *Angew. Chem. Int. Ed.* **2005,** *44* (7), 1012-1044; (c) Yoo, H.-D.; Nam, S.-J.; Chin, Y.-W.; Kim, M.-S., Misassigned natural products and their revised structures. *Arch. Pharm. Res.* **2016,** *39* (2), 143-153.
279. Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J., Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014,** *48* (4), 2097-2098.
280. Sumner, L.; Lei, Z.; Nikolau, B.; Saito, K.; Roessner, U.; Trengove, R., Proposed quantitative and alphanumeric metabolite identification metrics. *Metabolomics* **2014,** *10* (6), 1047-1049.
281. Everett, J. R., A New Paradigm for Known Metabolite Identification in Metabonomics/Metabolomics: Metabolite Identification Efficiency. *Comput. Struct. Biotechnol. J.* **2015,** *13*, 131-144.
282. Bruhn, T.; Schaumlöffel, A.; Hemberger, Y.; Bringmann, G., SpecDis: Quantifying the Comparison of Calculated and Experimental Electronic Circular Dichroism Spectra. *Chirality* **2013,** *25* (4), 243-249.

283. Audoin, C.; Cocandeau, V.; Thomas, O. P.; Bruschini, A.; Holderith, S.; Genta-Jouve, G., Metabolome consistency: additional parazoanthines from the Mediterranean zoanthid *Parazoanthus axinellae*. *Metabolites* **2014,** *4* (2), 421-432.
284. Schymanski, E.; Neumann, S., The Critical Assessment of Small Molecule Identification (CASMI): Challenges and Solutions. *Metabolites* **2013,** *3* (3), 517.
285. Nikolic, D.; Jones, M.; Sumner, L.; Dunn, W., CASMI 2014: Challenges, Solutions and Results. *Curr. Metabolomics* **2017,** *5* (1), 5-17.
286. Nikolić, D., CASMI 2016: A manual approach for dereplication of natural products using tandem mass spectrometry. *Phytochem. Lett.* **2017,** *21*, 292-296.
287. Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M., Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal. Chem.* **2016,** *88* (16), 7946-58.
288. Wolfender, J.-L.; Queiroz, E. F.; Hostettmann, K., Phytochemistry in the microgram domain—a LC–NMR perspective. *Magn. Reson. Chem.* **2005,** *43* (9), 697-709.
289. Albert, K., *On-line LC-NMR and related techniques*. John Wiley & Sons: Chichester, 2002.
290. Sturm, S.; Seger, C.; Godejohann, M.; Spraul, M.; Stuppner, H., Conventional sample enrichment strategies combined with high-performance liquid chromatography-solid phase extraction-nuclear magnetic resonance analysis allows analyte identification from a single minuscule Corydalis solida plant tuber. *J. Chromatogr. A* **2007,** *1163* (1-2), 138-144.
291. Corcoran, O.; Spraul, M., LC-NMR-MS in drug discovery. *Drug Discov. Today* **2003,** *8* (14), 624-631.
292. (a) Exarchou, V.; Krucker, M.; van Beek, T. A.; Vervoort, J.; Gerothanassis, I. P.; Albert, K., LC-NMR coupling technology: recent advancements and applications in natural products analysis. *Magn. Reson. Chem.* **2005,** *43* (9), 681-687; (b) van der Hooft, J. J.; Mihaleva, V.; de Vos, R. C.; Bino, R. J.; Vervoort, J., A strategy for fast structural elucidation of metabolites in small volume plant extracts using automated MS-guided LC-MS-SPE-NMR. *Magn. Reson. Chem.* **2011,** *49 Suppl 1*, S55-60.
293. Guillarme, D.; Nguyen, D. T. T.; Rudaz, S.; Veuthey, J. L., Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *Eur. J. Pharm. Biopharm.* **2008,** *68* (2), 430-440.
294. Bohni, N.; Queiroz, E. F.; Wolfender, J.-L., On-line and At-line Liquid Chromatography Nuclear Magnetic Resonance and Related Micro-Nuclear Magnetic Resonance Methods in Natural Product Analysis. In *Encyclopedia of Analytical Chemistry (Plant Analysis: Chemical and Biological)*, Hostettmann, K.; Stuppner, H., Eds. John Wiley & Sons: Chichester, UK, 2014; pp 1-31.
295. van der Hooft, J. J. J.; Akermi, M.; Ünlü, F. Y.; Mihaleva, V.; Roldan, V. G.; Bino, R. J.; de Vos, R. C. H.; Vervoort, J., Structural Annotation and Elucidation of Conjugated Phenolic Compounds in Black, Green, and White Tea Extracts. *J. Agric. Food Chem.* **2012,** *60* (36), 8841-8850.
296. Sumner, L. W.; Lei, Z.; Nikolau, B. J.; Saito, K., Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. *Nat. Prod. Rep.* **2015,** *32* (2), 212-29.
297. Gomes, N. G. M.; Pereira, D. M.; Valentao, P.; Andrade, P. B., Hybrid MS/NMR methods on the prioritization of natural products: Applications in drug discovery. *J. Pharmaceut. Biomed. Anal.* **2018,** *147*, 234-249.
298. Liu, B.; Kongstad, K. T.; Sun, Q.; Nyberg, N. T.; Jager, A. K.; Staerk, D., Dual High-Resolution alpha-Glucosidase and Radical Scavenging Profiling Combined with HPLC-HRMS-SPE-NMR for Identification of Minor and Major Constituents Directly from the Crude Extract of Pueraria lobata. *J. Nat. Prod.* **2015,** *78* (2), 294-300.
299. Lima, R. D. L.; Gramsbergen, S. M.; Van Staden, J.; Jager, A. K.; Kongstad, K. T.; Staerk, D., Advancing HPLC-PDA-HRMS-SPE-NMR Analysis of Coumarins in Coleonema album by Use of Orthogonal Reversed -Phase C-18 and Pentafluorophenyl Separations. *J. Nat. Prod.* **2017,** *80* (4), 1020-1027.
300. Bingol, K.; Brüschweiler, R., Knowns and unknowns in metabolomics identified by multidimensional NMR and hybrid MS/NMR methods. *Curr. Opin. Biotechnol.* **2017,** *43*, 17-24.

301. Bingol, K.; Bruschweiler-Li, L.; Li, D.; Zhang, B.; Xie, M.; Brüschweiler, R., Emerging new strategies for successful metabolite identification in metabolomics. *Bioanalysis* **2016,** *8* (6), 557-573.
302. van der Hooft, J. J.; de Vos, R. C.; Mihaleva, V.; Bino, R. J.; Ridder, L.; de Roo, N.; Jacobs, D. M.; van Duynhoven, J. P.; Vervoort, J., Structural elucidation and quantification of phenolic conjugates present in human urine after tea intake. *Anal. Chem.* **2012,** *84* (16), 7263-71.
303. Bingol, K.; Bruschweiler-Li, L.; Yu, C.; Somogyi, A.; Zhang, F.; Bruschweiler, R., Metabolomics Beyond Spectroscopic Databases: A Combined MS/NMR Strategy for the Rapid Identification of New Metabolites in Complex Mixtures. *Anal. Chem.* **2015,** *87* (7), 3864-70.
304. Bingol, K.; Bruschweiler, R., NMR/MS Translator for the Enhanced Simultaneous Analysis of Metabolomics Mixtures by NMR Spectroscopy and Mass Spectrometry: Application to Human Urine. *J. Proteome Res.* **2015,** *14* (6), 2642-2648.
305. Bingol, K.; Zhang, F. L.; Bruschweiler-Li, L.; Bruschweiler, R., Carbon Backbone Topology of the Metabolome of a Cell. *J. Am. Chem. Soc.* **2012,** *134* (21), 9006-9011.
306. Bakiri, A.; Plainchont, B.; Emerenciano, V. D.; Reynaud, R.; Hubert, J.; Renault, J. H.; Nuzillard, J. M., Computer-aided Dereplication and Structure Elucidation of Natural Products at the University of Reims. *Mol. Inf.* **2017,** *36* (10), 1700027.
307. Bertrand, S.; Azzollini, A.; Nievergelt, A.; Boccard, J.; Rudaz, S.; Cuendet, M.; Wolfender, J.-L., Statistical correlations between HPLC activity-based profiling results and NMR/MS microfractions data to deconvolute bioactive compounds in mixture. *Molecules* **2016,** *21* (3), 259.
308. Zhokhov, A. K.; Loskutov, A. Y.; Rybal'chenko, I. V., Methodological Approaches to the Calculation and Prediction of Retention Indices in Capillary Gas Chromatography. *J. Anal. Chem.* **2018,** *73* (3), 207-220.
309. Stanstrup, J.; Neumann, S.; Vrhovsek, U., PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal. Chem.* **2015,** *87* (18), 9421-8.
310. Eugster, P. J.; Boccard, J.; Debrus, B.; Bréant, L.; Wolfender, J.-L.; Martel, S.; Carrupt, P.-A., Retention time prediction for dereplication of natural products (CxHyOz) in LC-MS metabolite profiling. *Phytochemistry* **2014,** *108*, 196-207.
311. Randazzo, G. M.; Tonoli, D.; Hambye, S.; Guillarme, D.; Jeanneret, F.; Nurisso, A.; Goracci, L.; Boccard, J.; Rudaz, S., Prediction of retention time in reversed-phase liquid chromatography as a tool for steroid identification. *Anal. Chim. Acta* **2016,** *916*, 8-16.
312. Hall, L. M.; Hill, D. W.; Bugden, K.; Cawley, S.; Hall, L. H.; Chen, M. H.; Grant, D. F., Development of a Reverse Phase HPLC Retention Index Model for Nontargeted Metabolomics Using Synthetic Compounds. *J. Chem. Inf. Model.* **2018,** *58* (3), 591-604.
313. Zhou, Z. W.; Shen, X. T.; Tu, J.; Zhu, Z. J., Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2016,** *88* (22), 11084-11091.
314. Boycheva, S.; Daviet, L.; Wolfender, J. L.; Fitzpatrick, T. B., The rise of operon-like gene clusters in plants. *Trends Plant Sci.* **2014,** *19* (7), 447-59.
315. Blin, K.; Wolf, T.; Chevrette, M. G.; Lu, X.; Schwalen, C. J.; Kautsar, S. A.; Suarez Duran, H. G.; de los Santos, Emmanuel L C.; Kim, H. U.; Nave, M.; Dickschat, J. S.; Mitchell, D. A.; Shelest, E.; Breitling, R.; Takano, E.; Lee, S. Y.; Weber, T.; Medema, M. H., antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.* **2017,** *45* (Web Server issue), W36-W41.
316. Skinnider, M. A.; Dejong, C. A.; Rees, P. N.; Johnston, C. W.; Li, H.; Webster, Andrew L. H.; Wyatt, M. A.; Magarvey, N. A., Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Res.* **2015,** *43* (20), 9645-62.
317. Medema, M. H.; Fischbach, M. A., Computational approaches to natural product discovery. *Nat. Chem. Biol.* **2015,** *11* (9), 639-648.
318. Schneider, B.; Hölscher, D., Laser microdissection and cryogenic nuclear magnetic resonance spectroscopy: an alliance for cell type-specific metabolite profiling. *Planta* **2007,** *225* (3), 763-770.
319. Moussaieff, A.; Rogachev, I.; Brodsky, L.; Malitsky, S.; Toal, T. W.; Belcher, H.; Yativ, M.; Brady, S. M.; Benfey, P. N.; Aharoni, A., High-resolution metabolic mapping of cell types in plant roots. *Proc. Natl. Acad. Sci. U.S.A.* **2013,** *110* (13), E1232.

320. Watanabe, R.; Sugai, C.; Yamazaki, T.; Matsushima, R.; Uchida, H.; Matsumiya, M.; Takatsu, A.; Suzuki, T., Quantitative Nuclear Magnetic Resonance Spectroscopy Based on PULCON Methodology: Application to Quantification of Invaluable Marine Toxin, Okadaic Acid. *Toxins* **2016,** *8* (10), 294.
321. (a) Awad, H.; Khamis, M. M.; El-Aneed, A., Mass Spectrometry, Review of the Basics: Ionization. *Applied Spectroscopy Reviews* **2014,** *50* (2), 158-175; (b) Kauppila, T. J.; Syage, J. A.; Benter, T., Recent developments in atmospheric pressure photoionization-mass spectrometry. *Mass Spectrom. Rev.* **2015,** *36* (3), 423-449.
322. Wernisch, S.; Pennathur, S., Evaluation of coverage, retention patterns, and selectivity of seven liquid chromatographic methods for metabolomics. *Anal. Bioanal. Chem.* **2016,** *408* (22), 6079-6091.
323. Strehmel, N.; Kopka, J.; Scheel, D.; Böttcher, C., Annotating unknown components from GC/EI-MS-based metabolite profiling experiments using GC/APCI(+)-QTOFMS. *Metabolomics* **2013,** *10* (2), 324-336.
324. Wakimoto, T., Toward the Dark Matter of Natural Products. *Chem. Record* **2017,** *17* (11), 1124-1134.
325. Matsuda, Y.; Mitsuhashi, T.; Lee, S.; Hoshino, M.; Mori, T.; Okada, M.; Zhang, H.; Hayashi, F.; Fujita, M.; Abe, I., Astellifadiene: Structure Determination by NMR Spectroscopy and Crystalline Sponge Method, and Elucidation of its Biosynthesis. *Angew. Chem. Int. Ed.* **2016,** *55* (19), 5785-5788.
326. (a) Gruene, T.; Wennmacher, J. T. C.; Zaubitzer, C.; Holstein, J. J.; Heidler, J.; Fecteau-Lefebvre, A.; De Carlo, S.; Muller, E.; Goldie, K. N.; Regeni, I.; Li, T.; Santiso-Quinones, G.; Steinfeld, G.; Handschin, S.; van Genderen, E.; van Bokhoven, J. A.; Clever, G. H.; Pantelic, R., Rapid structure determination of microcrystalline molecular compounds using electron diffraction. *Angew. Chem. Int. Ed.* **2018**, DOI: 10.1002/anie.201811318; (b) Christopher, G. J.; Michael W., M.; Johan, H.; Tyler J., F.; Brian M., S.; Jose A., R.; Hosea, N.; Tamir, G., The CryoEM Method MicroED as a Powerful Tool for Small Molecule Structure Determination. *ChemRChiv* **2018**, https://chemrxiv.org/articles/The_CryoEM_Method_MicroED_as_a_Powerful_Tool_for_Small_Molecule_Structure_Determination/7215332.