Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification

Pierre-Marie Allard[1], Grégory Genta-Jouve[2] and Jean-Luc Wolfender[1]

Natural products (NPs) research is changing and rapidly adopting cutting-edge tools, which radically transform the way to characterize extracts and small molecules. With the innovations in metabolomics, early integration of deep metabolome annotation information allows to efficiently guide the isolation of valuable NPs only and, in parallel, to generate massive metadata sets for the study of given extracts under various perspectives. This is the case for chemotaxonomy studies where common biosynthetic traits among species can be evidenced, but also for drug discovery purpose where such traits, in combination with bioactivity studies on extracts, may evidence bioactive molecules even before their isolation. One of the major bottlenecks of such studies remains the level of accuracy at which NPs can be identified. We discuss here the advancements in LC–MS and associated mining methods by addressing what would be ideal and what is achieved today. We propose future developments for reinforcing generic NPs databases both in the spectral and structural dimensions by heading towards a virtuous metabolite identification cycle allowing annotation of both known and unreported metabolites in an iterative manner. Such approaches could significantly accelerate and improve our knowledge of the huge chemodiversity found in nature.

Addresses
[1] School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, CMU—Rue Michel Servet 1, 1211 Geneva 11, Switzerland
[2] Equipe C-TAC, UMR CNRS 8638 COMETE—Université Paris Descartes, 4 Avenue de l'observatoire, 75006 Paris, France

Corresponding author: Wolfender, Jean-Luc (jean-luc.wolfender@unige.ch)

Current Opinion in Chemical BiologyCurrent Opinion in Chemical Biology 2017, **36**:40–49

This review comes from a themed issue on **Omics**

Edited by **Georg Pohnert** and **Frank C Schroeder**

http://dx.doi.org/10.1016/j.cbpa.2016.12.022

1367-5931/Published by Elsevier Ltd.

Introduction

In Natural Products (NPs) research, a major task is the lengthy process of *de novo* identification of the diverse metabolites occurring in the metabolomes of plants, microorganisms and other biological sources. This process may be driven by the search for new NPs to populate screening libraries or for chemotaxonomy and chemical ecological studies. It can also result from bioguided isolation and focus on bioactive hits only. Metabolomics, by providing a holistic survey of natural extract composition, has thus gained increasing importance in NPs research, but the main bottleneck still lies in the unambiguous identification of the highlighted chemicals.

With the recent progress made in metabolite profiling [1], the possibility of generating high-quality spectroscopic data on natural extracts has increased tremendously, especially with the introduction of high-resolution mass spectrometry (HRMS) [2] coupled to ultra-high-performance liquid chromatography (UHPLC) [3]. Such platforms can generate structural information for hundreds to thousands of metabolites in crude natural extracts [4]. Today, the challenge lies in annotating this high-quality spectral information [5]. This process, in addition to simple molecular formula (MF) determination [6], requires maximizing the application of orthogonal filters (e.g., chemotaxonomy, retention time [7], fragmentation patterns [8,9] . . .) to generate hypotheses with scoring results that should document the validity of the automated interpretation.

This paper presents recent tools for data acquisition, treatment, organization and annotation in NPs research. Current and ideal strategies for each stage are discussed, and a conceptual workflow for annotation of known and unreported metabolites is proposed.

Data acquisition

For a pertinent metabolome annotation, a first key step is to generate high-quality metabolite profiling data. Even if natural extracts can be directly analyzed (NMR [10], Direct Infusion HRMS [11]), the complexity of the samples often requires preliminary chromatographic separation step. In the case of MS detection, this step is crucial for the separation of isomers and can reduce ion suppression effects. LC–MS currently represents the most versatile and commonly used approach for profiling

extracts [4], and we will focus only on data generated by this type of platform.

LC

The ideal LC separation should provide resolution of all metabolites in a given matrix, particularly for the separation of isomers. It should be versatile enough to address the wide range of physicochemical characters of NPs (ranging from polar, permanently charged compounds and sugars to lipophilic and aliphatic compounds)

The development of UHPLC has made it possible to attain high peak capacities within shorter analysis times than conventional HPLC by using sub-2 μm particle phases [12]. On the other hand, the availability of numerous phase chemistries (e.g., RP18, HILIC) provides adequate retention of most types of metabolome constituents. The range of metabolites profiled using a single method can be extended by the use of a supercritical mobile phase in UHPLC [13]. The coupling of RPLC and HILIC has also been shown to offer greatly increased coverage of the metabolome when analyzing plasma and urine matrices [14]. At the moment, no single analytical method can profile all metabolites in a given sample. The combination of orthogonal separation modes (2D-LC or ion mobility) should provide better metabolome coverage [15•].

One of the main limitations of LC is its lack of genericity. While generic linear gradients are conventionally applied in metabolomics, the retention behaviors are strongly-dependent on the column type and the chromatographic system. An approach called "retention projection" aims at standardization of retention time (RT) across laboratories [16•]. The development and adoption of this type of generic retention parameter should be helpful in the metabolite annotation process for LC–MS data.

MS

The ideal MS system should be able to ionize and detect all metabolites over an extensive dynamic range. It should provide high-quality data for unambiguous molecular weight determination, subsequent MF assignment and should generate information-rich fragmentation spectra on every analytes

Despite the impressive evolution of MS in terms of resolution, sensitivity and acquisition rates, each actual type of MS architecture (ionization source/collision cell type/analyzers) requires compromises. For example, HRMS is achieved at the expense of a high scan speed. In addition, ionization is compound-dependent, and various sources (ESI, APCI, APPI) exhibit different selectivity and yield different degenerate features for [M] (e.g., (de)protonated ions, dimers, adducts, in-source fragments) or no detection at all. Thus, development of instruments able to rapidly switch polarity modes and ionization sources within a single LC run would be of interest. Fragmentation is also dependent on the collision mode (CID, HCD, ETD) and applied energy [17]. Here also, there is a strong need to find ways to normalize the acquisition of fragmentation spectra across laboratories.

Although the current technology provides non-exhaustive metabolome coverage with only partial fragmentation information, the enormous amount of data generated requires efficient automated treatment methods to extract meaningful information.

Data treatment

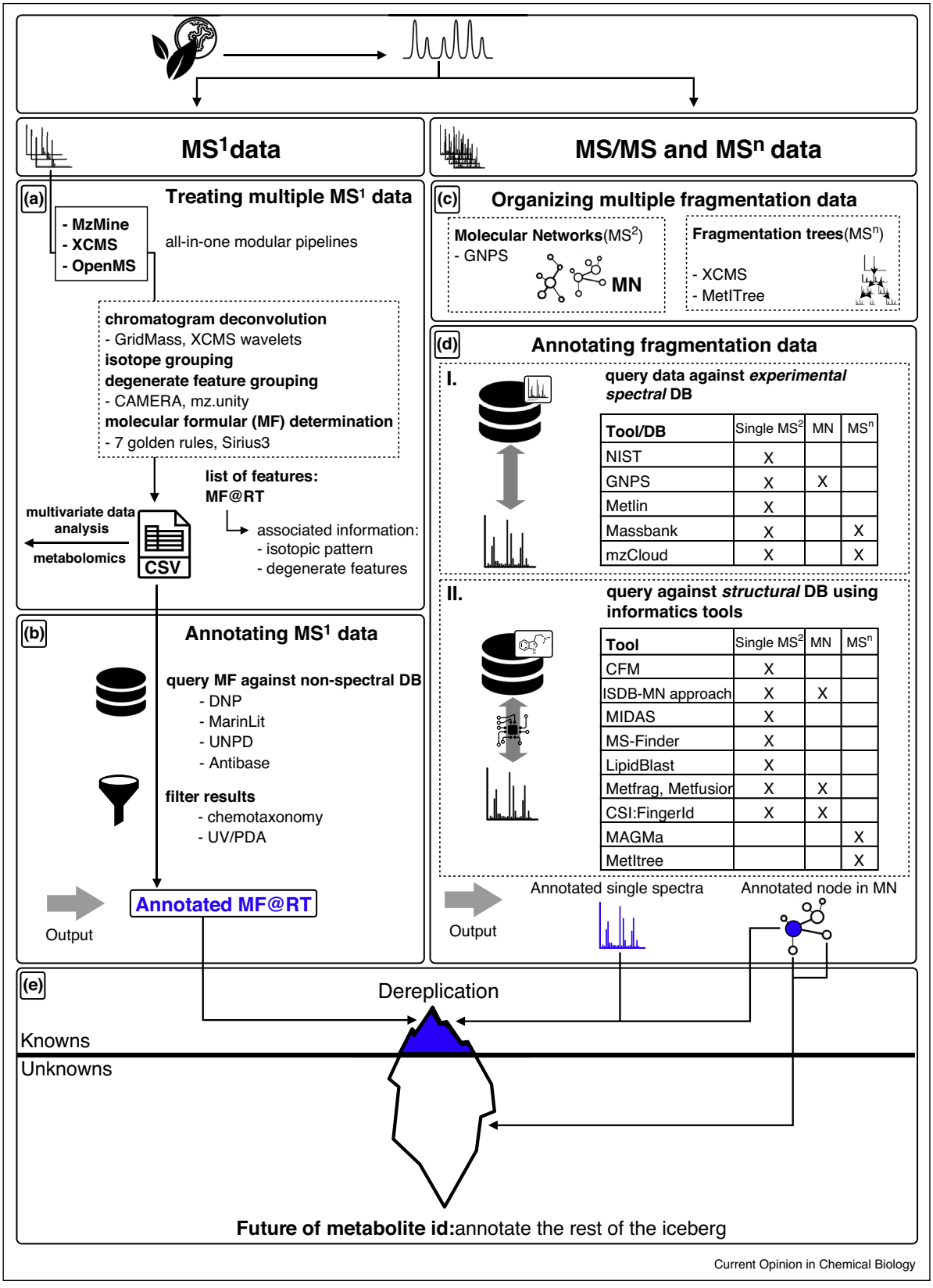
The perfect LC–MS data treatment workflow should provide a list of all analytes with a maximum amount of information associated to each entry: a unique MF, a generic parameter linked to RT, a list of all analytes' degenerate features and the associated fragmentation spectra. This data treatment solution should be instrument-independent and modular

At present, a number of software programs have been designed to address each step of the LC–MS data treatment procedure (Figure 1), and various open-source pipelines integrate these modular solutions and allow the conversion of raw data to a clean annotated peak list (e.g., XCMS, MZmine or OpenMS). (Step A in Figure 1.)

An important step is the disambiguation of molecular features (e.g., molecular ions, adducts, in-source fragments, dimers). The mz.unity algorithm is a recently developed open solution which allows efficient grouping of these degenerate features [18•]. Drawing relations between these peaks allows a better use of resources during the data treatment through cleaning of datasets. This tool is a perfect example illustrating that a superior level of information can be obtained by the establishment of a relationship between elements of a dataset without additional data acquisition. MF determination can be considered as the first step of the metabolite identification process. The increased resolution and mass accuracy of modern MS have allowed easier access to MF. The combination of analyte isotopic patterns, heuristic rules [19] and/or fragmentation information [6] can now render this process straightforward. Except in special cases, MF determination should now not be a limiting step in the metabolite identification process.

For the acquisition of additional spectral information on the analytes, MS/MS is commonly used. Two main acquisition modes can be differentiated: Data-Dependent Analysis (DDA) and Data-Independent Analysis (DIA). DDA provides high-quality data but fewer MS/MS acquisitions during the run, whereas DIA allows better metabolite coverage at the expense of losing the direct link with the parent ion. Recently, various informatics approaches (e.g., MS-DIAL, MetDIA) have been tailored for the treatment of DIA data in metabolomics

Figure 1



Main tools to treat, organize and annotate LC–MS data.

and will undoubtedly be increasingly used for deeper metabolome coverage [20,21].

Automated methods are necessary to organize and interpret these data. The following sections will discuss the available solutions to organize fragmentation data and link the fragmentation pattern to a molecular structure.

Data organization

The optimal MS-data organization tool should link all previously listed analytes (MF@RT) according to selectable parameters: spectral similarity, structural similarity, presence of a particular structural moiety, presence of a particular atom or a specific physico-chemical property. It should be possible to map multiple information layers (bioactivities, taxonomy, gene sequences) on top of the organized data. The tool should be Free, open-source, compatible with existing databases (DBs) and should allow the import of data being acquired by different teams in a traceable manner

Today, the major massive MS/MS data organization tool is the Global Natural Products Social molecular networking platform (GNPS) [22••]. GNPS is a hybrid tool making it possible to organize MS/MS data as spectral or molecular networks (MNs) (Step C in Figure 1). MNs are generated by a spectral alignment algorithm establishing a similarity score between individual spectra. Since fragmentation spectra generally reflect the chemical structures of the fragmented ions, it becomes possible to represent a whole metabolome as subgroups (clusters) of similar structures. Additionally MNs of researchers worldwide can be compared and annotated by querying spectral DBs present on the platform. This way of representing a metabolome is radically new and offers a real paradigm shift in NPs chemistry. The network representation gives the ability to overlay information in a visual manner, allowing the representation of multiple levels of information within a single MN. Today, MNs link *m/z* features, but tools making it possible to refine the spectral network to a MN of unique identifiers per analyte (MF@RT) would be desirable.

Also, MS^n data can be organized, processed, shared, visualized and compared using solutions such as MetiTree (Step C in Figure 1) [23].

Once the LC–MS data are organized as a network of features (ideally unique identifiers) and their associated acquired metadata (e.g., MS^2, MS^n, sample preparation, biological origin), the next step is the annotation of these data.

Data annotation

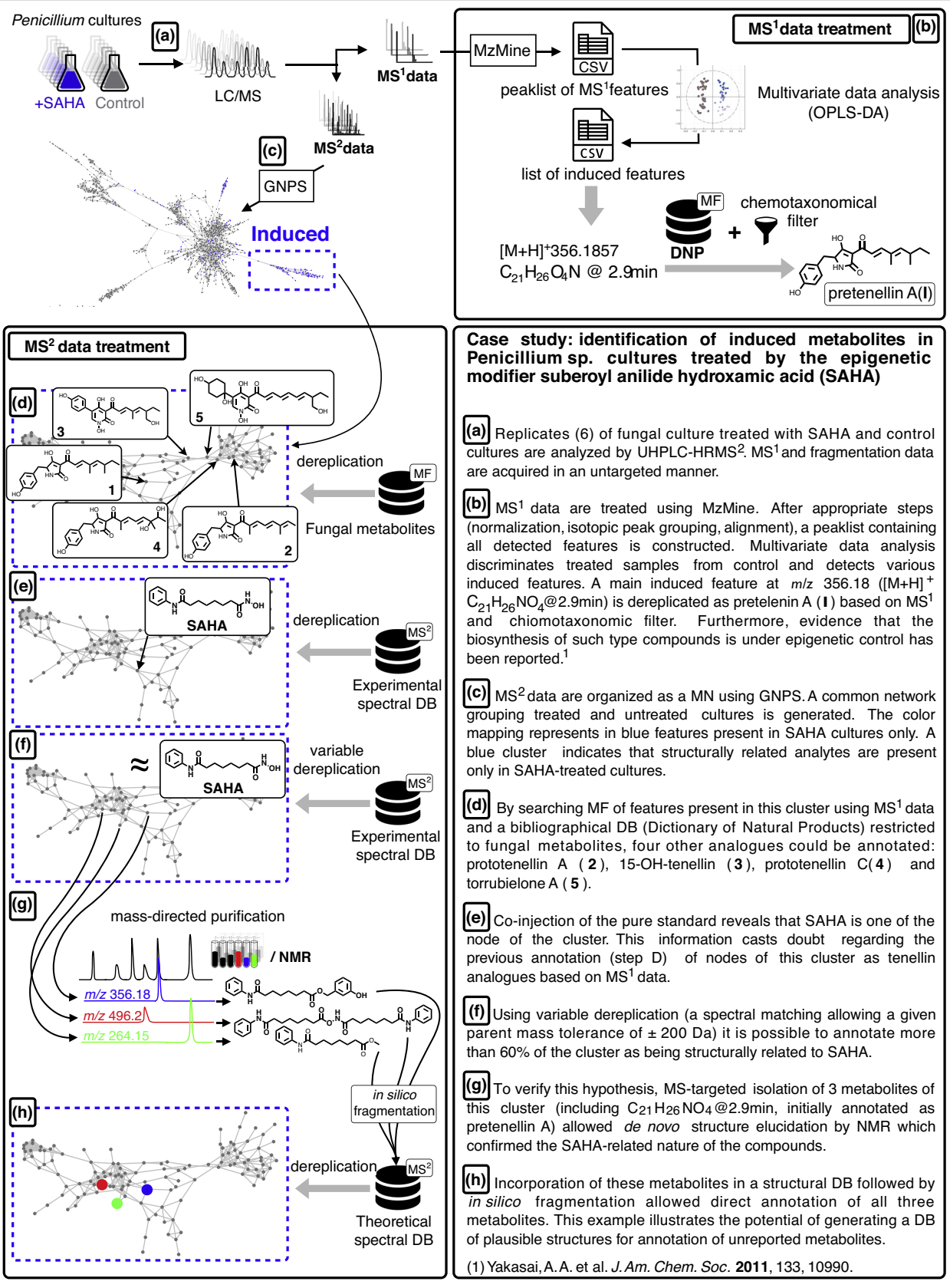
Data annotation should ideally link unique identifiers to a single molecular structure, unambiguously.

When a unique MF is established, annotation can be undertaken by matching this MF against a DB without requiring further spectral information. Various DBs of NPs exist, and some, such as the Universal Natural Products Database (UNPD, http://pkuxxj.pku.edu.cn/UNPD/), are freely accessible (for an extended list, see Ref. [24•]). Nevertheless, the isomeric nature of numerous NPs hampers further precise annotation, and the main challenge then lies in the scoring of their annotation. To reduce the numbers of hits, chemotaxonomic filters can be applied, which may drastically limit the number of possible structures for a given MF (Step B in Figure 1). For further isomer ranking, additional fragmentation information needs to be considered by (i) searching an experimental spectral DB by direct spectral match (Step D1 in Figure 1) or (ii) searching a structural DB via a computational approach linking the experimental spectra to the structures (Step D2 in Figure 1). A direct spectral match requires access to experimental spectral DB, which are still limited in size (ca. 20 000 compounds) [25] compared to the actual number of known NPs (> 200 000). To overcome these limits, a number of tools have been developed [8,26•]. Some make it possible to generate theoretical fragmentation spectra from a structure (e.g., CFM-ID [27], MS-Finder [28], Metfrag [29]), while others compute a fingerprint from the experimental fragmentation spectrum and then match it against a structural DB (CSI:FingerID [25]). These tools are modular and can be combined into different annotation pipelines [30]. They have proven to perform well (see CASMI 2016 contest, http://casmi-contest.org) and will clearly be crucial parts of metabolite identification pipelines for NPs chemists in the coming years.

Using one of these tools (CFM-ID) and spectral matching (Tremolo [31]), we developed an integrated pipeline, ISDB-MN (*In Silico* Database-Molecular Networking), which allows the batch annotation of full MNs against an *in silico* fragmented DB of NPs [32•]. This pipeline and the full *in silico* fragmentation DB are freely available at http://oolonek.github.io/ISDB/. An example of the process followed for the identification of induced metabolites in fungal cultures elicited by the epigenetic modifier SAHA is shown in Box 1 . This case study illustrates the value of considering fragmentation data and MN relationships in addition to MS^1 data during the annotation process. MS-targeted isolation followed by *de novo* NMR identification confirmed the putative class annotation of the variable dereplication process [32].

In general, regarding data annotation, improvements are needed in (i) the *scoring of the annotation results* and (ii) the *annotation of unknowns* (compounds absent from experimental spectral AND structural DBs). We will address both points in the following sections.

Box 1 Case study: identification of induced metabolites in *Penicillium* sp. cultures treated by the epigenetic modifier suberoyl anilide hydroxamic acid (SAHA).



Case study: identification of induced metabolites in Penicillium sp. cultures treated by the epigenetic modifier suberoyl anilide hydroxamic acid (SAHA)

(a) Replicates (6) of fungal culture treated with SAHA and control cultures are analyzed by UHPLC-HRMS^2. MS^1 and fragmentation data are acquired in an untargeted manner.

(b) MS^1 data are treated using MzMine. After appropriate steps (normalization, isotopic peak grouping, alignment), a peaklist containing all detected features is constructed. Multivariate data analysis discriminates treated samples from control and detects various induced features. A main induced feature at *m/z* 356.18 ($[M+H]^+$ $C_{21}H_{26}NO_4$@2.9min) is dereplicated as pretelenin A (**I**) based on MS^1 and chiomotaxonomic filter. Furthermore, evidence that the biosynthesis of such type compounds is under epigenetic control has been reported.[1]

(c) MS^2 data are organized as a MN using GNPS. A common network grouping treated and untreated cultures is generated. The color mapping represents in blue features present in SAHA cultures only. A blue cluster indicates that structurally related analytes are present only in SAHA-treated cultures.

(d) By searching MF of features present in this cluster using MS^1 data and a bibliographical DB (Dictionary of Natural Products) restricted to fungal metabolites, four other analogues could be annotated: prototenellin A (**2**), 15-OH-tenellin (**3**), prototenellin C(**4**) and torrubielone A (**5**).

(e) Co-injection of the pure standard reveals that SAHA is one of the node of the cluster. This information casts doubt regarding the previous annotation (step D) of nodes of this cluster as tenellin analogues based on MS^1 data.

(f) Using variable dereplication (a spectral matching allowing a given parent mass tolerance of ± 200 Da) it is possible to annotate more than 60% of the cluster as being structurally related to SAHA.

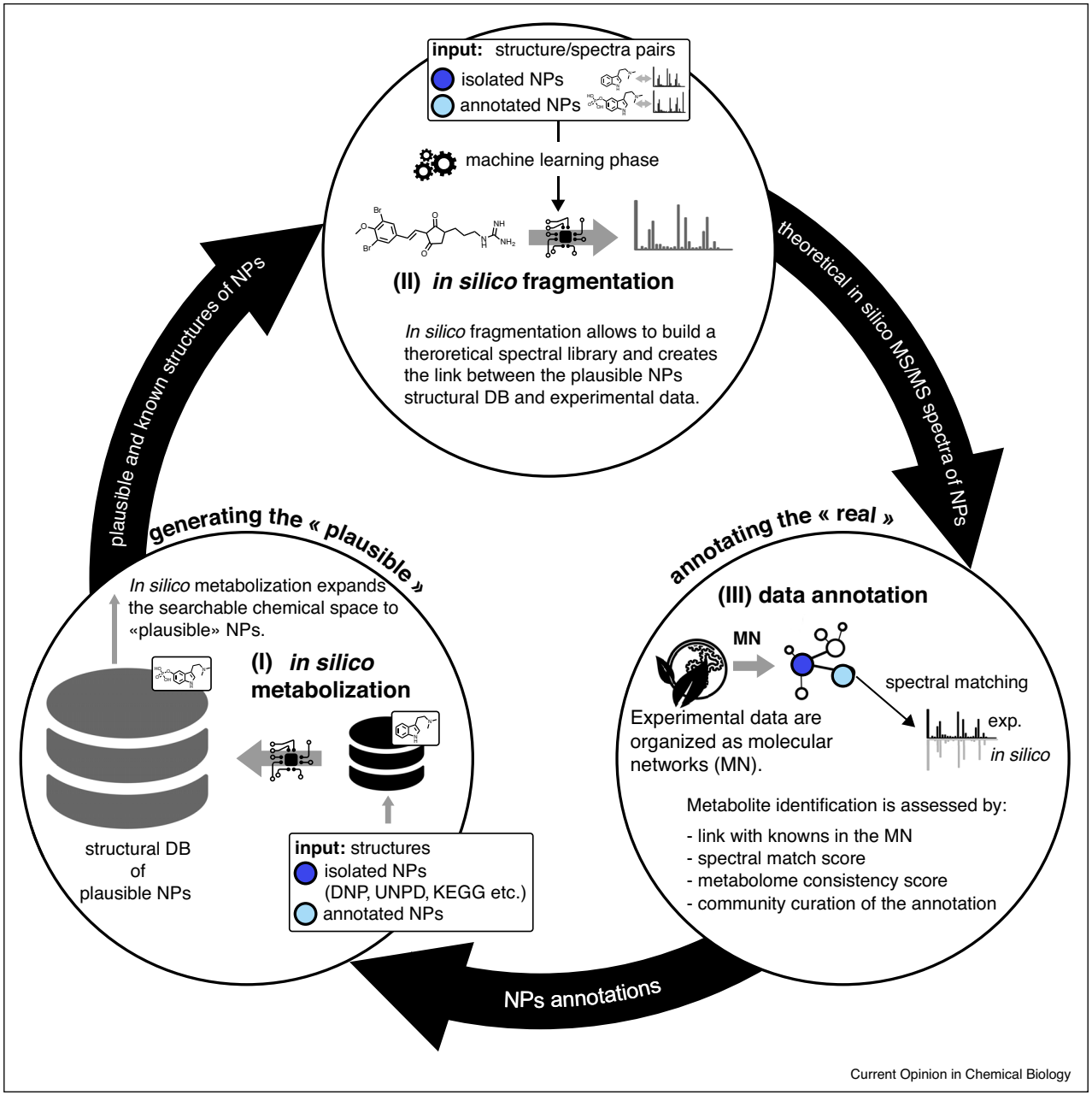
(g) To verify this hypothesis, MS-targeted isolation of 3 metabolites of this cluster (including $C_{21}H_{26}NO_4$@2.9min, initially annotated as pretenellin A) allowed *de novo* structure elucidation by NMR which confirmed the SAHA-related nature of the compounds.

(h) Incorporation of these metabolites in a structural DB followed by *in silico* fragmentation allowed direct annotation of all three metabolites. This example illustrates the potential of generating a DB of plausible structures for annotation of unreported metabolites.

(1) Yakasai, A. A. et al. *J. Am. Chem. Soc.* **2011**, 133, 10990.

Current Opinion in Chemical Biology

Figure 2



Conceptual virtuous cycle of metabolite identification.

Scoring results

An annotation is valuable only if it can be scored within an established reference system to rate its confidence. Ideally, for LC–MS data of natural extracts, this score should rate spectral comparison but should also assess other aspects leading to the annotation, such as the retention behavior of the analyte or the consistency of the annotated structure with its metabolic context. Ideally, a meta-score integrating these various aspects should be established. Constitutive subscores should be accessible and offer a way to trace back all information used to generate the annotation

One common method of comparison is to calculate a similarity score between two spectra (e.g., experimental vs. database). In MS, typical spectral matching scores are the result of operations such as the cosine similarity or the dot product. They are used for EIMS and MS/MS DB searching and applied in MS/MS spectra tools such as

NIST MS Search, CFM-ID [33] or GNPS [22]. Numerous calculations are possible to establish similarity measurement between two objects and other equations than cosine similarity or dot product might be better suited for comparison of experimental and *in silico* fragmented MS/MS spectra [34]. An evaluation of such equations would be of interest. A perfect spectral score is obtained when spectra are compared to a DB of experimental spectra of standards acquired under the same experimental conditions. When the comparison is performed against *in silico* spectral DB, where lower value scores are expected, complementary scores should be considered. During the generation of a MN, a first spectral match scoring is performed between similar compounds fragmented under the same experimental conditions allowing their clustering, the quantification of their similarities and the propagation of the annotation. During the annotation of a MN, a second spectral match can be performed against either experimental spectral or *in silico*-generated spectra calculated from a structural DB (Step D in Figure 1). To restrict the search at this stage, chemotaxonomic filters can also be applied.

Once a MN is annotated, and a *spectral* network is thus converted into a *structural* network, it becomes possible to apply other types of scoring to refine the annotation results. Physicochemical properties can be extracted from the structures, thus allowing the possibility of simulating, for example, RT behavior for comparison against experimental ones. Generic tools for this purpose are lacking, but the approach has been successfully applied to a limited set of compounds (e.g., steroids) [7]. Metabolome consistency (the coherence of a metabolite within a biochemical pathway) [35•] can also be inferred by checking structural similarity. This task is usually performed by visual inspection of the structures, but calculating structural comparison scores such as the Tanimoto index [36] or using solutions such as ChemGPS-NP [37] should provide ways to automate this stage.

All these scores should ultimately be considered and combined into a meta-score to assess the global relevance of the annotation.

Identifying the real within the possible

In deep metabolome annotation, the next challenge will consist in the annotation of the unreported NPs, which still represent an important part of the metabolome (E in Figure 1) [38]. For this, a DB of plausible unknowns must be generated. Organic molecule structures can be systematically generated using algorithms [39], but in the case of NPs research, it is much more desirable to have a plausible structural DB, which is ideally constructed according to biosynthetic rules. This approach has been recently developed to expand the chemodiversity starting from several structural DBs, using an algorithm called Biochemical Network Integrated Computational Explorer (BNICE) [40••]. Using this approach (Metabolic *In silico* Network Expansions (MINES)), the initial DBs were expanded by a factor 50. This type of expanded structural DB can be further converted into spectral DB for matching experimental data. For example, starting from a set of 75 metabolites from green tea, a DB of 27 170 compounds has been generated, allowing the annotation of previously unreported tea metabolites in urine [41]. Likewise, the whole human metabolome DB (HMDB) has been expanded by more than 40 times and fragmented *in silico* (MyCompoundID) [42]. In NPs research, a similar approach has been applied to a marine zoanthid and allowed the description of five previously unreported hydantoin alkaloids through the comparison of fragmentation patterns with the patterns of previously isolated parazoathines and the MS/MS spectra simulation of *in silico*-predicted compounds according to metabolome consistency [35]. In addition to biochemical pathways/reactions that can be stored in existing DBs (e.g., KEGG [43] MetaCyc [44], CathaCyc [45]), genome mining approaches with bioinformatics tools such as antiSMASH, PKMiner or SMURF [46] can also be considered for guiding the *in silico* expansion of structural DBs. Genome mining in combination with MS profiling has proven to be an efficient method, for example, for the discovery of columbamides in *Moorea bouilonii* [47,48].

By gathering the tools described throughout this paper, we propose to move towards a virtuous cycle of metabolite annotation. A conceptual workflow is presented in Figure 2. It consists of (i) expansion of a structural DB of all known NPs with plausible metabolites generated *in silico* according to known biosynthetic rules or genome-inferred information; (ii) transformation of this expanded structural DB into a spectral DB using *in silico* fragmentation tools; and (iii) massive annotation of the experimental data by spectral matching and adequate filtering. Once the metabolites have been annotated, this information can iteratively feed the annotation cycle through machine learning processes at the *in silico* metabolization stage and/or at the *in silico* fragmentation stage. To avoid the risk of transforming this virtuous cycle into a vicious cycle, annotation validation is mandatory. This task should be performed by the development of a robust meta-score, as described above, combined with MS-targeted micro-isolations of well selected metabolites to establish solid anchor points in given metabolite clusters.

Conclusion and perspectives

The rapid developments of LC–HRMS/MS metabolite profiling and metabolomics of natural extracts have deeply changed the possibilities to conduct NPs research. In our opinion, several classical bioactivity-guided isolation studies, considered by some as 'fishing expeditions', can now be converted into a 'cherry picking' of valuable NPs only. As mentioned, much effort in metabolomics must still be made to reach the level at which NPs can be

unambiguously identified. Compared to proteomics, the physical nature of small metabolites requires other routes of spectral interpretation. An efficient combination of LC-HRMS/MS information obtained on extended sets of extracts through adapted mining methods will efficiently reveal the richness of natural metabolomes and provide ways to carry directed investigation in a much more rational and efficient way.

Valuable NPs annotation requires multiple scores from orthogonal types of information (HR MS/MS, retention behaviors, metabolome consistency). One key type of experimental information, which is acquired but often ignored, is analyte retention times. Despite some preliminary work, many efforts must be made on the use of this parameter because it indirectly reflects the physicochemical properties of an analyte. It is thus important to find ways to correlate RT information with a generic parameter that can be compared and matched among analytical platforms. Efforts in the same direction, towards a standardization of ion mobility behavior through cross-sections [15•], must also be performed.

Following such a deep annotation process, metabolomes of diverse species should ideally be compared, grouped, aligned and queried within a meta-metabolomics DB. This data matrix could be structured according to the taxonomic relationships of the analyzed samples on one side and structural relations between metabolites on the other. Co-occurring metabolites between closely related species should be easily evidenced in such a matrix and offer a cross-validation of the annotation results. Such a meta-metabolome DB could also be queried from other perspectives, and, for example, statistical correlation with bioassay results might highlight the occurrence of a common metabolite or the presence of given NPs classes responsible for a specific bioactivity. This DB should be designed as a public DB instead of being developed for local use only. As reported very recently, the dissemination of original spectral data [49•] and standardized reporting of spectral information [50] should increase the reproducibility of chemical research and even lead to the easier curation of possible errors that might have been stored in DBs in the past [51]. This model is the one followed by the GNPS initiative [22••] and seems to be the ideal starting point for organizing and sharing MS data of small metabolites.

As shown, the technical and bioinformatics developments should allow to easily gather and share knowledge on natural metabolomes chemodiversity. Nevertheless, this public data sharing also raises ethical and legal issues in line with the recently implemented Nagoya Protocol regarding access to genetic resources and the fair and equitable sharing of benefits [52]. Common reflection on these prospects and issues should be performed within the NP research community.

Acknowledgement

JLW is thankful to the Swiss National Science Foundation for the support of metabolomics projects (SNF grants 310030E-164289 and 9 316030_164095).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Seger C, Sturm S, Stuppner H: **Mass spectrometry and NMR spectroscopy: modern high-end detectors for high resolution separation techniques—state of the art in natural product HPLC-MS, HPLC-NMR, and CE-MS hyphenations**. *Nat Prod Rep* 2013, **30**:970-987.

2. Xian F, Hendrickson CL, Marshall AG: **High resolution mass spectrometry**. *Anal Chem* 2012, **84**:708-719.

3. Rathahao-Paris E, Alves S, Junot C, Tabet J-C: **High resolution mass spectrometry for structural identification of metabolites in metabolomics**. *Metabolomics* 2016, **12**.

4. 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.

5. da Silva RR, Dorrestein PC, Quinn RA: **Illuminating the dark matter in metabolomics**. *Proc Natl Acad Sci U S A* 2015, **112**:12549-12550.

6. Meusel M, Hufsky F, Panter F, Krug D, Muller R, Bocker S: **Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns**. *Anal Chem* 2016, **88**:7556-7566.

7. Randazzo GM, 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.

8. Hufsky F, Bocker S: **Mining molecular structure databases: identification of small molecules based on fragmentation mass spectrometry data**. *Mass Spectrom Rev* 2016 http://dx.doi.org/10.1002/mas.21489.

9. Ridder L, van der Hooft JJ, Verhoeven S, de Vos RC, van Schaik R, Vervoort J: **Substructure-based annotation of high-resolution multistage MS(n) spectral trees**. *Rapid Commun Mass Spectrom* 2012, **26**:2461-2471.

10. Markley JL, Brüschweiler R, Edison AS, Eghbalnia HR, Powers R, Raftery D, Wishart DS: **The future of NMR-based metabolomics**. *Curr Opin Biotechnol* 2017, **43**:34-40.

11. Hao J, Liebeke M, Sommer U, Viant MR, Bundy JG, Ebbels TMD: **Statistical correlations between NMR spectroscopy and direct infusion FT-ICR mass spectrometry aid annotation of unknowns in metabolomics**. *Anal Chem* 2016, **88**:2583-2589.

12. Fekete S, Veuthey JL, Guillarme D: **Comparison of the most recent chromatographic approaches applied for fast and high resolution separations: theory and practice**. *J Chromatogr A* 2015, **1408**:1-14.

13. Perrenoud AGG, Guillarme D, Boccard J, Veuthey JL, 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**:101-111.

14. Contrepois K, Jiang L, Snyder M: **Optimized analytical procedures for the untargeted metabolomic profiling of human urine and plasma by combining hydrophilic interaction (HILIC) and reverse-phase liquid chromatography (RPLC)-mass spectrometry**. *Mol Cell Proteomics* 2015, **14**:1684-1695.

15. • Ortmayr K, Causon TJ, Hann S, Koellensperger G: **Increasing selectivity and coverage in LC–MS based metabolome analysis**. *Trends Anal Chem* 2016, **82**:358-366.
A review of current strategies for increasing selectivity and metabolome coverage in LC–MS based metabolomics studies.

16. • Abate-Pella D, Freund DM, Ma Y, Simon-Manso Y, Hollender J, Broeckling CD, Huhman DV, Krokhin OV, Stoll DR, Hegeman AD *et al.*: **Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods**. *J Chromatogr A* 2015, **1412**:43-51.
An inter-lab evaluation of the retention projection methodology, which could provide an interesting standardization solution for taking into consideration LC parameters for metabolite identification purposes.

17. Johnson AR, Carlson EE: **Collision-induced dissociation mass spectrometry: a powerful tool for natural product structure elucidation**. *Anal Chem* 2015, **87**:10668-10678.

18. • Mahieu NG, Spalding JL, Gelman SJ, Patti GJ: **Defining and detecting complex peak relationships in mass spectral data: the Mz.unity algorithm**. *Anal Chem* 2016, **88**:9037-9046.
An in-depth study of sources of degeneracy in MS, and an algorithm for grouping of such degenerate features.

19. Kind T, Fiehn O: **Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry**. *BMC Bioinformatics* 2007, **8**:105.

20. Li H, Cai Y, Guo Y, Chen F, Zhu ZJ: **MetDIA: targeted metabolite extraction of multiplexed MS/MS spectra generated by data-independent acquisition**. *Anal Chem* 2016, **88**:8757-8764.

21. 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 Methods* 2015, **12**:523-526.

22. •• Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T *et al.*: **Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking**. *Nat Biotechnol* 2016, **34**:828-837.
Presentation of the Global Natural Products Social Molecular Networking (GNPS) project, the first living repository of MS raw data and molecular networking platform: a game-changer for NPs chemists and researchers dealing with MS data.

23. Rojas-Chertó M, van Vliet M, Peironcely JE, van Doorn R, Kooyman M, te Beek T, van Driel MA, Hankemeier T, Reijmers T: **MetiTree: a web application to organize and process high-resolution multi-stage mass spectrometry metabolomics data**. *Bioinformatics 2012*, **28**:2707-2709.

24. • Johnson SR, Lange BM: **Open-access metabolomics databases for natural product research: present capabilities and future potential**. *Front Bioeng Biotechnol* 2015, **3**:22.
A panorama of open-access databases for both MS and NMR data in NPs research.

25. Duehrkop K, Shen H, Meusel M, Rousu J, Boecker S: **Searching molecular structure databases with tandem mass spectra using CSI:FingerID**. *Proc Natl Acad Sci U S A* 2015, **112**:12580-12585.

26. • Hufsky F, Scheubert K, Bocker S: **New kids on the block: novel informatics methods for natural product discovery**. *Nat Prod Rep* 2014, **31**:807-817.
In this review, the newcomers in the informatic toolbox of NPs chemists are presented.

27. Allen F, Greiner R, Wishart D: **Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification**. *Metabolomics* 2015, **11**:98-110.

28. 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**:7946-7958.

29. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S: **MetFrag relaunched: incorporating strategies beyond in silico fragmentation**. *J Cheminform* 2016, **8**:1-16.

30. Verdegem D, Lambrechts D, Carmeliet P, Ghesquière B: **Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization**. *Metabolomics* 2016, **12**.

31. Wang M, Bandeira N: **Spectral library generating function for assessing spectrum-spectrum match significance**. *J Proteome Res* 2013, **12**:3944-3951.

32. • Allard PM, Peresse T, Bisson J, Gindro K, Marcourt L, Pham VC, Roussi F, Litaudon M, Wolfender JL: **Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication**. *Anal Chem* 2016, **88**:3317-3323.
An integrative metabolite annotation pipeline combining molecular networks for data organization and spectral matching against an in silico fragmented DB of >170 000 NPs.

33. 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**:W94-W99.

34. Wijaya SH, Afendi FM, Batubara I, Darusman LK, Altaf-Ul-Amin M, Kanaya S: **Finding an appropriate equation to measure similarity between binary vectors: case studies on Indonesian and Japanese herbal medicines**. *BMC Bioinformatics* 2016, **17**:520.

35. • Audoin C, Cocandeau V, Thomas OP, Bruschini A, Holderith S, Genta-Jouve G: **Metabolome consistency: additional parazoanthines from the mediterranean zoanthid parazoanthus axinellae**. *Metabolites* 2014, **4**:421-432.
Principles of metabolome consistency and how in silico metabolization followed by in silico fragmentation can be applied to the annotation of previously unreported metabolites.

36. Bajusz D, Rácz A, Héberger K: **Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?** *J Cheminform* 2015, **7**:1-13.

37. Larsson J, Gottfries J, Muresan S, Backlund A: **ChemGPS-NP: tuned for navigation in biologically relevant chemical space**. *J Nat Prod* 2007, **70**:789-794.

38. Pettit RK: **Small-molecule elicitation of microbial secondary metabolites**. *Microb Biotechnol* 2011, **4**:471-478.

39. Reymond JL: **The chemical space project**. *Acc Chem Res* 2015, **48**:722-730.

40. •• Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, Hanson AD, Fiehn O, Tyo KEJ, Henry CS: **MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics**. *J Cheminform* 2015:7.
Authors propose extensions of existing structural DBs to plausible metabolites via Metabolic In silico Network Expansions (MINEs). Using a dedicated algorithm and expert-curated reaction rules, extended DBs are shown to present a high number of NP-like metabolites. Such DBs will pave the way for unreported metabolite annotation.

41. Ridder L, van der Hooft JJ, Verhoeven S, de Vos RC, Vervoort J, Bino RJ: **In silico prediction and automatic LC–MS(n) annotation of green tea metabolites in urine**. *Anal Chem* 2014, **86**:4767-4774.

42. Huan T, Tang CQ, Li RH, Shi Y, Lin GH, Li L: **MyCompoundID MS/MS Search: Metabolite Identification Using a Library of Predicted Fragment-Ion-Spectra of 383830 Possible Human Metabolites**. *Anal Chem* 2015, **87**:10619-10626.

43. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 1999, **27**:29-34.

44. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A *et al.*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2014, **42**:D459-471.

45. Van Moerkercke A, Fabris M, Pollier J, Baart GJ, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM, Goossens A: **CathaCyc, a metabolic pathway database built from Catharanthus roseus RNA-Seq data**. *Plant Cell Physiol* 2013, **54**:673-685.

46. Weber T: **In silico tools for the analysis of antibiotic biosynthetic pathways**. *Int J Med Microbiol* 2014, **304**:230-235.

47. Kleigrewe K, Almaliti J, Tian IY, Kinnel RB, Korobeynikov A, Monroe EA, Duggan BM, Di Marzo V, Sherman DH, Dorrestein PC *et al.*: **Combining mass spectrometric metabolic profiling with genomic analysis: a powerful approach for discovering natural products from cyanobacteria**. *J Nat Prod* 2015, **78**:1671-1682.

48. Moss NA, Bertin MJ, Kleigrewe K, Leao TF, Gerwick L, Gerwick WH: **Integrating mass spectrometry and genomics for cyanobacterial metabolite discovery**. *J Ind Microbiol Biotechnol* 2016, **43**:313-324.

49. • Bisson J, Simmler C, Chen S-N, Friesen JB, Lankin DC, McAlpine JB, Pauli GF: **Dissemination of original NMR data enhances reproducibility and integrity in chemical research**. *Nat Prod Rep* 2016, **33**:1028-1033.
An advocacy for dissemination and publication of original NMR raw data (FID). Application of such guidelines for NMR as well as MS data should be largely beneficial for NP research.

50. Wohlgemuth G, Mehta SS, Mejia RF, Neumann S, Pedrosa D, Pluskal T, Schymanski EL, Willighagen EL, Wilson M, Wishart DS *et al.*: **SPLASH, a hashed identifier for mass spectra**. *Nat Biotechnol* 2016, **34**:1099-1101.

51. Reddy DS, Kutateladze AG: **Structure revision of an acorane sesquiterpene cordycepol A**. *Org Lett* 2016, **18**:4860-4863.

52. Kamau EC, Fedder B, Winter G: **The Nagoya Protocol on access to genetic resources and benefit sharing: what is new and what are the implications for provider and user countries and the scientific community?** *Law Dev J (LEAD)* 2010, **6**:248-263.