Computational and Structural Biotechnology Journal 19 (2021) 72–85

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

Seeing the forest for the trees: Retrieving plant secondary biochemical pathways from metabolome networks

Sandrien Desmet [a,b,1], Marlies Brouckaert [a,b,1], Wout Boerjan [a,b,2,*], Kris Morreel [a,b,2,*]

[a] Ghent University, Department of Plant Biotechnology and Bioinformatics, Ghent, Belgium
[b] VIB Center for Plant Systems Biology, Ghent, Belgium

ARTICLE INFO

Article history:
Received 4 September 2020
Received in revised form 26 November 2020
Accepted 28 November 2020
Available online 03 December 2020

ABSTRACT

Over the last decade, a giant leap forward has been made in resolving the main bottleneck in metabolomics, i.e., the structural characterization of the many unknowns. This has led to the next challenge in this research field: retrieving biochemical pathway information from the various types of networks that can be constructed from metabolome data. Searching putative biochemical pathways, referred to as biotransformation paths, is complicated because several flaws occur during the construction of metabolome networks. Multiple network analysis tools have been developed to deal with these flaws, while *in silico* retrosynthesis is appearing as an alternative approach. In this review, the different types of metabolome networks, their flaws, and the various tools to trace these biotransformation paths are discussed.

1. Introduction

At the start of this millennium, biochemical research was marked by the retrieval of the Arabidopsis [1] and human full genome sequences [2,3], and by the development of metabolomics [4–6] as a third cornerstone next to transcriptomics and proteomics, to support functional genomics. Functional genomics was expected to enable the high-throughput functional annotation of the many unknown genes that were obtained via genome sequencing. However, halfway 2016, the molecular function of 47% of the Arabidopsis genes was still unknown owing mainly to the often excessive time necessary to obtain experimental proof for the function of a gene (https://www.arabidopsis.org/portals/genAnnotation/genome_snapshot.jsp).

In parallel to unraveling gene functions, the omics technologies brought hope to finally understand the system-wide organization and regulation of the metabolism in living systems. Admittedly, systems biology entered full maturation at the moment that biological networks, such as transcriptional regulatory, protein–protein interaction, and metabolic networks, could be constructed from large amounts of omics data [7–9]. These large-scale networks offered a chance to understand how the molecular basis of life is governed. Accordingly, network analysis has provided insight into frequently occurring node connectivity patterns or motifs in biological networks [10], and revealed how, e.g., the control of gene regulation in living systems is designed [11]. As compared to other biological networks, the information displayed in metabolic networks is far less complete. The two main reasons for this lack of knowledge are enzyme promiscuity, and the many unknown pathways in secondary, also called specialized, metabolism. Enzyme promiscuity, occurring when enzymes recognize multiple substrates and/or catalyze multiple reactions [12–14], could be attributed to 37% of all metabolic enzymes in *Escherichia coli*, and a similar proportion has been observed in other prokaryotic and eukaryotic microorganisms [15]. Concerning the second reason, as opposed to primary metabolism, which is common to all organisms, secondary metabolism is not required for general growth, development, and reproduction, but provides the host with distinct fitness benefits in a specific ecosystem, hence, the specialized metabolic pathways often vary substantially among species. This increases dramatically the number of pathways that have to be unraveled, explaining the slow progress in pathway elucidation and the functional annotation of the responsible genes.

1.1. Metabolic networks provide a basis to study the control and regulation of metabolism

Before attempting to study prevailing patterns in a metabolic network, a network that accurately reflects metabolism has to be assembled. For the construction of a metabolic network, informa-

* Corresponding authors at: VIB-UGent Center for Plant Systems Biology, Technologiepark 71, 9052 Ghent, Belgium.

E-mail addresses: wout.boerjan@psb.vib-ugent.be (W. Boerjan), kris.morreel@psb.vib-ugent.be (K. Morreel).

[1] Shared first authors.

[2] Shared last authors.

https://doi.org/10.1016/j.csbj.2020.11.050
2001-0370/

tion can be readily gathered from pathway databases (Fig. 1A), such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16–18], BioCyc [19], the Plant Reactome database [20,21] and the Gramene database [22], yet these databases might be incomplete or insufficiently detailed. When attempting to complete the metabolic network of a particular species, gaps representing unknown reactions have to be filled in, which demands an exhaustive search for all gene–protein–reaction associations by concatenating gene–protein and protein–reaction information from different databases. Subsequently, a final curation via constraint-based modeling (CBM, see Glossary) [23–25] yields a so-called genome-scale metabolic model (GSMM) [26–28]. Metabolic networks are mainly displayed either as a homogenous network (nodes and edges reflecting metabolites and reactions) or as a bipartite graph characterized by two types of nodes (representing metabolites and enzymes), in which the edges represent links between substrates/products and their respective enzymes. Regardless of the layout, metabolic networks typically contain many poorly connected nodes interconnected by a few heavily connected nodes (the hubs), the latter being especially associated with cofactors such as ATP, NADH, glutamate and coenzyme A [29]. Consequently, the node connectivity, defined as the number of edges per node, shows a heavy-tailed probability distribution [30–36]. Furthermore, metabolic networks have a non-random topology and are likely organized in a hierarchical modular structure (Fig. 1A) [37], in which network modules (see Glossary) that sometimes represent particular biochemical pathways, are nested.

Compared to transcriptional regulatory networks, the search for motifs in metabolic networks is more slowly pursued, in part because the relationship between a metabolic network motif and metabolic regulation or control (see Glossary for the difference between metabolic control and regulation) is not always clear [38]. Nevertheless, targeted searches for well-known motifs have been performed, for example, to exhaustibly enumerate all substrate cycles [39]. The same authors also screened for feedback inhibition loops following enrichment of the metabolic networks with metabolite–enzyme regulatory (inhibition/activation) interactions [40]. The latter type of analysis can only be efficiently performed when (i) all such interactions are known, and (ii) the metabolic network adequately reflects the biochemical pathway architecture [41]. Both prerequisites are also important when searching for network motifs concerning the transcriptional regulation of biochemical pathways, which necessitates combining metabolic networks with transcriptional regulatory networks [42–44]. Obviously, the more the metabolic network accurately reflects metabolism, the better it is suited, in combination with other types of biological networks, to search for metabolic control/regulatory motifs.

Nodes representing primary metabolites were shown to be connected via multiple network paths [45]. In part, this is due to the many cycles (e.g., Krebs cycle, pentose phosphate pathway, etc.) operating in primary metabolism. Hence, blocking a particular pathway in primary metabolism is not necessarily associated with the loss of the downstream metabolites. This explains, at least partially, why quantitative trait locus (QTL) analyses of the concentrations of primary metabolites in natural or mapping populations yield associations with many loci [46–48]. Furthermore, these QTL data also point to the shared control of metabolic fluxes by all reaction steps as opposed to control exerted by a single rate-limiting step [49,50]. The fact that the levels of primary metabolites are controlled by many steps explains why significant correlations between the abundances of any pair of primary metabolites across biological replicates, i.e., abundance-based correlations, are rarely encountered [51]. Positive abundance-based correlations necessitate that (i) most of the control of the levels of both metabolites occurs via the same set of reactions, and, assuming this is indeed the case, that (ii) each reaction affects the covariation between the levels of both metabolites in the same direction [52]. These are rather stringent conditions that are not expected to readily occur. Therefore, the few positive correlations that have been reported in primary metabolism likely arise when (i) metabolite levels covary because a single step controls most of the flux to both metabolites (asymmetric control), (ii) the rate of one step is varying over a much larger range than any of the other steps (outlying variation), or (iii) both metabolites are in chemical equilibrium [53]. Whereas a chemical equilibrium can only explain significant correlations between biochemically proximal metabolites, e.g., between glucose-6-phosphate and fructose-6-phosphate, asymmetric control and outlying variation also explain the observed correlations between biochemically distant metabolites in primary metabolism [54].

Although negative abundance-based correlations are less likely to be observed compared to positive abundance-based correlations, they have been reported in primary metabolism [51]. They arise whenever mass conservation exists between two metabolites, for example when they are part of a moiety-conserved cycle (e.g., a reaction in which NADH is oxidized will be coupled to other reactions that reduce NAD^+, leading to a negative correlation between NADH and NAD^+ levels) [52]. From this perspective, searching motifs representing moiety-conserved cycles in metabolic networks that are supplemented with abundance-based correlations using metabolome data would be a powerful approach to gain insight into the importance of moiety-conserved cycles in metabolic control/regulation. The same approach also allows to search for groups of nodes that are mutually showing high positive abundance-based correlations in metabolic networks. This approach enables pointing to regions in the network in which the flux is mainly affected by one particular enzymatic reaction. Likely, such regions are more prevalent in secondary metabolism.

In contrast to the many QTLs typically obtained for the abundance of primary metabolites, the levels of plant-derived secondary metabolites are often associated with only few loci [46,47,55]. Sometimes, enzyme-encoding loci can be functionally annotated based on the absence of a particular secondary metabolite in a subset of the population [56,57]. Furthermore, the guilt-by-association principle, in which correlations are sought between metabolite accumulation and transcript expression profiles, has been especially successful in the study of plant secondary metabolism [58–60]. Consequently, it seems that secondary metabolic pathways are much more subjected to transcriptional control or affected by rate-limiting steps than primary metabolic pathways. Asymmetric control would be expected in case of rate-limiting steps affecting the levels of different metabolites, leading to positive abundance-based correlations between biochemically related metabolites. In agreement, in a study of the aromatic metabolism in the rosette leaves of *Arabidopsis thaliana*, Morreel et al. (2014) [61] noticed a positive relationship between the abundance-based correlation coefficient and the mass spectrometry (MS) fragmentation spectral similarity. As Shen et al. (2019) [62] observed that similar MS fragmentation spectra especially appear between neighboring metabolites in a biochemical pathway, the study of Morreel et al. (2014) [61] suggests that biochemically proximal metabolites might more readily show a positive abundance-based correlation in secondary metabolism than in primary metabolism. Gaquerel et al. (2013) [63] also noticed that more positive correlations are observed between the abundances of compounds belonging to the same secondary biochemical class than between compounds belonging to different classes. However, more studies are necessary to understand the distribution of abundance-based correlations across secondary metabolic networks, and how this distribution differs from that of abundance-based correlations in primary metabolic networks. Nevertheless, these few observations

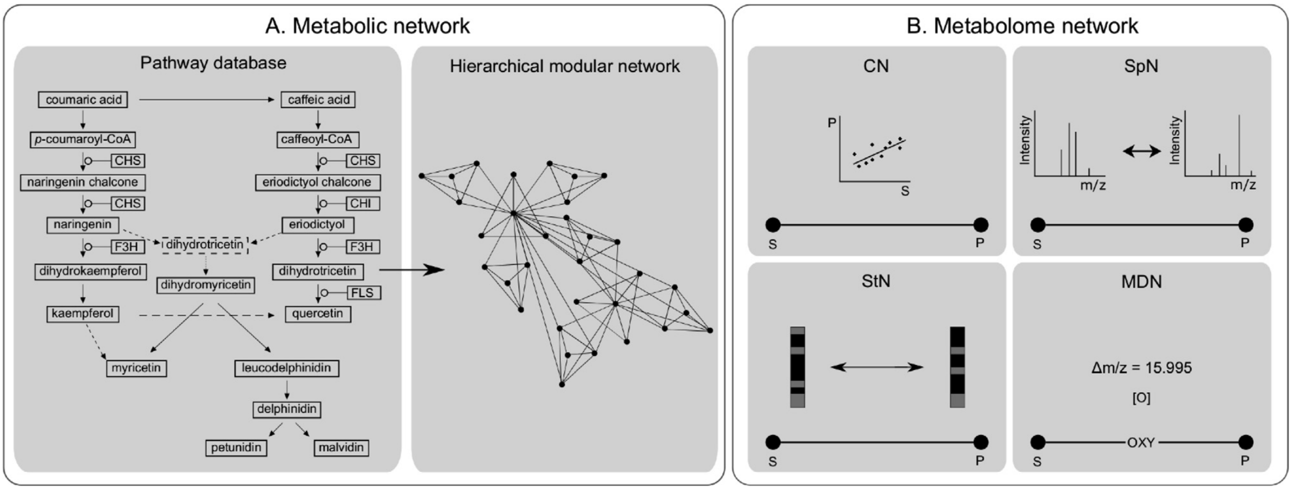

Fig. 1. Metabolic versus metabolome network. (A) Scheme reflecting the construction of a homogeneous metabolic network (nodes and edges reflecting metabolites and enzymatic conversions) using data from the KEGG database. The metabolic network displays a hierarchical modular structure (adapted from Ravasz et al. (2002) [37]). (B) A metabolome network is constructed from metabolome data (nodes represent features). Dependent on whether edges reflect a correlation, a mass difference, a spectral similarity, or a structural similarity (using a binary vector of molecular descriptors for each structure), the network is referred to as a correlation network (CN), a mass difference network (MDN), a spectral similarity network (SpN), or a structural similarity network (StN). OXY, oxygenation; P, product; S, substrate.

indicate the importance of studying, besides the global metabolic network, the differences between primary and secondary metabolic networks.

Many differences between primary and secondary metabolism can be explained at a transcriptional regulation level, but arise as well from differences in the network topology of primary and secondary metabolism, such as the presumed higher frequency of cycles and branching in primary metabolism as compared to secondary metabolism. However, a GSMM approach will fail to elucidate the differences between primary and secondary metabolic networks, because insufficient pathway knowledge is available concerning the secondary metabolism of most species. The main source of information on secondary metabolic networks is generated by the comprehensive profiling of the secondary metabolome via liquid chromatography (LC) or capillary gas chromatography (CGC) coupled to MS. Such profiling data yields tens of thousands of features (see Glossary), annotated by a retention time and a mass-to-charge (m/z) value, that represent the metabolites.

1.2. What type of information is gained via MS-based metabolomics?

Following data processing, GC- and LC-MS platforms yield a list of features. These features represent ions that are usually generated via (de)protonation of the metabolite to yield the negative or positive pseudo-molecular ion of the metabolite (mainly in LC-MS), or by electron loss to yield the radical cation of the metabolite (mainly in GC–MS). Below, the pseudo-molecular ion or radical cation are collectively referred to as the precursor ion. In addition to the feature representing the precursor ion, each profiled metabolite is associated with features representing (i) natural isotopes, (ii) potential adducts, (iii) fragment ions produced inside the ionization source (in-source fragmentation; ISF) [64], and (iv) in-source reactions [65]. From an MS perspective, all features (their m/z values and their intensities) belonging to the same metabolite can be viewed in an MS^1 spectrum. Below, the terms features and ions are interchangeably used.

In GC–MS-based metabolomics, extensive fragmentation in the ionization source occurs upon electron ionization (EI), yielding very reproducible ISF spectra that can be matched against large EI spectral databases such as the Wiley Registry of Mass Spectral Data [66], which currently contains EI spectra of almost 700,000 compounds. In LC-MS-based metabolomics, the fragmentation spectrum is generated in the MS analyzer rather than in the ionization source, mainly via low energy collision-induced dissociation (CID). CID leads to either MS^n spectra or an MS/MS spectrum when profiling occurs using, e.g., an ion trap (IT) or a quadrupole-time-of-flight (Q-TOF) MS. In MS^n fragmentation, the generated MS^2 product ions can each be further fragmented to MS^3 second order product ions, which can themselves be subjected to MS^4 fragmentation, and so on. Therefore, the MS^n spectra of an ionized compound consist of an MS^2 spectrum and, optionally, several MS^3, MS^4, . . . spectra. This permits the construction of MS^n spectral trees that reflect the subsequent fragmentations of the ionized compound in the gas phase [67–69]. In contrast with MS^n fragmentation, MS/MS fragmentation yields more product ions, including low-mass product ions. Independently of the type, i.e., MS^n or MS/MS, the CID spectrum depends on the instrument voltage and temperature settings, as well as the instrument configuration, and, thus, is often not reproducible between labs. Consequently, as compared to the construction of EI spectral databases, the construction of CID spectral databases, such as MassBank [70] or mzCloud (HighChem, Ltd. Bratislava, Slovakia), have been lagging behind. CID spectral databases have been developed relatively recently, i.e., during the last decade, but are only slowly increasing in numbers [71].

From the GC- or LC-MS profiles, biochemical insights can be gained by the construction of a mass difference network (MDN), a spectral similarity network (SpN), a structural similarity network (StN), or an abundance-based correlation network (CN) in which the features are represented as nodes. None of these networks (generally referred to as metabolome networks; Fig. 1B) accurately display the metabolic network (representing the total of all known biochemical pathways), yet they provide a starting point for further curation. Below, we describe the current status in generating biochemical pathway information from metabolome data.

2. The metabolome network as a metabolic network surrogate

The first networks constructed from metabolome data were CNs [51] (Fig. 1B). In a CN, the nodes are features that are connected by edges whenever their abundance-based correlation is high. Assuming a metabolic steady state, these abundance-based correlations arise from concentration fluctuations that propagate through metabolism because of the interplay between a continuously

Table 1
Software/algorithms for metabolome network construction.

Software/algorithm^2	Type	Included procedures	Reference
MetaNetter	MDN		[86]
	CN	Pathway database mapping Compound-based feature grouping	[101]
MI-Pack	MDN^1		[88]
Metscape 2	MDN^1	Addition of non-detected intermediates	[198]
	MDN	Atom mass differences	[199]
GNPS	MN		[81]
MetaMapp	MDN^1	StN and MN networks	[84]
mummichog	MDN^1	Differential feature clustering onto network modules	[150]
mzGroupAnalyzer	MDN		[200]
CSPP	MDN	Retention time order CID spectral similarity Abundance-based correlation	[61]
MetaMapR	MDN^1	StNs, SpNs and MNs	[151]
	MN	*in silico* CID spectral database	[201]
PIUMet	MDN^1	Differential feature clustering onto network modules	[149]
BioCAn	MDN^1	Differential feature clustering onto network modules + neighboring node connectivity CID spectral matching *in silico* spectral elucidation	[148]
MetaNetter 2	MDN	Compound-based feature annotation	[102]
	MDN	Pathway database mapping (KEGG)	[202]
MetCirc	MN		[80]
NAP	MN	*in silico* CID spectral elucidation	[112]
MetGem	MN		[203]
MolNetEnhancer	MN	*in silico* CID spectral elucidation (several tools)	[113]
	MN	Gas phase fragmentation rules	[114]
MetWork	MN	*in silico* CID spectral elucidation Biotransformations	[204]
MetNet	MDN	Retention time order Multiple association statistics	[205]
COBRA Toolbox v.3.0	MDN^3	Multi-omics data	[176]
REMI	MDN^3	Multi-omics data	[173]
MetaBridge	MDN^3	Multi-omics data	[177]

Due to size restrictions, a restricted number of software/algorithms is given. [1] Nodes in the MDN are profiled compounds that can be retrieved, together with their enzymatic reactions which represent the MDN edges, from pathway databases. [2] A name is lacking for some algorithms. [3] Rather than an MDN, these software/algorithms employ a GSMM which can sometimes be manually curated.

changing environment, the pathway architecture and the complex regulation of metabolism [54]. Therefore, the covariations between metabolite concentrations enable, at least in theory and when the metabolic network is fully known, to compute all elasticities of the various reactions [72]. As defined in metabolic control analyses [49,50], the elasticity refers to the change in reaction rate upon a small change in the concentration of a metabolite; in case of a substrate, the elasticity represents a system-wide version of the Michaelis-Menten constant. Hence, CNs contain information of the physiological state [73]. Comparing CNs might enable, e.g., to point to reactions that are considerably changed between the two experimental conditions [74]. Various methods for the differential analysis of CNs have been developed [75–77].

In order to unravel metabolic pathways, the question arises to what extent CNs display the pathway architecture. While metabolites that are in chemical equilibrium will yield “correlated” features, many neighboring pathway intermediates do not show high abundance-based correlations. Furthermore, significant correlations can also be generated between biochemically distant metabolites, at least in case the CN is based on Pearson or Spearman correlation coefficient computations [52–54]. Therefore, little pathway information can be retrieved from such a CN. However, CNs that are created using partial correlation coefficients, so-called Gaussian Graphical Models (GMMs), were argued to reflect the biochemical pathway architecture much better [78].

The main goal of SpNs is to facilitate structural characterization by grouping features representing metabolites for which similar MS fragmentation spectra were recorded [79–81] (Fig. 1B). Assuming that CID spectral similarity reflects structural similarity, all CID spectra within an experiment are mutually compared, an edge is drawn whenever the spectral similarity surpasses a preset threshold, and the mass difference between the two involved features is computed. Structural characterization can then be performed via network propagation; starting from a node representing a known compound (e.g., via spectral library matching), the structure associated with an adjacent node can be elucidated taking the spectral similarity into account. This procedure has been implemented in the Global Natural Products Social Molecular Networking (GNPS) database (Table 1); here, the SpN is referred to as a molecular network (MN) [79,81]. Besides computing the MN, the GNPS database allows to match CID spectra against existing CID spectral libraries, and to curate or elucidate CID spectra that are already present in the database. The on-going curation of the GNPS database is referred to as ‘living data’ by the GNPS developers [82]. Clearly, many of the edges in an SpN will represent concatenations of multiple enzymatic reactions.

StNs are the most recently defined of all types of metabolome networks [83,84] (Fig. 1B). They are constructed using those features for which a structure could be retrieved from a compound or pathway database. Structural similarity computations are not based on the molecular structures themselves, but are obtained following the ‘translation’ of each molecular structure into a set of molecular descriptors. Different sets of molecular descriptors are publicly available, such as the ‘fingerprint’ set of the PubChem compound database, which contains almost 900 descriptors. Each of the molecular descriptors within a set represents a chemical substructure. Consequently, the compound’s structure is represented by a molecular descriptor vector of binary elements, in which each element represents the absence/presence of a particular chemical substructure. The structural similarity between two compounds, represented by, e.g., the Tanimoto coefficient, is then computed using their corresponding molecular descriptor vectors.

Structural characterization can also be aided via the construction of MDNs, assuming that the profiled compounds are rather closely biochemically related. In this approach, pairs of features are searched, of which the mass differences correspond to those of well-known enzymatic reactions, e.g., 14.015 Da in case of a methylation (Fig. 1B). Consequently, MDN edges represent mass differences and, thus, putative biotransformations (see Glossary). Structural characterization proceeds via network propagation taking the mass difference, i.e., the biochemical conversion, into account. Such MDNs employ MS^1 level information, and were first applied on direct infusion-MS profiles [85,86]. By mapping the mass differences to those present in pathway databases such as KEGG, a more accurate computation of the chemical formula of the features was possible [87,88]. Using reversed phase LC-MS profiling, this approach has been elaborated by including the elution order between the candidate substrate and product features of the considered biotransformation as, e.g., the candidate product of a methylation is expected to elute later than the candidate substrate. Hence, as retention time information is also included, such MDNs are rather referred to as Candidate Substrate Product Pair (CSPP) networks [61] (Table 1). Additional support for the biochemical validity of a particular CSPP might be obtained by computing for the candidate substrate and product features, (i) the abundance-based correlation, and (ii) their CID spectral similarity whenever CID spectra are available [61]. CSPP networks have been useful to gain insight into cellular [89] and sub-cellular metabolism [90], and to annotate the *in vivo* activity of unknown enzymes [91]. Despite the apparent redundancy of SpN and MDN approaches for structural characterization, both approaches are partially complementary: whereas SpNs may reveal unexpected biochemical conversions, MDNs will display structurally similar compounds even when their CID spectra are dissimilar. While sometimes not realized, it has indeed been shown that spectral similarity implies structural similarity, but not necessarily vice versa (e.g., see structural versus spectral similarity plots in Rasche et al. (2012) [92] and Rojas-Cherto et al. (2012) [93]) (left and middle spectrum in Fig. 2).

Despite the help offered by metabolome networks in understanding, e.g., metabolic changes upon perturbations, they cannot be analyzed via the same systems biology approaches that are applied to metabolic networks. Several flaws prevent turning them into an approximation of metabolic networks: (i) metabolome networks are highly redundant due to the presence of multiple features representing each compound, (ii) being the most important bias, most of the network nodes are not annotated because of the large number of unknown compounds in the secondary metabolism, and (iii) no one-to-one relationships, i.e., a bijection, exist between biotransformations and enzymatic reactions.

2.1. Flaw 1: metabolome networks are highly redundant

Ideally, only the feature representing the precursor ion of each metabolite should be included in the metabolome network by discarding all other redundant features. This would need MS^1 spectrum construction as the final data processing stage, implying the grouping of all features at a particular retention time that are associated with the same compound, yet preventing the inclusion of features associated with co-eluting compounds. Currently, several programs are available for this purpose [94–99]. Following the construction of the MS^1 spectrum, the m/z value of the precursor ion is annotated by taking the m/z values of the adduct ions into account [100]. However, (i) precursor ion annotation is sometimes difficult when adduct ions are absent, (ii) some compounds are detected as adduct ions rather than as precursor ions, and (iii) the ion type corresponding to some MS^1 features cannot be annotated, hence, questioning whether the feature is not rather representing a co-eluting compound. Therefore, although allowing the construction of a less redundant metabolome network, prior MS^1 spectral filtering will inevitably provide a biased network. A possible solution exists in pinpointing or clustering, within the metabolome network, features belonging to the same MS^1 spectrum. Because an abundance-based correlation is often implemented in the MS^1 spectrum construction procedure, both Gipson et al. (2008) [101] and Gaquerel et al. (2013) [63] applied a further clustering of MS^1 features following CN construction. To deal with this problem in MDN construction, two types of edges have been implemented in MetaNetter 2 (Table 1) that address either a biotransformation or an MS^1-based adduct formation [102]. Whereas the CN-based approach allows pinpointing all features belonging to the same MS^1 spectrum, the MDN-based approach enables annotating the ion type of the MS^1 spectral features whenever possible. Optimally, the benefits of both methods are combined.

2.2. Flaw 2: poor node annotation

The low number of known metabolites in plant secondary metabolism is an immediate consequence of the labor-intensive purification or chemical synthesis of the compound that is necessary to identify a compound by nuclear magnetic resonance (NMR) analysis or via spiking experiments. To date, this bottleneck remains unsolved as researchers are increasingly realizing the complexity of the plant metabolome, with predictions of up to 1 million metabolites being present in the plant kingdom [103,104]. Contrasting with the size of the transcriptome or proteome of a species, the metabolome size cannot be predicted from genome information. Many enzymes are promiscuous; even enzymes from central metabolic pathways, such as glycolysis, cannot avoid side reactions [105]. This is further complicated by the *in planta* chemical degradation of metabolites [106].

As compared to spiking- or NMR-based identification, a more high-throughput structural characterization pipeline employs the metabolome data itself. Typically, this starts with computing the chemical formula of the precursor ions [88,107–110]. When available, spectroscopic data, e.g., UV/VIS absorption spectra, and/or the ion mobility-based drift time can be consulted, yet most information on the structural moieties is obtained by analyzing the CID or EI spectrum. Except for a limited number of compounds that can be structurally annotated by matching against spectral databases, *de novo* spectral elucidation will have to be attempted for most of the unknown compounds. From a spectral matching perspective, especially CID spectra will need *de novo* spectral elucidation (see '*1.2 What type of information is gained via MS-based metabolomics?*'). Tools for the latter purpose have been developed [111], and subsequently combined with MNs to facilitate network propagation [112–114].

Most *de novo* spectral elucidation tools need candidate structures for which the CID spectra can be predicted and compared to that of the unknown compound. Such candidate structures can be retrieved from a compound database (e.g., the PubChem database containing currently over 110 million compounds) using the mass or the chemical formula of the unknown compound [115]. However, the number of plant metabolites represents only a small fraction of any large compound database, as most entries will be non-biological compounds such as, e.g., drugs. Consequently, only a minority of the profiled compounds can be traced back in these compound databases. This bottleneck has been tackled by creating *in silico* compound databases, i.e., by subjecting the metabolites *in silico* to well-known biotransformations, such as methylation and glucosylation among others [116]. With each applied biotransformation, this approach readily leads to an exponentially increasing number of *in silico* generated compounds, hence, some programs only perform the *in silico* biotransformations on a selected set of

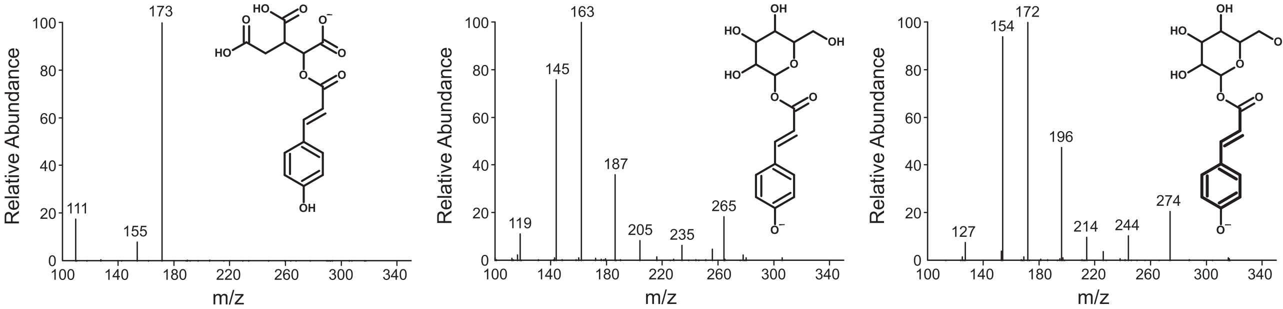

Fig. 2. Negative ionization CID spectra. MS^2 spectra recorded on an IT-MS of two *p*-coumaroyl esters: *p*-coumaroyl isocitrate (left, precursor ion at *m*/*z* 337.06) and *p*-coumaroyl glucose (middle, *m*/*z* 325.09). On the right, the MS^2 spectrum is displayed of [$^{13}C_9$]-labeled *p*-coumaroyl glucose (*m*/*z* 334.12) generated during a time course feeding experiment with [$^{13}C_9^{15}N_1$]-phenylalanine.

metabolites, of which the core structures are highly likely to occur in the generated metabolite profiles [117–121].

2.2.1. Combining analytical chemical approaches for structural characterization via data fusion

Via data fusion, the complementary information present in different sources of spectroscopic data can be integrated together. This process involves either online- or offline-coupled instrumentation. Examples of the former are the various combinations of UV/VIS absorption, infrared (IR) absorption and fluorescence spectroscopy, and MS coupled to LC [122]. However, except for MS, all these detectors record spectra of which the detection response at each wavelength results from contributions of all co-eluting compounds. Disentangling the spectra of the individual compounds from these composite spectra has been a main objective in the development of chemometry [123]. To avoid the often time-consuming (and sometimes biased) chemometry-associated machine learning approaches, IR spectroscopy has been recently assembled onto IT-MS; upon isolation of a particular ion in the mass analyzer, its IR absorption and CID spectrum can then be recorded almost simultaneously [124]. As an alternative approach to the online detector coupling or the use of machine learning-based data fusion, the redundant information present in spectra can be used to fuse data from different instrument platforms. For example, in a CSPP-like approach, high-resolution MS data has been fused to LC low-resolution MS data, hence, combining the accurate *m*/*z* values recorded on the former instrument with the retention times obtained on the latter platform [125].

NMR spectroscopy has also been online hyphenated to LC-MS. In the presence of a magnetic field, NMR measures the differences between the ^{1}H nuclei of a compound with respect to the precession frequencies of their magnetic moment vectors; these differences are expressed in chemical shift values. Compared to MS, NMR has a low sensitivity, compromising the coupling of NMR to LC-MS [126]. Because of the issues concerning online LC-MS-NMR, offline coupling of NMR to LC-MS has more frequently been pursued. By combining the offline data from NMR and MS, structural elucidation has been successful even when spectra were acquired on a mixture of compounds [127]. Some of these approaches could be called ‘supervised’ NMR as the NMR spectrum is inspected for the expected peak combination that supports the MS-based structural characterization [120].

2.2.2. Data fusion using retention time alignment: spectral metadata analysis

The above-described data fusion via an offline approach is cumbersome and/or sensitive to the mismatching of different types of spectra, but can be improved when the different spectroscopic analyses are preceded by the same chromatographic method, assuming that the retention time of a particular compound hardly differs between the different instrument platforms. Data fusion then relies on chromatographic alignment. Among the more straightforward approaches is the data fusion of MS/MS and MS^n spectra, which we here refer to as spectral metadata analysis. Both of these CID spectral types yield complementary information but are sufficiently redundant to aid the chromatographic alignment, i.e., by associating the same compound in both chromatograms using both retention time correspondence and CID spectral matching. Such an MS spectral metadata analysis allows also to consult multiple spectral databases and to implement multiple *de novo* spectral elucidation tools for structural characterization, because these databases and tools are sometimes rather dedicated to a particular type of CID spectrum. These advantages of combining different types of CID spectra have led to, e.g., the construction of the widely used IT-orbitrap MS analyzers [128], which record MS/MS-like higher-energy collision dissociation (HCD) spectra, as well as MS^n spectra. However, even when performing metabolite profiling on such instruments, spectral metadata analysis is still necessary whenever CID spectra have to be combined that were recorded using different instrument settings (e.g., positive or negative ionization, or involving metal complexation), or using different wet lab conditions (e.g., when performing stable isotope-labeled (SIL) precursor feeding, or applying post-column derivatization).

Sometimes more detailed structural information can be derived from CID spectra using particular instrument settings. For example, the default metabolomics-based LC-MS procedures often cannot handle the precise linkage position between substructures, such as the linkage position of an aglycone to its sugar in case of glycosylation. Such linkage positions can often be determined by considering the CID spectrum of either the metal ion complexes of the glycosides [129] or of the glycoside anions [130–132]; both types of CID spectra sometimes yield complementary information. Furthermore, following elucidation of the linkage position of the aglycone onto the sugar, the stereoisomeric configuration still has to be determined. Such information might be derived via MS^n analysis [133]. Thus, both in-depth MS^n analysis and dedicated CID spectral settings aid in getting sufficient information to resolve the structures of unknown compounds.

Alternatively, to get sufficient CID spectral information, SIL precursor feeding or post-column derivatization can be performed. Upon feeding plants with a SIL biochemical precursor followed by an LC-MS analysis, the CID spectra of the isotopologs (see Glossary) of any unknown compound can be visualized together for structural characterization (compare middle and right spectrum in Fig. 2) [134]: any CID product ion representing a substructure that is derived from the SIL biochemical precursor will appear at a different *m*/*z* value in the CID spectra of the isotopologs. Additionally, isotopologs of an analyte can also be generated via hydrogen– deuterium exchange (HDX) [135]. The observed mass increments of precursor and product ions due to HDX aid in locating functional groups having exchangeable protons [136].

Structural characterization can also be facilitated by including information obtained following derivatization of the metabolite. To permit alignment of LC-MS data generated from the underivatized and the derivatized analytes, post-column derivatization is necessary. Post-column derivatization might involve mixing the eluent with a derivatization reagent before it enters the MS ionization source. For example, mixing the eluent with acetone under UV light converts double bonds to oxetane moieties [137]. Following post-column derivatization, the resulting mass increments of the precursor and product ions facilitate the localization of double bonds. Post-column derivatization can also be pursued in-source by infusing the derivatization reagent directly into the MS ionization source [138].

2.3. Flaw 3: biotransformations do not adequately reflect enzymatic reactions

When inspecting metabolome networks, it is often observed that two nodes that are connected by an edge do not truly correspond to the substrate and the product of an enzymatic reaction. This anomaly occurs most frequently in CNs and SpNs, in which edges are defined based on abundance-based correlation or CID spectral similarity coefficients, and where mass difference computation, if performed, only occurs as a second step. Therefore, an MDN that is built using only the features representing the precursor ions is most suited to search for biochemical pathways. Nevertheless, information from CNs and SpNs can be helpful in deriving an MDN that reflects the metabolism. In an MDN, four potential causes for this anomaly have been published (Fig. 3) [61]: (i) biotransformations that reflect a linear sequence of enzymatic reactions (referred to as 'reaction combinations', Fig. 3A), sometimes occurring because the pathway intermediates are not detected in the metabolome data; (ii) biotransformations that can be explained by different combinations of enzymatic reactions ('multiple reaction paths', Fig. 3B); (iii) biotransformations that reflect true enzymatic reactions, yet some of the corresponding edges connect the derivatized forms (e.g., glucosylated derivatives) rather than the *in vivo* substrate and product ('reaction displacements', Fig. 3C); and (iv) biotransformations of which some of the corresponding edges end up in a node corresponding to a structural isomer of the expected product ('isomer displacements', Fig. 3D). The latter can only be properly addressed via structural characterization (see *1.2. What type of information is gained via MS-based metabolomics?*') or, at least in part, by considering the MS fragmentation spectral similarity.

Some of the anomalies can be partially solved using information from pathway databases such as KEGG; this approach has led to the development of several metabolome database visualization programs [88,101,139–146] (Table 1). These programs typically annotate features by matching their computed molecular weights to those of the compounds in a pathway database. Metabolomics results can then be interpreted based on the pathways in which the features are mapped (referred to as metabolite mapping) [147]. These pathway visualization strategies are especially interesting when analyzing primary metabolism, because primary metabolites are well-covered in pathway databases. However, each feature will be mapped to all of its corresponding isomers in the pathway database. To select the correct isomer, comparative metabolomics experiments can be carried out. Here, feature annotation results from mapping the differential features mutually as close as possible onto metabolic pathways or metabolic network modules [148,149]. Methods such as *mummichog* [150] (Table 1) assume that the metabolic network generated from a GSMM shows a hierarchical modular topology (see '*1.1 Metabolic networks provide a basis to study the control and regulation of metabolism*'), and that the differential features resulting from applying a treatment or a genetic modification are biochemically related and, thus, end up in one or a few neighboring modules. Alternatively, including retention time and (predicted) CID spectral information may improve the isomer mapping. Such a strategy is implemented in MetDNA [62] (Table 1), and is based on the assumption that neighboring pathway intermediates have similar MS fragmentation spectra. The algorithm starts with feature identification via spectral matching against MS spectral databases; the identified features are referred to as 'seed metabolites'. Next, the seed metabolites are mapped onto KEGG, and KEGG compounds that are linked via a reaction to one of the seed metabolites are retrieved. The precursor m/z values for these KEGG-retrieved compounds are computed, their retention times predicted (based on a retention time prediction model constructed using the initial set of seed metabolites) and their putative MS fragmentation spectra borrowed from the seed metabolite to which they were linked via an enzymatic reaction. Searching the metabolome data for features having (i) an identical precursor m/z value, (ii) a similar retention time, and (iii) a similar MS fragmentation spectrum to those associated with one of the KEGG-retrieved compounds, results in additional feature annotations in a first round. After the set of seed metabolites has been augmented with the newly annotated features, the whole procedure can be repeated. This recursive method allowed the authors to annotate more than 1000 compounds, yet, as is the case for *mummichog* and related approaches, annotations involved only compounds that are already present in pathway databases. Thus, by mapping of all annotated features onto a metabolome network, only known pathways will be displayed.

Many of the identified, especially secondary, metabolites are not present in pathway databases, but might be found in compound databases [151,152]. Barupal et al. (2012) [84] combined information from both pathway and compound databases to construct networks: MDNs based on pathway databases were overlaid with StNs, in which edges represented structural similarity. In addition, the authors included the unknown compounds as well by constructing and overlaying an SpN, thus, connecting unknown features with known features via the SpN. Rather than overlaying different types of metabolome networks, Morreel et al. (2014) [61] added CID spectral similarity as an additional attribute to the edges in the MDN, allowing to directly judge the biochemical validity of the biotransformation displayed by each of the edges via their CSPP algorithm.

2.3.1. Dealing with pathways that are not yet present in any database

Tackling the system-wide annotation of pathways by connecting completely unknown compounds that are neither present in pathway databases nor in compound databases might benefit from SIL precursor feeding experiments (see '*2.2.2 Data fusion using retention time alignment: spectral metadata analysis*') or from generating differential metabolome networks, e.g., from samples taken throughout development [81], across different tissues [153], derived following an abiotic or a biotic stress application, or based on (sub)cellular profiling. Differential metabolome networks facilitate the suggestion of putative pathways, because compounds of a particular class are often synthesized at a specific developmental stage, in a specific tissue, cell or cellular compartment, or upon a particular treatment. Notably, dedicated instruments are sometimes needed, such as MS imaging or nano-LC-MS instruments, for tissue-specific or cellular metabolomics [154,155]. The differential biotransformations in these comparative metabolome experiments might be derived by computing all possible mass differences and comparing their frequencies between developmental stages, tissues/cells/compartments, or treatments.

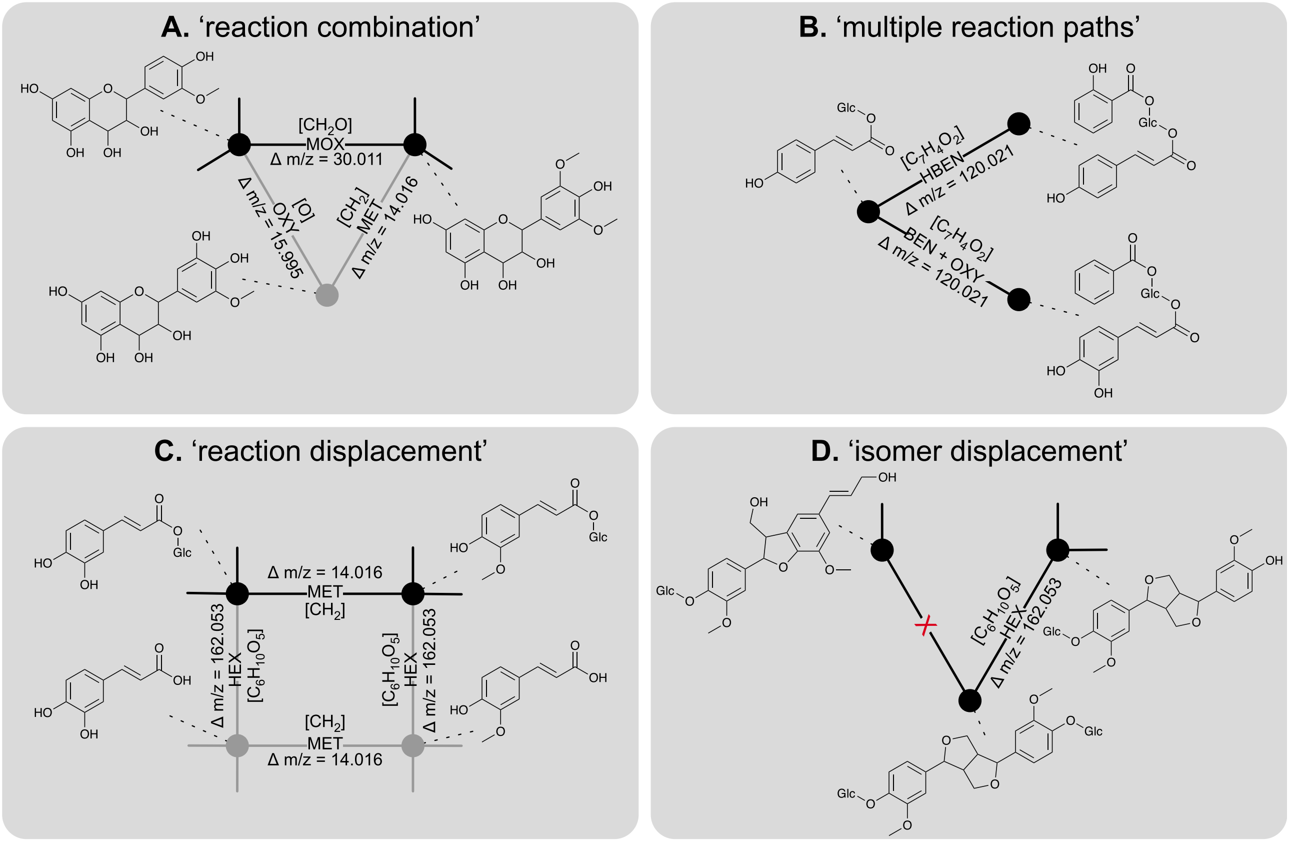

Fig. 3. Biotransformation–enzymatic conversion anomalies. See text for explanation (*'2.3 Flaw 3: biotransformations do not adequately reflect enzymatic reactions'*). Edges and black nodes represent putative biotransformations and features. Gray nodes represent putative, non-detected pathway intermediates. BEN, benzoylation; HBEN, hydroxybenzoylation; HEX, hexosylation; MET, methylation; MOX, methoxylation; OXY, oxygenation.

Rather than constructing putative biotransformation paths by an in-depth analysis of the metabolome data, biotransformation paths between two pre-specified compounds, i.e., precursor and end-product, might be computationally generated via a so-called retrosynthesis. To predict biotransformation paths, several biotransformation databases and tools have been established [156–161]. Using these databases and tools, a variety of *in silico* enzymatic products of a particular compound can be generated, and by performing this iteratively, a network (directed graph) can be constructed with nodes and edges reflecting *in silico*-generated compounds and biotransformations, respectively. As mentioned above, such a strategy allows creating *in silico* compound databases, yet also enables tracing putative reaction paths between precursor and end-product in the network via, e.g., graph theory-based pathway searching algorithms. To prevent combinatorial explosion, restrictions on the biotransformations and the generated biotransformation paths need to be imposed [162]. For example, in the BNICE (Biochemical Network Integrated Computational Explorer) framework, biotransformations are only allowed on compounds having particular substructures; in addition, only the thermodynamically feasible biotransformation paths are retained [163]. However, even when using constraints, multiple biotransformation paths are still predicted. Further optimization involves atom mapping [164–167] and/or CBM [23,24,168]. Atom mapping only accepts biotransformation paths comprised of steps in which an atom is transferred between the substrate and product. For example, when considering the hexokinase reaction in the network, the biotransformation path containing a step in which glucose is connected to glucose-6-phosphate will be retained by atom mapping, whereas the alternative path in which glucose is connected with ADP will be rejected. CBM enables selecting the mass-balanced pathways; a CBM approach using *in silico*-generated metabolic networks has been developed [43]. The various graph theory-based path searching algorithms and the implemented constraints have been reviewed by several groups [169,170].

Although metabolome networks might allow predicting biochemical pathways, the enzymes/genes responsible for the individual reactions cannot be identified without additional information. Annotating the correct enzymes/genes associated with each reaction step is important as it provides considerable support for the *in vivo* existence of the pathway. Enzyme/gene information can be retrieved from other omics technologies [171]. Such multi-omics data are often generated in a time-course experiment [172], or in a comparative experiment involving particular treatments or different tissues, cells, or cellular compartments [173]. These comparative experimental set-ups might be especially relevant for secondary metabolism. Opposite to the production of primary metabolites which are precursors for a variety of growth and stress physiological processes, selective sets of secondary metabolites are coordinately produced in a spatially and temporally controlled manner under particular stress conditions [174]. Taking also into account that many biochemical pathways operate within enzyme complexes [175], metabolic network construction benefits most from combining metabolome networks with gene co-expression and/or protein–protein interaction networks [176,177], and using, e.g., the guilt-by-association principle (see '*1.1 Metabolic networks provide a basis to study the control and regulation of metabolism*') [178–180].

In addition, metabolomics can be combined with genetic/biochemical screens for gene annotation. Forward genetics screens are based on generating a large random collection of mutants, and screening them for the (dis)appearance of particular metabolites via metabolomics [181]. Alternatively, as the levels of many metabolites vary quantitatively rather than in a Mendelian way in natural or mapping populations, genes or genetic loci that govern the metabolite abundances can be revealed by subjecting population-wide metabolome data to QTL analysis [182–185] or genome-wide association studies (GWAS) [186,187]. QTL/GWAS studies might involve metabolome network construction, for example, in a bipartite network in which nodes represent either features or genetic loci. New biochemical pathways can then be proposed based on the indirect connection between known and unknown features via their association with the same genetic locus [188]. In a reverse genetics approach, gene overexpressing or down-regulated lines can be compared with control lines. This is referred to as *ex vivo* metabolome profiling [189]. The latter term also encompasses the use of inhibitors as a way to inactivate a particular enzyme or enzyme class. If *ex vivo* metabolome profiling fails to annotate the function of a particular gene/enzyme, *in vitro* activity-based metabolome profiling can be attempted. Compared to *in vitro* enzymatic assays that involve the addition of the putative substrate to the enzyme, a metabolite extract is fed to the enzyme and new features are searched for [190]; *in vitro* activity-based metabolome profiling is high-throughput and offers much more chance of finding the *in vivo* substrate as opposed to traditional enzyme assays. Strategies to combine metabolomics with genetics to resolve biochemical pathways have been reviewed [189,191].

3. Summary and outlook

To understand how plant secondary metabolism is organized and controlled/regulated, and how it compares with primary metabolism, it is necessary to fully comprehend its overall pathway architecture. Despite that pathways of secondary metabolism are insufficiently represented in current pathway databases and GSMMs, metabolomics data represent a rich resource to discover the various secondary metabolites that are present in a particular species via metabolome network construction. Expectedly, such networks will better reflect the metabolic network when they are constructed from a combination of CNs, SpNs, StNs and MDNs, and taking information from pathway databases or existing GSMMs into account.

Including retrosynthesis, which is currently mainly used in the synthetic biology field, into metabolome network studies offers a powerful strategy to predict specialized metabolic pathways. To further support the *in silico*-predicted optimal enzyme-catalyzed path towards a particular product, the intermediates can be searched for in the metabolome network. Recent progress in the annotation of metabolite substructure based on CID spectra might help in annotating unknown peaks as pathway intermediates [192]. However, retrosynthesis, performed on a multitude of precursor–product metabolite pairs, might boost the metabolic network complexity. In this case, a similar strategy could be applied as currently performed in the curation of GSMM, i.e., applying CBM to remove non-essential edges on the one hand and to point pathway gaps on the other hand [25]. Further support for particular biotransformation paths can also be provided via another GSMM-applied strategy: pathway mapping using multi-omics data, especially when derived from time-course or comparative studies. Currently, pathway mapping of comparative multi-omics data hardly considers pathway compartmentalization, despite the fact that topological differences are evident between global and compartment-specific metabolic networks [193]. Fortunately, such information has increasingly become available during the last decade due to the construction of pathway databases containing tissue-, cell-, and cellular compartment-specific information [20,21].

Of utmost importance for unraveling secondary metabolic pathways is the continued effort to improve the annotation of metabolome network nodes in case retrosynthesis-based annotation is not successful. Despite the advances obtained by the development of databases and *in silico* tools, it is increasingly recognized that further progress is hindered by the rather slow growth of annotated, metabolite-associated, MS spectral data. In addition, such data are available in a variety of MS spectral types, such as MS/MS and MS^n spectra, hence data fusion and spectral metadata analysis will become increasingly important for structural characterization. The latter approaches are rather seldomly pursued as (i) instrument platforms in many labs are over-occupied, (ii) an extensive structural characterization is unnecessary for many default metabolomics approaches, and, perhaps most importantly, (iii) only few labs have access to different spectroscopic instruments recording MS/MS as well as MS^n. Consequently, dedicated software that handles the alignment of CID spectra across different instrument platforms or recorded under different experimental settings (e.g., following metal complexation, derivatization, isotope labeling, or different CID spectral settings), is currently underdeveloped, as are spectral databases connecting these different types of CID spectra.

By considering different metabolome network metrics (abundance-based correlation, spectral similarity, etc.), retrosynthesis, multi-omics data, and final authentication via genetic/biochemical screening experiments, an improved metabolome-to-metabolic network transition is expected. Primordial is the inclusion of information from pathway databases, yet the same databases have to be updated with new insight gained from the metabolic networks, and new spectral databases are needed that integrate the different types of spectra, as obtained via data fusion and mass spectral metadata analysis. As the number of metabolic networks will continue to rise, non-curated data might increasingly enter these databases. Therefore, concomitant with the generation of metabolic networks, an investment in database curation will be necessary [194,195]. Consequently, elucidating plant secondary metabolism at an increasing pace requires a better collaboration between plant specialists, computational biologists, computer scientists, database developers, and the GSMM and metabolite identification research communities.

Glossary

Abundance-based correlation. The correlation between the abundances of two features across biological replicates, by default computed as the Pearson or Spearman correlation coefficient.

Biotransformation. Term derived from the computational field that refers to the reactions interconverting biochemical compounds; biotransformations reflect tentative enzymatic conversions.

Compound database. Publicly available database that contains mainly structural and physicochemical data of compounds. Biochemical information is often lacking or rudimentary.

Constraint-based modeling (CBM). Approach to determine the flux space by solving the system equations (see Glossary) assuming a metabolic (pseudo-)steady state, i.e., by solving $\boldsymbol{N}v$=0 with $\boldsymbol{N}$ being the stoichiometric matrix (rows and columns are metabolites and reactions), and v being the flux vector (in case of a steady state, the flux for each enzymatic conversion equals its reaction rate). The flux space is formed by the set of non-trivial solutions for v. Besides the (pseudo-)steady state, additional constraints such as reaction thermodynamics, and setting lower and upper bonds to the enzymatic reaction rates, can be included in the model.

Feature. A term derived from machine learning. The ions detected via MS are referred to as features following processing of the chromatogram peaks. Each feature is characterized by a *m/z* value and a retention time.

Isotopologs. Molecules that only differ in their isotopic composition.

Metabolic control/regulation. Conceptual definitions for metabolic regulation and control are based on Fell (1997) [196]. Metabolic regulation is defined within the context of homeostasis, i.e., in order to stabilize metabolite concentrations and pathway fluxes. Metabolic control defines the change in pathway flux upon a change in reaction rate of a particular enzyme. For example, a linear pathway in which the first enzyme is feedback inhibited by a pathway intermediate, will be regulated by the first enzyme, yet the pathway flux will be mostly affected by the enzymes downstream of the intermediate performing the feedback inhibition [197].

Metabolic network. Nodes and edges represent metabolites and enzymatic conversions based on information from pathway databases or from a genome-scale metabolic model (GSMM).

Metabolome network. Network constructed from metabolome data; nodes represent features. Edges reflect an abundance-based correlation (correlation network, CN), a spectral similarity computed using, e.g., the dot product (spectral similarity network, SpN), a structural similarity computed as, e.g., the Tanimoto coefficient (structural similarity network, StN), or a mass difference (mass difference network, MDN).

Network module. Sub-network of which the nodes are involved in the same biological function that is different from the biological functions of other modules, for example, membership of a particular biochemical pathway.

Network motif. A recurring, significant pattern of node interconnections.

Pathway database. A database containing predominantly biochemical pathway information, i.e., the metabolites and the enzymes working on them. Often species-specific pathway information is present and sometimes gene/protein data.

System equations. A system equation is proposed for each metabolite, and defines how its concentration is determined by the difference between its rates of synthesis and degradation. At (pseudo-)steady state, the synthesis rate is balanced by the degradation rate, yielding a stable metabolite concentration for which the instantaneous rate of change is zero.

CRediT authorship contribution statement

Sandrien Desmet: Writing - original draft, Writing - review & editing. **Marlies Brouckaert:** Writing - original draft, Writing - review & editing. **Wout Boerjan:** Writing - original draft, Writing - review & editing. **Kris Morreel:** Writing - original draft, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. Annick Bleys for proofreading and preparing the manuscript. The authors received funding from the ERC-Advanced Grant 2019 POPMET. Sandrien Desmet and Marlies Brouckaert are indebted to the FWO-SB for predoctoral fellowships (151562 and 1S38920N). We also thank the Bijzonder Onderzoeksfonds-Zware Apparatuur of Ghent University for the Fourier transform ion cyclotron resonance mass spectrometer (174PZA05) and the Hercules program of Ghent University for the Synapt Q-Tof (Grant AUGE/014).

References

[1] Initiative TAG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000;408:796–815.

[2] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.

[3] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. The sequence of the human genome. Science 2001;291:1304–51.

[4] Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, et al. Metabolite profiling for plant functional genomics. Nat Biotechnol 2000;18:1157–61.

[5] Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 1999;29:1181–9.

[6] Oliver SG. Yeast as a navigational aid in genome analysis. Microbiology 1997;143:1483–7.

[7] Almaas E. Biological impacts and context of network theory. J Exp Biol 2007;210:1548–58.

[8] Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. SIAM Rev 2009;51:661–703.

[9] Newman MEJ. The structure and function of complex networks. SIAM Rev 2003;45:167–256.

[10] Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet 2007;8:450–61.

[11] Peter IS. The function of architecture and logic in developmental gene regulatory networks. Curr Top Dev Biol 2020;139:267–95.

[12] Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. Nat Chem Biol 2013;9:72–80.

[13] Peracchi A. The limits of enzyme specificity and the evolution of metabolism. Trends Biochem Sci 2018;43:984–96.

[14] Schwab W. Metabolome diversity: too few genes, too many metabolites?. Phytochemistry 2003;62:837–49.

[15] Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, et al. Network context and selection in the evolution to enzyme specificity. Science 2012;337:1101–4.

[16] Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res 2019;47:D590–5.

[17] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 2000;28:27–30.

[18] Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci 2019;28:1947–51.

[19] Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform 2019;20:1085–93.

[20] Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, et al. Plant Reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Res 2017;45:D1029–39.

[21] Naithani S, Gupta P, Preece J, D'Eustachio P, Elser JL, et al. Plant Reactome: a knowledgebase and resource for comparative pathway analysis. Nucleic Acids Res 2020;48:D1093–103.

[22] Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, et al. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res 2018;46:D1181–9.

[23] Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO. Metabolic pathways in the post-genome era. Trends Biochem Sci 2003;28:250–8.

[24] Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. J Biol Syst 1994;2:165–82.

[25] Volkova S, Matos MRA, Mattanovich M, Marín de Mas I. Metabolic modelling as a framework for metabolomics data integration and analysis. Metabolites 2020.

[26] Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA 2007;104:1777–82.

[27] Poolman MG, Miguet L, Sweetlove LJ, Fell DA. A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol 2009;151:1570–81.

[28] Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. Genome Biol 2019;20:121.

[29] Pfeiffer T, Soyer OS, Bonhoeffer S. The evolution of connectivity in metabolic networks. PLoS Biol 2005;3:e228.

[30] Broido AD, Clauset A. Scale-free networks are rare. Nat Commun 2019;10:1017.

[31] Fell D, Wagner A. Structural properties of metabolic networks: implications for evolution and modelling of metabolism, in Animating the Cellular Map, J. H. Hofmeyr, J.M. Rohwer, and J.L. Snoep, Editors. 2000, Stellenbosch University Press: Stellenbosch, South Africa. p. 79-85.

[32] Gamermann D, Triana-Dopico J, Jaime R. A comprehensive statistical study of metabolic and protein–protein interaction network properties. Phys A 2019;534:122204.

[33] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási A-L. The large-scale organization of metabolic networks. Nature 2000;407:651–4.

[34] Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. Mol Biosyst 2009;5:1482–93.

[35] Takemoto K. Metabolic networks are almost nonfractal: a comprehensive evaluation. Phys Rev E 2014;90:022802.

[36] Winterbach W, Wang H, Reinders M, Van Mieghem P, de Ridder D. Metabolic network destruction: relating topology to robustness. Nano CommunNetw 2011;2:88–98.

[37] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi A-L. Hierarchical organization of modularity in metabolic networks. Science 2002;297:1551–5.

[38] Lacroix V, Fernandes CG, Sagot M-F. Motif search in graphs: application to metabolic networks. IEEE/ACM Trans Comput Biol Bioinform 2006;3:360–8.

[39] Sridharan GV, Ullah E, Hassoun S, Lee K. Discovery of substrate cycles in large scale metabolic networks using hierarchical modularity. BMC Syst Biol 2015;9:5.

[40] Sridharan GV, Hassoun S, Lee K. Identification of biochemical network modules based on shortest retroactive distances. PLoS Comput Biol 2011;7:e1002262.

[41] Gutteridge A, Kanehisa M, Goto S. Regulation of metabolic networks by small molecule metabolites. BMC Bioinf 2007;8:88.

[42] Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. Nat Biotechnol 2004;22:86–92.

[43] Kumar S, Mahajan S, Jain S. Feedbacks from the metabolic network to the genetic network reveal regulatory modules in E. coli and B. subtilis. PLoS ONE 2018;13:e0203311.

[44] Yeang C-H. Integration of metabolic reactions and gene regulation. Mol Biotechnol 2011;47:70–82.

[45] Kim P, Lee D-S, Kahng B. Biconnectivity of the cellular metabolism: a cross-species study and its implication for human diseases. Sci Rep 2015;5:15567.

[46] Alseekh S, Tohge T, Wendenberg R, Scossa F, Omranian N, et al. Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. Plant Cell 2015;27:485–512.

[47] Fernie AR, Tohge T. The genetics of plant metabolism. Annu Rev Genet 2017;51:287–310.

[48] Wu S, Alseekh S, Cuadros-Inostroza Á, Fusari CM, Mutwil M, et al. Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. PLoS Genet 2016;12:e1006363.

[49] Heinrich R, Rapoport TA. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. Eur J Biochem 1974;42:89–95.

[50] Kacser H, Burns JA. The control of flux. Symp Soc Exp Biol 1973;27:65–104.

[51] Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, et al. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 2001;13:11–29.

[52] Camacho D, de la Fuente A, Mendes P. The origin of correlations in metabolomics data. Metabolomics 2005;1:53–63.

[53] Steuer R, Gross T, Selbig J, Blasius B. Structural kinetic modeling of metabolic networks. Proc Natl Acad Sci USA 2006;103:11868–73.

[54] Steuer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in metabolomic networks. Bioinformatics 2003;19:1019–26.

[55] Wu S, Tohge T, Cuadros-Inostroza Á, Tong H, Tenenboim H, et al. Mapping the Arabidopsis metabolic landscape by untargeted metabolomics at different environmental conditions. Mol Plant 2018;11:118–34.

[56] Li X, Bergelson J, Chapple C. The ARABIDOPSIS accession Pna-10 is a naturally occurring sng1 deletion mutant. Mol Plant 2010;3:91–100.

[57] Tohge T, Wendenburg R, Ishihara H, Nakabayashi R, Watanabe M, et al. Characterization of a recently evolved flavonol-phenylacyltransferase gene provides signatures of natural light selection in Brassicaceae. Nat Commun 2016;7:12399.

[58] Li H, Lv Q, Ma C, Qu J, Cai F, et al. Metabolite profiling and transcriptome analyses provide insights into the flavonoid biosynthesis in the developing seed of tartary buckwheat (Fagopyrum tataricum). J Agric Food Chem 2019;67:11262–76.

[59] Saito K. Phytochemical genomics – a new trend. Curr Opin Plant Biol 2013;16:373–80.

[60] Wisecaver JH, Borowsky AT, Tzin V, Jander G, Kliebenstein DJ, et al. A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. Plant Cell 2017;29:944–59.

[61] Morreel K, Saeys Y, Dima O, Lu F, Van de Peer Y, et al. Systematic structural characterization of metabolites in Arabidopsis via candidate substrate-product pair networks. Plant Cell 2014;26:929–45.

[62] Shen X, Wang R, Xiong X, Yin Y, Cai Y, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nat Commun 2019;10:1516.

[63] Gaquerel E, Kotkar H, Onkokesung N, Galis I, Baldwin IT. Silencing an N-acyltransferase-like involved in lignin biosynthesis in Nicotiana attenuata dramatically alters herbivory-induced phenolamide metabolism. PLoS ONE 2013;8:e62336.

[64] Xu Y-F, Lu W, Rabinowitz JD. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography–mass spectrometry-based metabolomics. Anal Chem 2015;87:2273–81.

[65] Lei Z, Jing L, Qiu F, Zhang H, Huhman D, et al. Construction of an ultrahigh pressure liquid chromatography-tandem mass spectral library of plant natural products and comparative spectral analyses. Anal Chem 2015;87:7373–81.

[66] Oberacher H, Whitley G, Berger B. Evaluation of the sensitivity of the 'Wiley registry of tandem mass spectral data, MSforID'with MS/MS data of the 'NIST/NIH/EPA mass spectral library'. J Mass Spectrom 2013;48:487–96.

[67] Cao M, Fraser K, Rasmussen S. Computational analyses of spectral trees from electrospray multi-stage mass spectrometry to aid metabolite identification. Metabolites 2013;3:1036–50.

[68] Kasper PT, Rojas-Chertó M, Mistrik R, Reijmers T, Hankemeier T, et al. Fragmentation trees for the structural characterisation of metabolites. Rapid Commun Mass Spectrom 2012;26:2275–86.

[69] Vaniya A, Fiehn O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. Trends Anal Chem 2015;69:52–61.

[70] Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703–14.

[71] Vinaixa M, Schymanski EL, Neumann S, Navarro M, Salek RM, et al. Mass spectral databases for LC/MS-and GC/MS-based metabolomics: state of the field and future prospects. Trends Anal Chem 2016;78:23–35.

[72] Sun X, Weckwerth W. COVAIN: a toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. Metabolomics 2012;8:81–93.

[73] Fukushima A, Kusano M, Redestig H, Arita M, Saito K. Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. BMC Syst Biol 2011;5:1.

[74] Nägele T, Weckwerth W. A workflow for mathematical modeling of subcellular metabolic pathways in leaf metabolism of Arabidopsis thaliana. Front Plant Sci 2013;4:541.

[75] Rosato A, Tenori L, Cascante M, Ramon De Atauri Carulla P, Martins dos Santos VAP, et al. From correlation to causation: analysis of metabolomics data using systems biology approaches. Metabolomics 2018;14:37.

[76] Costello CA, Hu T, Liu M, Zhang W, Furey A, et al. Differential correlation network analysis identified novel metabolomics signatures for non-responders to total joint replacement in primary osteoarthritis patients. Metabolomics 2020;16:61.

[77] Jahagirdar S, Saccenti E. On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis. Metabolites 2020;10:171.

[78] Krumsiek J, Suhre K, Illig T, Adamski J, Theis FJ. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. BMC Syst Biol 2011;5:21.

[79] Aron AT, Gentry EC, McPhail KL, Nothias L-F, Nothias-Esposito M, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. Nat Protoc 2020;15:1954–91.

[80] Naake T, Gaquerel E. MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. Bioinformatics 2017;33:2419–20.

[81] Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci USA 2012;109:E1743–52.

[82] Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol 2016;34:828–37.

[83] Barupal DK, Fiehn O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep 2017;7:14567.
[84] Barupal DK, Haldiya PK, Wohlgemuth G, Kind T, Kothari SL, et al. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. BMC Bioinf 2012;13:99.
[85] Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. Metabolomics 2006;2:155–64.
[86] Jourdan F, Breitling R, Barrett MP, Gilbert D. MetaNetter: inference and visualization of high-resolution metabolomic networks. Bioinformatics 2008;24:143–5.
[87] Rogers S, Scheltema RA, Girolami M, Breitling R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. Bioinformatics 2009;25:512–8.
[88] Weber RJM, Viant MR. MI-Pack: increased confidence of metabolite identification in mass spectra by integrating accurate masses and metabolic pathways. Chemometrics Intell Lab Syst 2010;104:75–82.
[89] Laitinen T, Morreel K, Delhomme N, Gauthier A, Schiffthaler B, et al. A key role for apoplastic H_2O_2 in Norway spruce phenolic metabolism. Plant Physiol 2017;174:1449–75.
[90] Dima O, Morreel K, Vanholme B, Kim H, Ralph J, et al. Small glycosylated lignin oligomers are stored in Arabidopsis leaf vacuoles. Plant Cell 2015;27:695–710.
[91] Niculaes C, Morreel K, Kim H, Lu F, McKee LS, et al. Phenylcoumaran benzylic ether reductase prevents accumulation of compounds formed under oxidative conditions in poplar xylem. Plant Cell 2014;26:3775–91.
[92] Rasche F, Scheubert K, Hufsky F, Zichner T, Kai M, et al. Identifying the unknowns by aligning fragmentation trees. Anal Chem 2012;84:3417–26.
[93] Rojas-Cherto M, Peironcely JE, Kasper PT, van der Hooft JJJ, de Vos RCH, et al. Metabolite identification using automated comparison of high-resolution multistage mass spectral trees. Anal Chem 2012;84:5524–34.
[94] Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. Anal Chem 2014;86:6812–7.
[95] Domingo-Almenara X, Montenegro-Burke JR, Guijas C, Majumder EL-W, Benton HP, et al. Autonomous METLIN-guided in-source fragment annotation for untargeted metabolomics. Anal Chem 2019;91:3246–53.
[96] Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, et al. Deconvolution gas chromatography/mass spectrometry of urinary organic acids–potential for pattern recognition and automated identification of metabolic disorders. Rapid Commun Mass Spectrom 1999;13:279–84.
[97] Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem 2012;84:283–9.
[98] 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;88:9037–46.
[99] Senan O, Aguilar-Mogas A, Navarro M, Capellades J, Noon L, et al. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. Bioinformatics 2019;35:4089–97.
[100] De Vijlder T, Valkenborg D, Lemière F, Romijn EP, Laukens K, et al. A tutorial in small molecule identification via electrospray ionization-mass spectrometry: the practical art of structural elucidation. Mass Spectrom Rev 2018;37:607–29.
[101] Gipson GT, Tatsuoka KS, Sokhansanj BA, Ball RJ, Connor SC. Assignment of MS-based metabolomic datasets via compound interaction pair mapping. Metabolomics 2008;4:94–103.
[102] Burgess KEV, Borutzki Y, Rankin N, Daly R, Jourdan F. MetaNetter 2: A Cytoscape plugin for ab initio network analysis and metabolite feature classification. J Chromatogr B 2017;1071:68–74.
[103] Dixon RA, Strack D. Phytochemistry meets genome analysis, and beyond. Phytochemistry 2003;62:815–6.
[104] Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. Metabolite profiling: from diagnostics to systems biology. Nat Rev Mol Cell Biol 2004;5:763–9.
[105] Collard F, Baldin F, Gerin I, Bolsée J, Noël G, et al. A conserved phosphatase destroys toxic glycolytic side products in mammals and yeast. Nat Chem Biol 2016;12:601–7.
[106] de Crécy-Lagard V, Haas D, Hanson AD. Newly-discovered enzymes that function in metabolite damage-control. Curr Opin Chem Biol 2018;47:101–8.
[107] Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics 2009;25:218–24.
[108] Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinf 2007;8:105.
[109] Neumann S, Böcker S. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. Anal Bioanal Chem 2010;398:2779–88.
[110] Rojas-Chertó M, Kasper PT, Willighagen EL, Vreeken RJ, Hankemeier T, et al. Elemental composition determination based on MSn. Bioinformatics 2011;27:2376–83.
[111] Hufsky F, Böcker S. Mining molecular structure databases: identification of small molecules based on fragmentation mass spectrometry data. Mass Spectrom Rev 2017;36:624–33.
[112] da Silva RR, Wang M, Nothias L-F, van der Hooft JJJ, Caraballo-Rodríguez AM, et al. Propagating annotations of molecular networks using in silico fragmentation. PLoS Comput Biol 2018;14:e1006089.
[113] Ernst M, Kang KB, Caraballo-Rodríguez AM, Nothias L-F, Wandy J, et al. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. Metabolites 2019;9:144.
[114] Pilon AC, Gu H, Raftery D, da Silva Bolzani V, Peporine Lopes N, et al. Mass spectral similarity networking and gas-phase fragmentation reactions in the structural analysis of flavonoid glycoconjugates. Anal Chem 2019;91:10413–23.
[115] Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinf 2010;11:148.
[116] Li L, Li R, Zhou J, Zuniga A, Stanislaus AE, et al. MyCompoundID: using an evidence-based metabolome library for metabolite identification. Anal Chem 2013;85:3401–8.
[117] Hadadi N, Hafner J, Shajkofci A, Zisaki A, Hatzimanikatis V. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. ACS Synth Biol 2016;5:1155–66.
[118] Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Cheminformatics 2015;7:44.
[119] Menikarachchi LC, Hill DW, Hamdalla MA, Mandoiu II, Grant DF. In silico enzymatic synthesis of a 400,000 compound biochemical database for nontargeted metabolomics. J Chem Inf Model 2013;53:2483–92.
[120] Qiu F, Fine DD, Wherritt DJ, Lei Z, Sumner LW. PlantMAT: a metabolomics tool for predicting the specialized metabolic potential of a system and for large-scale metabolite identifications. Anal Chem 2016;88:11373–83.
[121] Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RCH, Vervoort J, et al. In silico prediction and automatic LC–MSn annotation of green tea metabolites in urine. Anal Chem 2014;86:4767–74.
[122] 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;1382:136–64.
[123] Biancolillo A, Marini F. Chemometric methods for spectroscopy-based pharmaceutical analysis. Front Chem 2018;6:576.
[124] van Outersterp RE, Houthuijs KJ, Berden G, Engelke UF, Kluijtmans LAJ, et al. Reference-standard free metabolite identification using infrared ion spectroscopy. Int J Mass Spectrom 2019;443:77–85.
[125] Forcisi S, Moritz F, Lucio M, Lehmann R, Stefan N, et al. Solutions for low and high accuracy mass spectrometric data matching: a data-driven annotation strategy in nontargeted metabolomics. Anal Chem 2015;87:8917–24.
[126] Wolfender J-L, Queiroz EF, Hostettmann K. Phytochemistry in the microgram domain — a LC-NMR perspective. Magn Reson Chem 2005;43:697–709.
[127] Boiteau RM, Hoyt DW, Nicora CD, Kinmonth-Schultz HA, Ward JK, et al. Structure elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction. Metabolites 2018;8:8.
[128] Ichou F, Schwarzenberg A, Lesage D, Alves S, Junot C, et al. Comparison of the activation time effects and the internal energy distributions for the CID, PQD and HCD excitation modes. J Mass Spectrom 2014;49:498–508.
[129] Schaller-Duke RM, Bogala MR, Cassady CJ. Electron transfer dissociation and collision-induced dissociation of underivatized metallated oligosaccharides. J Am Soc Mass Spectrom 2018;29:1021–35.
[130] Dallinga JW, Heerma W. Fast atom bombardment mass spectrometry of the D-aldohexoses and some deoxyaldohexoses. Biomed Environ Mass Spectrom 1989;18:363–72.
[131] March RE, Stadey CJ. A tandem mass spectrometric study of saccharides at high mass resolution. Rapid Commun Mass Spectrom 2005;19:805–12.
[132] Mulroney B, Peel JB, Traeger JC. Theoretical study of deprotonated glucopyranosyl disaccharide fragmentation. J Mass Spectrom 1999;34:856–71.
[133] Fang TT, Zirrolli J, Bendiak B. Differentiation of the anomeric configuration and ring form of glucosyl-glycolaldehyde anions in the gas phase by mass spectrometry: isomeric discrimination between m/z 221 anions derived from disaccharides and chemical synthesis of m/z 221 standards. Carbohydr Res 2007;342:217–35.
[134] Bueschl C, Kluger B, Neumann NKN, Doppler M, Maschietto V, et al. MetExtract II: a software suite for stable isotope-assisted untargeted metabolomics. Anal Chem 2017;89:9518–26.
[135] Kostyukevich Y, Acter T, Zherebker A, Ahmed A, Kim S, et al. Hydrogen/deuterium exchange in mass spectrometry. Mass Spectrom Rev 2018;37:811–53.
[136] Lam W, Ramanathan R. In electrospray ionization source hydrogen/deuterium exchange LC-MS and LC-MS/MS for characterization of metabolites. J Am Soc Mass Spectrom 2002;13:345–53.
[137] Murphy RC, Okuno T, Johnson CA, Barkley RM. Determination of double bond positions in polyunsaturated fatty acids using the photochemical Paternò-Büchi reaction with acetone and tandem mass spectrometry. Anal Chem 2017;89:8545–53.
[138] Cheng, S.C., Bhat, S.M., Shiea, J. Flame atmospheric pressure chemical ionization coupled with negative electrospray ionization mass spectrometry for ion molecule reactions. J Am Soc Mass Spectrom 2017;28:1473-1481.
[139] Cottret L, Wildridge D, Vinson F, Barrett MP, Charles H, et al. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. Nucleic Acids Res 2010;38:W132–7.

[140] García-Alcalde F, García-López F, Dopazo J, Conesa A. Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. Bioinformatics 2011;27:137–9.
[141] Junker BH, Klukas C, Schreiber F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. BMC Bioinf 2006;7:109.
[142] Karnovsky A, Weymouth T, Hull T, Tarcea VG, Scardoni G, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. Bioinformatics 2012;28:373–80.
[143] Klukas C, Schreiber F. Integration of -omics data and networks for biomedical research with VANTED. J Integrative Bioinformatics 2010;7:112.
[144] Leader DP, Burgess K, Creek D, Barrett MP. Pathos: a web facility that uses metabolic maps to display experimental changes in metabolites identified by mass spectrometry. Rapid Commun Mass Spectrom 2011;25:3422–6.
[145] Suhre K, Schmitt-Kopplin P. MassTRIX: mass translator into pathways. Nucleic Acids Res 2008;36:W481–4.
[146] Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, et al. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. BMC Bioinf 2019;20:614.
[147] Booth SC, Weljie AM, Turner RJ. Computational tools for the secondary analysis of metabolomics experiments. Comp Struct Biotechnol J 2013;4:e201301003.
[148] Alden N, Krishnan S, Porokhin V, Raju R, McElearney K, et al. Biologically consistent annotation of metabolomics data. Anal Chem 2017;89:13097–104.
[149] Pirhaji L, Milani P, Leidl M, Curran T, Avila-Pacheco J, et al. Revealing disease-associated pathways by network integration of untargeted metabolomics. Nat Methods 2016;13:770–6.
[150] Li S, Park Y, Duraisingham S, Strobel FH, Khan N, et al. Predicting network activity from high throughput metabolomics. PLoS Comput Biol 2013;9:e1003123.
[151] Grapov D, Wanichthanarak K, Fiehn O. MetaMapR: pathway independent metabolomic network analysis incorporating unknowns. Bioinformatics 2015;31:2757–60.
[152] Wanichthanarak K, Fan S, Grapov D, Barupal DK, Fiehn O. Metabox: a toolbox for metabolomic data analysis, interpretation and integrative exploration. PLoS ONE 2017;12:e0171046.
[153] Li D, Heiling S, Baldwin IT, Gaquerel E. Illuminating a plant's tissue-specific metabolic diversity using computational metabolomics and information theory. Proc Natl Acad Sci USA 2016;113:E7610–8.
[154] Ferguson CN, Fowler JWM, Waxer JF, Gatti RA, Loo JA. Mass spectrometry-based tissue imaging of small molecules. AdvExpMedBiol 2019;1140:99–109.
[155] Porta Siegel T, Hamm G, Bunch J, Cappell J, Fletcher JS, et al. Mass spectrometry imaging and integration with other imaging modalities for greater molecular understanding of biological tissues. Mol Imaging Biol 2018;20:888–901.
[156] Wicker J, Lorsbach T, Gütlein M, Schmid E, Latino D, et al. enviPath – The environmental contaminant biotransformation pathway resource. Nucleic Acids Res 2016;44:D502–8.
[157] Yousofshahi M, Manteiga S, Wu C, Lee K, Hassoun S. PROXIMAL: a method for prediction of xenobiotic metabolism. BMC Syst Biol 2015;9:94.
[158] Ellis LBM, Gao J, Fenner K, Wackett LP. The University of Minnesota pathway prediction system: predicting metabolic logic. Nucleic Acids Res 2008;36:W427–32.
[159] de Groot MJL, van Berlo RJP, van Winden WA, Verheijen PJT, Reinders MJT, et al. Metabolite and reaction inference based on enzyme specificities. Bioinformatics 2009;25:2975–82.
[160] Greene N, Judson PN, Langowski JJ, Marchant CA. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ Res 1999;10:299–314.
[161] Klopman G, Dimayuga M, Talafous JMETA. 1. A program for the evaluation of metabolic transformation of chemicals. J Chem Inf Comput Sci 1994;34:1320–5.
[162] Pertusi DA, Stine AE, Broadbelt LJ, Tyo KEJ. Efficient searching and annotation of metabolic networks using chemical similarity. Bioinformatics 2015;31:1016–24.
[163] Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, et al. Exploring the diversity of complex metabolic networks. Bioinformatics 2005;21:1603–9.
[164] Arita M. Metabolic reconstruction using shortest paths. Simul Pract Theory 2000;8:109–25.
[165] Blum T, Kohlbacher O. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. Bioinformatics 2008;24:2108–9.
[166] Frainay C, Jourdan F. Computational methods to identify metabolic sub-networks based on metabolomic profiles. Brief Bioinform 2017;18:43–56.
[167] Pitkänen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol 2009;3:103.
[168] Chowdhury A, Maranas CD. Designing overall stoichiometric conversions and intervening metabolic reactions. Sci Rep 2015;5:16009.
[169] Kim SM, Pena MI, Moll M, Bennett GN, Kavraki LE. A review of parameters and heuristics for guiding metabolic pathfinding. J Cheminformatics 2017;9:51.
[170] Wang L, Dash S, Ng CY, Maranas CD. A review of computational tools for design and reconstruction of metabolic pathways. Synth Syst Biotechnol 2017;2:243–52.
[171] Jamil IN, Remali J, Azizan KA, Nor Muhammad NA, Arita M, et al. Systematic multi-omics integration (MOI) approach in plant systems biology. Front Plant Sci 2020;11:944.
[172] Buchweitz LF, Yurkovich JT, Blessing C, Kohler V, Schwarzkopf F, et al. Visualizing metabolic network dynamics through time-series metabolomic data. BMC Bioinf 2020;21:130.
[173] Pandey V, Hadadi N, Hatzimanikatis V. Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models. PLoS Comput Biol 2019;15:e1007036.
[174] Caretto S, Linsalata V, Colella G, Mita G, Lattanzio V. Carbon fluxes between primary metabolism and phenolic pathway in plant tissues under stress. Int J Mol Sci 2015;16:26378–94.
[175] Zhang Y, Fernie AR. Metabolons, enzyme-enzyme assemblies that mediate substrate channeling, and their roles in plant metabolism. Plant Communications 2020. https://doi.org/10.1016/j.xplc.2020.100081.
[176] Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. Nat Protoc 2019;14:639–702.
[177] Blimkie T, Lee AHY, Hancock RE. MetaBridge: an integrative multi-omics tool for metabolite-enzyme mapping. Curr Protoc Bioinform 2020;70:e98.
[178] Saito K, Dixon RA, Willmitzer L. Plant metabolomics (Biotechnology in Agriculture and Forestry 57) Berlin 2006 Springer-Verlag Heidelberg
[179] Saito K, Hirai MY, Yonekura-Sakakibara K. Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'. Trends Plant Sci 2008;13:36–43.
[180] Rai A, Rai M, Kamochi H, Mori T, Nakabayashi R, et al. Multiomics-based characterization of specialized metabolites biosynthesis in Cornus officinalis. DNA Res 2020;27:dsaa009.
[181] Sévin DC, Fuhrer T, Zamboni N, Sauer U. Nontargeted in vitro metabolomics for high-throughput identification of novel enzymes in Escherichia coli. Nat Methods 2017;14:187–94.
[182] Keurentjes JJB, Fu J, de Vos CHR, Lommen A, Hall RD, et al. The genetics of plant metabolism. Nat Genet 2006;38:842–9.
[183] Li K, Wen W, Alseekh S, Yang X, Guo H, et al. Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. Plant J 2019;99:216–30.
[184] Morreel K, Goeminne G, Storme V, Sterck L, Ralph J, et al. Genetical metabolomics of flavonoid biosynthesis in Populus: a case study. Plant J 2006;47:224–37.
[185] Schauer N, Semel Y, Roessner U, Gur A, Balbo I, et al. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 2006;24:447–54.
[186] Fang C, Fernie AR, Luo J. Exploring the diversity of plant metabolism. Trends Plant Sci 2019;24:83–98.
[187] Fang C, Luo J. Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. Plant J 2019;97:91–100.
[188] Quell JD, Römisch-Margl W, Colombo M, Krumsiek J, Evans AM, et al. Automated pathway and reaction prediction facilitates in silico identification of unknown metabolites in human cohort studies. J Chromatogr B 2017;1071:58–67.
[189] Prosser GA, Larrouy-Maumus G, de Carvalho LPS. Metabolomic strategies for the identification of new enzyme functions and metabolic pathways. EMBO Rep 2014;15:657–69.
[190] Vanholme R, Cesarino I, Rataj K, Xiao Y, Sundin L, et al. Caffeoyl shikimate esterase (CSE) is an enzyme in the lignin biosynthetic pathway in *Arabidopsis*. Science 2013;341:1103–6.
[191] Prosser JI. Dispersing misconceptions and identifying opportunities for the use of'omics' in soil microbial ecology. Nat Rev Microbiol 2015;13:439–46.
[192] Liu Y, Mrzic A, Meysman P, De Vijlder T, Romijn EP, et al. MESSAR: automated recommendation of metabolite substructures from tandem mass spectra. PLoS ONE 2020;15:e0226770.
[193] Waller TC, Berg JA, Lex A, Chapman BE, Rutter J. Compartment and hub definitions tune metabolic networks for metabolomic interpretations. Gigascience 2020;9:giz137.
[194] Rodriguez-Esteban R. Biocuration with insufficient resources and fixed timelines. Database 2015;2015:bav116.
[195] Naithani S, Gupta P, Preece J, Garg P, Fraser V, et al. Involving community in genes and pathway curation. Database 2019;2019:bay146.
[196] Fell D. Understanding the control of metabolism. London: Portland Press; 1997. p. 301.
[197] Sauro HM. Control and regulation of pathways via negative feedback. J R Soc Interface 2017;14:20160848.
[198] Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, et al. Metscape: a Cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics 2010;26:971–3.
[199] Tziotis D, Hertkorn N, Schmitt-Kopplin P. Kendrick-analogous network visualisation of ion cyclotron resonance Fourier transform mass spectra: improved options for the assignment of elemental compositions and the classification of organic molecular complexity. Eur J Mass Spectrom 2011;17:415–21.
[200] Doerfler H, Sun X, Wang L, Engelmeier D, Lyon D, et al. *mzGroupAnalyzer*-Predicting pathways and novel chemical structures from untargeted high-throughput metabolomics data. PLoS ONE 2014;9:e96188.
[201] Allard PM, Péresse T, Bisson J, Gindro K, Marcourt L, et al. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. Anal Chem 2016;88:3317–23.

[202] Moritz F, Kaling M, Schnitzler J-P, Schmitt-Kopplin P. Characterization of poplar metabotypes via mass difference enrichment analysis. Plant Cell Environ 2017;40:1057–73.

[203] Olivon F, Elie N, Grelier G, Roussi F, Litaudon M, et al. MetGem software for the generation of molecular networks based on the t-SNE algorithm. Anal Chem 2018;90:13900–8.

[204] Beauxis Y, Genta-Jouve G. MetWork: a web server for natural products anticipation. Bioinformatics 2019;35:1795–6.

[205] Naake T, Fernie AR. MetNet: Metabolite network prediction from high-resolution mass spectrometry data in R aiding metabolite annotation. Anal Chem 2019;91:1768–72.