Essays in Biochemistry (2023) EBC20230019
https://doi.org/10.1042/EBC20230019

PORTLAND PRESS

Review Article

Computational methods for processing and interpreting mass spectrometry-based metabolomics

Leonardo Perez de Souza[1] **and Alisdair R. Fernie**[1,2]

[1]Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany; [2]Center for Plant Systems Biology and Biotechnology, 4000 Plovdiv, Bulgaria

Correspondence: Alisdair R. Fernie (Fernie@mpimp-golm.mpg.de)

Metabolomics has emerged as an indispensable tool for exploring complex biological questions, providing the ability to investigate a substantial portion of the metabolome. However, the vast complexity and structural diversity intrinsic to metabolites imposes a great challenge for data analysis and interpretation. Liquid chromatography mass spectrometry (LC-MS) stands out as a versatile technique offering extensive metabolite coverage. In this mini-review, we address some of the hurdles posed by the complex nature of LC-MS data, providing a brief overview of computational tools designed to help tackling these challenges. Our focus centers on two major steps that are essential to most metabolomics investigations: the translation of raw data into quantifiable features, and the extraction of structural insights from mass spectra to facilitate metabolite identification. By exploring current computational solutions, we aim at providing a critical overview of the capabilities and constraints of mass spectrometry-based metabolomics, while introduce some of the most recent trends in data processing and analysis within the field.

Introduction

Omics technologies have revolutionized biological sciences by providing comprehensive insights into the intricate molecular workings of living systems. Among these, metabolomics stands out - being the closest link to phenotype and thereby representing a powerful tool for understanding cellular dynamics.

The vast chemical diversity of metabolites, with in excess of a million distinct chemical structures thought to be present across extant organisms [1,2], requires advanced analytical methods, and both the identification of unknown metabolites and the comprehensive coverage of the metabolome remains a great challenge. The diversity of structures exhibiting widely diverse physicochemical properties, complicates the achievement of comprehensive coverage using a single analytical technique. In contrast with nucleic acids and proteins that present regular structures built from the linear arrangement of a finite set of building blocks, the lack of structural regularity among metabolites results in a vastly larger search space of plausible structures. In fact, even considering only simple organic molecules such as alkanes, with a general molecular formula of C_nH_{2n+2}, a 20-carbon molecular formula can be assigned to over 3.3 million different stereoisomers [3].

Few analytical tools have the necessary features to tackle this challenging task. Among them, the integration of ultra-high-performance liquid chromatography (UHPLC) and high-resolution mass spectrometry provides the best coverage [4]. It provides high sensitivity to detect lowly abundant compounds, high dynamic range to cover a broad range of concentrations, high resolving power to allow quantification of multiple compounds in a complex matrix, and enough structural information to retrieve putative identities of measured metabolites. Additionally, LC-MS is the most flexible of all available technical platforms

Received: 15 September 2023
Revised: 10 November 2023
Accepted: 15 November 2023

Version of Record published:
24 November 2023

Essays in Biochemistry (2023) EBC20230019
https://doi.org/10.1042/EBC20230019

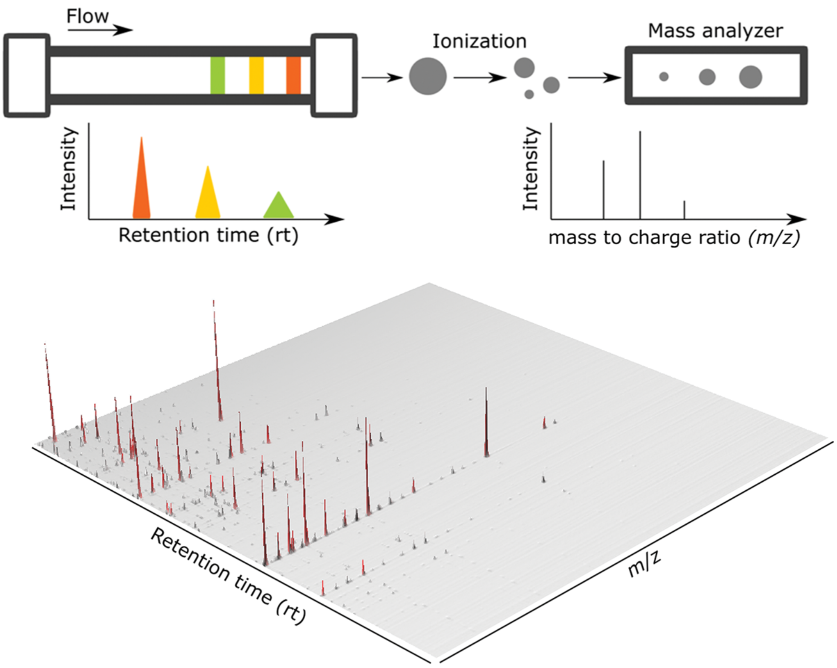

Figure 1. Exemplary overview of LC-MS mechanism and resulting mass chromatogram
In the chromatographic separation (top left) a mixture of compounds, represented by the different colors, elutes through the column. The differential interactions of the compounds with the mobile and stationary phases within the column separates them based on their retention time as represented by the differentially colored peaks in the chromatogram. All the efflux of the chromatographic system is ionized and ions are separated based on their mass to charge ratio (*m/z*) within the mass spectrometer (top right). The resulting multidimensional mass chromatogram is represented in the lower panel.

and can be adjusted for the detection of a wide variety of different compounds. Still, it is important to stress that no sole technique can cover the whole metabolome, and that much of what we can currently measure remains unidentified. Next, we discuss computational approaches that have been fundamental in assisting the interpretation of the complex datasets generated by LC-MS metabolomics experiments. They focus largely on two essential topics for exploratory metabolomics experiments. First, the unbiased detection of as many metabolites as possible. Second, the annotation of the large proportion of the metabolome that remains unidentified, colloquially termed the ‘dark’ metabolome.

LC-MS data overview

In a typical LC-MS run, compounds within a mixture differentially interact with the phase within a chromatographic column leading to their separation at distinct retention times (rt - glossary) (Figure 1). Analytes eluting off the column, are directly ionized via electrospray ionization (ESI). This process usually yields a mixture of the protonated molecular ion (depronated in negative ionization mode), as well as common adducts (glossary) formed with components of the mobile phase and environment (e.g. formic acid, acetic acid, Na^+, and K^+), and *in source* fragments (glossary). Subsequently, these ions make their way into the evacuated analyzer where they are separated based on their mass to charge ratio (*m/z*), prior to being registered by a detector. An additional step of fragmentation, termed tandem MS, is usually included inside the mass spectrometer providing second order (MS^2) spectra. MS^2 spectra provide the fragmentation pattern of the selected precursor ion, which is of great interest for structural elucidation. The resulting mass chromatograms emerge as highly complex multi-dimensional datasets, necessitating sophisticated computational strategies to fully leverage the biological insights attainable from them.

Data processing tools

Traditionally, manual data processing involves extensive and time-consuming assessment of the mass chromatograms. *In source* fragments and adducts must be identified and grouped for the selection and quantification of a single representative ion per analyte. This is the first computationally intensive process, and has been largely automated with the help of signal processing tools, and implementation of heuristic rules previously used by the analyst to classify detected signals. Several free data processing tools are available whose primary function is to automatically detect mass features generating well defined chromatographic peaks, and compare peak intensities across multiple samples, as a proxy for metabolite concentration. Among some of the most popular are XCMS [5], mzMine [6,7], OpenMS [8], and MS-DIAL [9]. For a more extensive list of the myriad of tools for this purpose we refer the reader to the more comprehensive reviews of Perez de Souza et al. [10] and Misra [11].

The automation of this peak processing step undoubtedly improves the unbiased analysis of the data but imposes some new challenges. *In source* fragments and adducts generate multiple redundant signals in the mass chromatograms. This is a major source of multicollinearity (glossary) in metabolomics datasets that should be removed to improve data analysis. There are several tools, based on correlation over narrow rt windows and automatic detection of relevant *m/z* differences, that can assist this annotation and often integrate directly with the aforementioned tools. These include CAMERA [12] and MS-DIAL [9]. However, a considerable effort in manual data curation is still required to access the results.

Processing optimization and curation

Another consequence of unbiased signal processing is the large proportion of poorly integrated signals. Automated processing of such complex datasets usually involves tradeoff between processing method sensitivity and the quality of the integrated data. Therefore, selecting the multiple parameters available for all the aforementioned tools is far from trivial. A few tools have tried to implement systematic methods for parameter optimization [13,14], and more recently, a dedicated processing tool (SLAW) was developed to provide self-optimizing processing workflows [15].

Considering how variable and sensitive data processing is, it is generally good practice to evaluate a workflow based on some previous knowledge of the sample, or some internal controls. A very interesting tool that facilitates this performance assessment is mzRAPP [16]. mzRAPP generates a benchmark subset from the analyzed dataset and uses it to return well-defined processing performance metrics.

Despite the availability of all these resources, careful curation of the processed data is strongly recommended. Such curation aims at classifying automatically processed data into poorly or well-integrated/defined peaks, removing the former from downstream analysis. Unsurprisingly, as any classification task, this can profit from modern machine learning tools. A few deep learning algorithms have been released that allow for a relatively quick dataset specific training based on user input of good and poorly integrated peaks [17–19]. These algorithms allow for the removal of such features with great success.

Second order spectra

Fragmentation patterns, together with accurate mass, provide the most important information regarding compound identification in mass spectrometry. It is important to highlight that *in source* fragmentation is still considered mostly a detrimental factor, since it has low reproducibility and increases dataset complexity. Still, it can provide some structural fragmentation, and there are attempts to explore it [20]. The use of tandem MS experiments is a preferred strategy though, providing more reproducible and interpretable data.

A traditional tandem MS experiment is performed inside the mass spectra, with target ions being isolated and fragmented. This traditional setup is often referred to as data dependent acquisition (DDA), given the relationship between the fragments (also known as daughter ions) and the original ion (also known as parental ion) is well stablished (Figure 2A). The isolation of each individual ion from the MS1 spectra demands scanning time from the detector. Therefore, the comprehensive coverage of all MS1 signals is normally impossible. Most metabolomics-oriented methods use the strategies of either automatically selecting TopN most intense ions on each MS1 scan event for immediate MS2 fragmentation in successive scans, or they perform an in-depth fragmentation study following repeated injections of representative samples. The former sacrifices coverage, while the latter sacrifices experiment throughput, since it demands extensive manual tuning of successive runs and posterior integration of data across different raw files to characterize a single sample.

Data independent analysis (DIA), in contrast, provides comprehensive fragmentation of all MS^1 ions, by using very large or no isolation windows, and fragmenting a mixture of multiple MS^1 ions (Figure 2B). The result is a highly multiplexed MS^2 spectra, with no clear precursor-product ion relationships, demanding considerably more computational effort to re-construct these relationships, and resulting in overall lower quality second order spectra. Most of the popular data processing tools, are currently developing and integrating alternatives to process DIA data, highlighting MS-DIAL that was developed with a particular focus toward that [21].

In an ideal scenario, in-depth analysis to obtain DDA data for most of the ions is the preferred approach. However, it is important to recognize the limitations imposed by the reality of most metabolomics facilities and experiments.

Essays in Biochemistry (2023) EBC20230019
https://doi.org/10.1042/EBC20230019

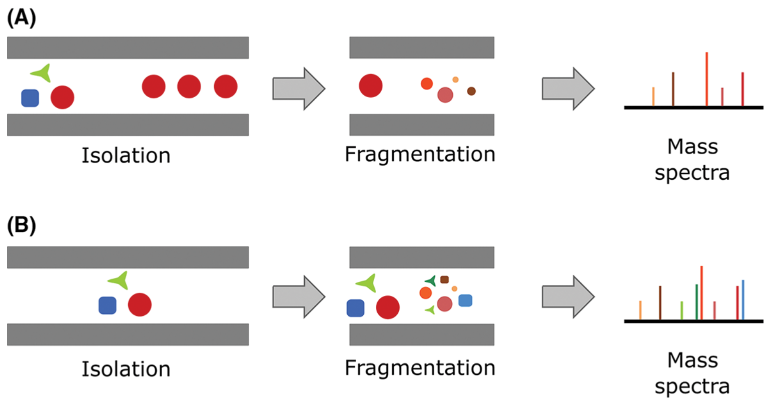

Figure 2. Diagram representing tandem MS experiments

Diagram representing tandem MS experiments in (**A**) DDA, and (**B**) DIA modes. The different geometric forms in the isolation chart represent a mixture of different ion populations and how they are transferred through the device on the different data acquisition modes. In (A) only the ion population represented by the red circle make their way through and is fragmented. All smaller fragments represented in the fragmentation chart are product ions of the same isolated precursor. In (B), all the different ion populations make their way through the isolation device and are fragmented. The smaller fragments represent product ions of one of the three different precursor ion populations within the fragmentation chart. The resulting mass spectra registered are visually similar, however, for (B) the precursor/product ion relationships are unknown.

Equipment time availability and the need for high throughput in order to run large association studies, for example, can impose significant limitations. In such situations, approaches such as TopN and DIA, provide the flexibility necessary to acquire valuable structural information while minimizing extensive fragmentation studies.

Metabolite annotation – spectral database search

After completing the initial data processing, the path to follow depends significantly on the goal of the experiment. Most general exploratory experiments, seek to comprehensively characterize the metabolic composition of the samples. In the following sections, we will focus into some essential concepts and tools that facilitate such untargeted and unbiased global annotation.

The most widely adopted approach for metabolite annotation relies on MS^2 spectral database searches. At the core of such methods lies the concept of quantifying the similarity between an experimental spectrum and that of an authentic standard stored within a spectral database [22]. The most commonly used score is the dot product, or cosine similarity [23]. This similarity measure computes the cosine of the angle between the two vectors representing each spectrum, ranging from 0 to 1, when both spectra are identical. It serves as an efficient metric to compare spectra with relatively high degree of purity. An alternative, often useful for contaminated spectra commonly found in metabolomics experiments, is the use of the reverse dot product, which omits from the calculation all peaks that do not match in the query spectra [22,23].

Although spectral database search appears straightforward, it faces significant limitations from two key factors. First, isomers frequently produce remarkably similar, often indistinguishable spectra. Second, the search is constrained by the relatively limited number of compounds available within spectral databases. The former challenge can be considerably amended by incorporating retention time information for dereplication of identical spectra, despite the great challenge that is to map retention times across different platforms. Quantitative structure-retention relationship (QSRR) models constitute a promising strategy to predict retention times and elution order of different compounds across systems [24]. Incorporating QSRR-based predictions as an additional scoring has a great potential aiding the selection between equally likely candidates based solely on spectrum similarity.

Metabolite annotation – network analysis

Perfect spectral database matches are not only rare due to the intrinsic variations of data generation, but also due to the huge diversity of the metabolome, for which only a minute proportion is represented with spectroscopic data in standard databases. Unsurprisingly, one area of metabolomics research that has observed growing interest in the recent years, is without question that of new compound elucidation. As highlighted by many authors, the vast majority of the detectable metabolome remains uncharacterized. This represents a huge untapped resource as many large-scale experiments rely on metabolite annotations to drive biological and mechanistic conclusions, often leaving aside strong associations with unknown compounds. Therefore, several alternative metabolite annotation approaches try to expand the reach of mass spectrometry-based annotation beyond previously characterized molecules.

Molecular networking is one such approach that became particularly popular since its introduction [25] and the later implementation of the GNPS platform for data analysis and community sharing of the annotated results [26]. The core idea behind molecular networking is again based on similarity scores, but here they are used to stablish connections between unknown experimental spectra that share scores above a certain threshold. The interpretation of these networks under the assumption that similar structures generate similar spectra allows for the propagation of a few characterized metabolites throughout the dataset. Regarding the similarity score, molecular networking traditionally implements a modified dot product that better captures similarities between spectra from compounds that undergo small functional changes [25,27]. This is achieved by accounting not only for perfect matches, but also for pairs of fragments that differ in the same m/z as the precursor ions.

Since the popularization of molecular networking several sophisticated network-based approaches for metabolite annotation have been developed extending these ideas. Many of these, such as the use of precursor mass differences to infer biochemical transformations [28], and *in silico* fragmentation are highly complementary [29], and some have been integrated within the GNPS environment through MolNetEnhancer [30]. Others remain as stand-alone tools, but integrating multiple network layers. NetID [31] provides a global network optimization approach using linear programming that incorporates information regarding MS2 spectra similarity, precursor mass shift, and retention time shifts. This strategy allows for the identification of putatively related compounds whilst also providing information regarding the biochemical relationships and identifying putative *in source* fragments and adducts. MetDNA [32,33] is in its second iteration, and similarly combines spectra similarity, knowledge based metabolic reactions, and correlations to annotate both putative compounds and potential adducts and fragments.

Establishing these connections between metabolites is also very useful from a data analysis and biological interpretation perspective. Many pathway enrichment tools such as metabolite set enrichment analysis, rely on pathway databases and therefore are mostly constrained to central metabolism. However, most of the metabolic diversity that characterizes the metabolome is not captured by such pathway databases, often being species specific or at least of limited distribution. Similarity and correlation networks can be used to implement enrichment analysis with little to no information regarding metabolite identities [34].

Metabolite annotation – *in silico* structure prediction

In silico structural predictions for *de novo* structural elucidation has long been a goal of metabolomics to overcome the limited metabolome coverage of spectral library databases. Recent developments in computational power, access to large high-quality datasets and machine learning, have paved the way to exciting developments in the past few years. In this scenario two contrasting conceptual frameworks have been explored, the first is translating molecular structures into spectra, while the second revolves around precisely the reverse process of translating mass spectra into molecular structures that are not yet present in databases [35].

Chemical structure databases such as PubChem [36] and Coconut [37] provide a vastly larger coverage of known chemical structures in relation to spectral databases, with the downside of lacking spectral information for most of them. Initial approaches to leverage the much higher coverage of these structural databases involved the development of tools that attempt to model the processes happening in the mass spectrometer to predict fragmentation solely from a chemical structure [38]. *In silico* fragmentation tools rely on diverse approaches including the combinatorial fragmentation of molecular bonds, and application of heuristic rules of fragmentation. Popular tools such as MAGMa+, MetFrag and MS-Finder [39–41] rely on different implementations of either one or a combination of both these approaches. Molecular fragmentation based on quantum chemistry calculations has also been explored as a possible solution for *in silico* fragmentation, and are available with tools such as QCMS^2 [42] and ChemFrag [43]. However, these methods demand considerable computational power which makes them less viable for the scale demanded by metabolomics studies. Machine learning models have also been used for *in silico* fragmentation. Some of the pioneering work on this field has culminated into the CFM-ID model [44]. CFM-ID predicts break tendencies of possible fragments and translates them into probabilities to grasp the competitive dynamics among potential breaks within the same molecule [38]. More recently, GRAFF-MS model has shown promising improvements with lower prediction error and considerably faster runtime using graph neural networks [45].

Alternatively, the prediction of molecular fingerprints from mass spectra has also been explored [46,47]. A molecular fingerprint is a long vector that encodes structural features of a molecule and it is a popular tool in chemoinformatic facilitating calculations involving molecular structures [48]. One of the best performing tools for metabolite annotation available in the past years, CSI:FingerID [47], integrated in the Sirius platform [49], uses predicted molecular formulas and fragmentation trees to generate a fingerprint and match this against structural databases. The main limitation of these approaches is that they are still restricted to the now much larger structural databases. Attempts to amend this dependency have shown promising results such as the integration of *in silico* structure database generation represented by the COSMIC workflow [50], however it still faces computational limitations once the chemical space explored becomes too large.

Recently, attempts to generate models capable of translating mass spectra into structures, dispensing the need of any database, seem to have finally taken off with the release of MSNovelist, Spec2Mol, MassGenie and MS2Mol [51–54] in the past 3 years. These methods have profited significantly from the development of deep learning methods to molecule generation such as the pioneering work by Gómez-Bombarelli et al. [55]. Perhaps it comes with no surprise that the machine learning models incorporated in these tools are also widely employed in natural language processing. Mass spectra analysts have long interpreted the fragmentation patterns in a similar way to a language. In fact, text mining has inspired spectra interpretation tools in the past. MS2LDA [56] for instance facilitates the extraction of fragmentation patterns linked to specific functional groups by leveraging latent Dirichlet allocation, originally designed to break down text documents into topics through the analysis of co-occurring words. Latest break throughs in natural language processing are likely to seamlessly translate into the field of mass spectral interpretation, significantly enhancing the efficiency of *de novo* structural prediction in the years to come.

Outlook

Mass spectrometry is a continuously evolving technique that has seen remarkable advancements in recent years. These advancements have resulted in significant enhancements in mass resolution, sensitivity, and detector speeds. Consequently, mass spectrometry now enables more informative and comprehensive data collection than ever before. Concomitantly, improved computational power, access to data and developments in machine learning are bringing unprecedented advancements in some of the most challenging aspects of metabolomics data analysis.

Metabolomics has become essential to a variety of different fields and it is adopted by researcher with widely different background. User friendliness has been a determining factor towards the popularization of these computational tools through intuitive graphical user interfaces, cross platform integration, self-optimizing pipelines and quality assessment tools. Data annotation platforms such as Sirius and GNPS, and processing tools like XCMS, mzMine, OpenMS, and MS-DIAL are prime examples of how user experience is a fundamental aspect of bringing these good ideas into everyday practice.

Metabolomics experiments are typically conducted to address highly targeted research inquiries, resulting in data that often exceeds the immediate scope of the study. Consequently, a substantial portion of the generated data remains largely unexplored. The emerging trend within the scientific community to openly share extensive raw datasets, along with the development of repository-scale data analysis pipelines, holds great promise and excitement for the future. This forward-looking approach is expected to provide novel insights into the metabolome and its intricate interactions in the years ahead.

Glossary

Adduct: An ion formed by interaction of two species within the ion source, often an analyte with components of the chromatographic solvent system, forming a single ion containing both constituents.

In source fragmentation: analyte fragmentation in the ionization source. It usually occurs as a byproduct of electrospray ionization.

Multicollinearity: Feature of a dataset where many variables are highly correlated. It usually results in a detrimental effect on statistical inferences.

Retention time (rt): The time a compound takes from the beginning of the run until it reaches the detector.

Summary

- Recent advances in computational metabolomics allow for the processing extensive and intricate datasets, establishment of complex relationships and application of heuristic rules to effectively aggregate redundant signals.
- Machine learning provides efficient and significantly faster alternatives for data quality curation, eliminating poorly integrated signals from data processing.
- The advancement of spectral similarity scores and innovative applications across samples is facilitating the exploration of increasingly intricate associations and the propagation the scarce knowledge about annotated metabolites through similarity networks.
- *In silico* structure prediction is improving at a fast pace. The representation of molecular structures as fingerprints allow for sophisticated calculations establishing relationships between spectra and structure in both directions.
- New machine learning models inspired by language processing are promising tools to generate putative structures directly from mass spectra.

Competing Interests

The authors declare that there are no competing interests associated with the manuscript.

Funding

L.P.S acknowledges the Bundesministerium für Bildung und Forschung (BMBF) [grant numbers 01DH21018, PROSIT PRIMA, and PSMOPF078] for financial support. A.R.F. acknowledges the European Union's Horizon 2020 research and innovation programme, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620) and the BG05M2OP001-1.003-001-C01 project, financed by the European Regional Development Fund through the Bulgarian 'Science and Education for Smart Growth' Operational Programme. Both authors acknowledge the support by the Max Planck Society.

Open Access

Open access for this article was enabled by the participation of Max Planck Digital Library in an all-inclusive *Read & Publish* agreement with Portland Press and the Biochemical Society.

Author Contribution

Both authors contributed with reviewing the literature and writing the manuscript.

Abbreviations

DDA, data dependent acquisition; DIA, data independent analysis; ESI, electrospray ionization; LC-MS, liquid chromatography mass spectrometry; QSRR, quantitative structure-retention relationship; UHPLC, ultra-high-performance liquid chromatography.

References

1 Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J., Ewald, J. et al. (2021) Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756, https://doi.org/10.1038/s41592-021-01197-1
2 Aharoni, A., Goodacre, R. and Fernie, A.R. (2023) Plant and microbial sciences as key drivers in the development of metabolomics research. *Proc. Natl. Acad. Sci.* **120**, e2217383120, https://doi.org/10.1073/pnas.2217383120
3 Paton, R.S. and Goodman, J.M. (2007) Exploration of the accessible chemical space of acyclic alkanes. *J. Chem. Inf. Model.* **47**, 2124–2132, https://doi.org/10.1021/ci700246b
4 Perez de Souza, L., Alseekh, S., Scossa, F. and Fernie, A.R. (2021) Ultra-high-performance liquid chromatography high-resolution mass spectrometry variants for metabolomics research. *Nat. Methods* **18**, 733–746, https://doi.org/10.1038/s41592-021-01116-4
5 Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R. and Siuzdak, G. (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.* **78**, 779–787, https://doi.org/10.1021/ac051437y
6 Pluskal, T., Castillo, S., Villar-Briones, A. and Orešič, M. (2010) MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **11**, 395, https://doi.org/10.1186/1471-2105-11-395

7 Schmid, R., Heuckeroth, S., Korf, A., Smirnov, A., Myers, O., Dyrlund, T.S. et al. (2023) Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* **41**, 447–449, https://doi.org/10.1038/s41587-023-01690-2

8 Röst, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F. et al. (2016) OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748, https://doi.org/10.1038/nmeth.3959

9 Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K. et al. (2015) MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523, https://www.nature.com/articles/nmeth.3393#supplementary-information, https://doi.org/10.1038/nmeth.3393

10 Perez de Souza, L., Naake, T., Tohge, T. and Fernie, A.R. (2017) From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web resources for mass spectral plant metabolomics. *GigaScience* **6**, 1–20, https://doi.org/10.1093/gigascience/gix037

11 Misra, B.B. (2021) New software tools, databases, and resources in metabolomics: updates from 2020. *Metabolomics* **17**, 49, https://doi.org/10.1007/s11306-021-01796-1

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

13 Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F. et al. (2015) IPO: a tool for automated optimization of XCMS parameters. *BMC Bioinformatics* **16**, 118, https://doi.org/10.1186/s12859-015-0562-8

14 McLean, C. and Kujawinski, E.B. (2020) AutoTuner: high fidelity and robust parameter selection for metabolomics data processing. *Anal. Chem.* **92**, 5724–5732, https://doi.org/10.1021/acs.analchem.9b04804

15 Delabriere, A., Warmer, P., Brennsteiner, V. and Zamboni, N. (2021) SLAW: a scalable and self-optimizing processing workflow for untargeted LC-MS. *Anal. Chem.* **93**, 15024–15032, https://doi.org/10.1021/acs.analchem.1c02687

16 El Abiead, Y., Milford, M., Salek, R.M. and Koellensperger, G. (2021) mzRAPP: a tool for reliability assessment of data pre-processing in non-targeted metabolomics. *Bioinformatics* **37**, 3678–3680, https://doi.org/10.1093/bioinformatics/btab231

17 Kantz, E.D., Tiwari, S., Watrous, J.D., Cheng, S. and Jain, M. (2019) Deep neural networks for classification of LC-MS spectral peaks. *Anal. Chem.* **91**, 12407–12413, https://doi.org/10.1021/acs.analchem.9b02983

18 Gloaguen, Y., Kirwan, J.A. and Beule, D. (2022) Deep learning-assisted peak curation for large-scale LC-MS metabolomics. *Anal. Chem.* **94**, 4930–4937, https://doi.org/10.1021/acs.analchem.1c02220

19 Stancliffe, E. and Patti, G.J. (2023) PeakDetective: a semisupervised deep learning-based approach for peak curation in untargeted metabolomics. *Anal. Chem.* **95**, 9397–9403, https://doi.org/10.1021/acs.analchem.3c00764

20 Seitzer, P.M. and Searle, B.C. (2019) Incorporating in-source fragment information improves metabolite identification accuracy in untargeted LC–MS data sets. *J. Proteome Res.* **18**, 791–796, https://doi.org/10.1021/acs.jproteome.8b00601

21 Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H. et al. (2020) A lipidome atlas in MS-DIAL 4. *Nat. Biotechnol.* **38**, 1159–1163, https://doi.org/10.1038/s41587-020-0531-2

22 Cai, Y., Zhou, Z. and Zhu, Z.-J. (2023) Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *TrAC, Trends Anal. Chem.* **158**, 116903, https://doi.org/10.1016/j.trac.2022.116903

23 Stein, S.E. and Scott, D.R. (1994) Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866, https://doi.org/10.1016/1044-0305(94)87009-8

24 Bach, E., Schymanski, E.L. and Rousu, J. (2022) Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. *Nat. Machine Intelligence* **4**, 1224–1237, https://doi.org/10.1038/s42256-022-00577-2

25 Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D. et al. (2012) Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci.* **109**, E1743–E1752, https://doi.org/10.1073/pnas.1203689109

26 Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y. et al. (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837, https://doi.org/10.1038/nbt.3597

27 Bittremieux, W., Schmid, R., Huber, F., van der Hooft, J.J.J., Wang, M. and Dorrestein, P.C. (2022) Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. *J. Am. Soc. Mass Spectrom.* **33**, 1733–1744, https://doi.org/10.1021/jasms.2c00153

28 Naake, T. and Fernie, A.R. (2019) MetNet: metabolite network prediction from high-resolution mass spectrometry data in R aiding metabolite annotation. *Anal. Chem.* **91**, 1768–1772, https://doi.org/10.1021/acs.analchem.8b04096

29 da Silva, R.R., Wang, M., Nothias, L.-F., van der Hooft, J.J.J., Caraballo-Rodríguez, A.M., Fox, E. et al. (2018) Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, e1006089, https://doi.org/10.1371/journal.pcbi.1006089

30 Ernst, M., Kang, K.B., Caraballo-Rodríguez, A.M., Nothias, L.-F., Wandy, J., Chen, C. et al. (2019) MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **9**, 144, https://doi.org/10.3390/metabo9070144

31 Chen, L., Lu, W., Wang, L., Xing, X., Chen, Z., Teng, X. et al. (2021) Metabolite discovery through global annotation of untargeted metabolomics data. *Nat. Methods* **18**, 1377–1385, https://doi.org/10.1038/s41592-021-01303-3

32 Shen, X., Wang, R., Xiong, X., Yin, Y., Cai, Y., Ma, Z. et al. (2019) Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* **10**, 1516, https://doi.org/10.1038/s41467-019-09550-x

33 Zhou, Z., Luo, M., Zhang, H., Yin, Y., Cai, Y. and Zhu, Z.-J. (2022) Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat. Commun.* **13**, 6656, https://doi.org/10.1038/s41467-022-34537-6

34 Barupal, D.K., Fan, S. and Fiehn, O. (2018) Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. *Curr. Opin. Biotechnol.* **54**, 1–9, https://doi.org/10.1016/j.copbio.2018.01.010

35 Ebbels, T.M.D., van der Hooft, J.J.J., Chatelaine, H., Broeckling, C., Zamboni, N., Hassoun, S. et al. (2023) Recent advances in mass spectrometry-based computational metabolomics. *Curr. Opin. Chem. Biol.* **74**, 102288, https://doi.org/10.1016/j.cbpa.2023.102288

36 Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A. et al. (2015) PubChem Substance and Compound databases. *Nucleic Acids Res.* **44**, D1202–D1213, https://doi.org/10.1093/nar/gkv951

37 Sorokina, M., Merseburger, P., Rajan, K., Yirik, M.A. and Steinbeck, C. (2021) COCONUT online: collection of open natural products database. *J. Cheminformatics* **13**, 2, https://doi.org/10.1186/s13321-020-00478-9

38 Krettler, C.A. and Thallinger, G.G. (2021) A map of mass spectrometry-based in silico fragmentation prediction and compound identification in metabolomics. *Brief. Bioinform.* **22**, bbab073, https://doi.org/10.1093/bib/bbab073

39 Verdegem, D., Lambrechts, D., Carmeliet, P. and Ghesquière, B. (2016) Improved metabolite identification with MIDAS and MAGMa through MS/MS spectral dataset-driven parameter optimization. *Metabolomics* **12**, 98, https://doi.org/10.1007/s11306-016-1036-3

40 Ruttkies, C., Neumann, S. and Posch, S. (2019) Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics* **20**, 376, https://doi.org/10.1186/s12859-019-2954-7

41 Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T. et al. (2016) Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.* **88**, 7946–7958, https://doi.org/10.1021/acs.analchem.6b00770

42 Cautereels, J., Claeys, M., Geldof, D. and Blockhuys, F. (2016) Quantum chemical mass spectrometry: ab initio prediction of electron ionization mass spectra and identification of new fragmentation pathways. *J. Mass Spectrom.* **51**, 602–614, https://doi.org/10.1002/jms.3791

43 Schüler, J.-A., Neumann, S., Müller-Hannemann, M. and Brandt, W. (2018) ChemFrag: chemically meaningful annotation of fragment ion mass spectra. *J. Mass Spectrom.* **53**, 1104–1115, https://doi.org/10.1002/jms.4278

44 Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R. and Wishart, D.S. (2021) CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* **93**, 11692–11700, https://doi.org/10.1021/acs.analchem.1c01465

45 Murphy, M., Jegelka, S., Fraenkel, E., Kind, T., Healey, D. and Butler, T. (2023) Efficiently predicting high resolution mass spectra with graph neural networks. *arXiv* 230111419, https://doi.org/10.48550/arXiv.2301.11419

46 Goldman, S., Wohlwend, J., Stražar, M., Haroush, G., Xavier, R.J. and Coley, C.W. (2023) Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat. Machine Intelligence*, https://doi.org/10.1038/s42256-023-00708-3

47 Dührkop, K., Shen, H., Meusel, M., Rousu, J. and Böcker, S. (2015) Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585, https://doi.org/10.1073/pnas.1509788112

48 Capecchi, A., Probst, D. and Reymond, J.-L. (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12**, 43, https://doi.org/10.1186/s13321-020-00445-4

49 Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M. et al. (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302, https://doi.org/10.1038/s41592-019-0344-8

50 Hoffmann, M.A., Nothias, L.-F., Ludwig, M., Fleischauer, M., Gentry, E.C., Witting, M. et al. (2022) High-confidence structural annotation of metabolites absent from spectral libraries. *Nat. Biotechnol.* **40**, 411–421, https://doi.org/10.1038/s41587-021-01045-9

51 Stravs, M.A., Dührkop, K., Böcker, S. and Zamboni, N. (2022) MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **19**, 865–870, https://doi.org/10.1038/s41592-022-01486-3

52 Shrivastava, A.D., Swainston, N., Samanta, S., Roberts, I., Wright Muelas, M. and Kell, D.B. (2021) MassGenie: a transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules* **11**, 1793, https://doi.org/10.3390/biom11121793

53 Butler, T.F.A., Lightheart, R., Bargh, B., Kerby, T., West, K. et al. (2023) MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. *ChemRxiv*, https://doi.org/10.26434/chemrxiv-2023-vsmpx-v3

54 Litsa, E.E., Chenthamarakshan, V., Das, P. and Kavraki, L.E. (2023) An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.* **6**, 132, https://doi.org/10.1038/s42004-023-00932-3

55 Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D. et al. (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Sci.* **4**, 268–276, https://doi.org/10.1021/acscentsci.7b00572

56 van der Hooft, J.J.J., Wandy, J., Barrett, M.P., Burgess, K.E.V. and Rogers, S. (2016) Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743, https://doi.org/10.1073/pnas.1608041113