Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

REVIEW

View Article Online
View Journal

Cite this: DOI: 10.1039/d5np00034c

Scalability of mass spectrometry-based metabolomics for natural extracts libraries exploration: current status, challenges, and opportunities

Adriano Rutz, *[a] Wout Bittremieux,[b] Robin Schmid,[cd] Olivier Cailloux, [e] Justin J. J. van der Hooft [fg] and Mehdi A. Beniddir *[h]

Received 1st May 2025

DOI: 10.1039/d5np00034c

rsc.li/npr

Covering: up to 2025

This review explores the potential of bioinformatics and chemoinformatics tools to advance the exploration of natural extracts libraries (NELs). Although metabolomics has become a term used routinely in natural product (NP) research, the field remains focused on individual molecules or small sets of compounds, which restricts scalability. This narrow focus is mirrored in the computational handling of generated data, limiting broader insights. By challenging the traditional molecule-first paradigm—a framework historically shaped by practical constraints—we present our vision of using computational approaches to unlock the full potential of NELs, now and in the future.

[a]*Institute of Molecular Systems Biology, ETH Zurich, Zürich, Switzerland. E-mail: rutz@imsb.biol.ethz.ch*
[b]*Department of Computer Science, University of Antwerp, Antwerp, Belgium*
[c]*Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Prague, Czechia*
[d]*mzio GmbH, Bremen, Germany*
[e]*LAMSADE, CNRS, Université Paris-Dauphine, Université PSL, 75016 Paris, France*
[f]*Bioinformatics Group, Wageningen University & Research, Wageningen, 6708PB, The Netherlands*
[g]*Department of Biochemistry, University of Johannesburg, Johannesburg 2006, South Africa*
[h]*Université Paris-Saclay, CNRS, BioCIS, 17 Avenue des Sciences, 91400 Orsay, France. E-mail: mehdi.beniddir@universite-paris-saclay.fr*

Adriano Rutz

Adriano Rutz is a Postdoctoral Researcher at ETH Zurich, developing computational and analytical methods in metabolomics. With a background in pharmaceutical sciences and phytochemistry, he integrates mass spectrometry, cheminformatics, and open science to explore the chemical diversity of life. After industry experience at Tradall SA (Bacardi Group), he completed his PhD from the University of Geneva. At ETH Zurich, he creates scalable, open-source tools for molecular annotation and metabolite discovery. He is a core contributor to the LOTUS initiative, a global effort to build a FAIR, community-driven knowledge base for natural products. His research fosters interdisciplinary collaboration at the intersection of data science, chemistry, and biology, driving innovation in metabolomics and beyond.

Wout Bittremieux

Wout Bittremieux is an Assistant Professor at the University of Antwerp, Belgium, where he leads a research group at the intersection of machine learning and computational mass spectrometry. His work centers on developing scalable algorithms and AI-driven approaches to interrogate large-scale metabolomics and proteomics datasets, often comprising millions to billions of mass spectra. By combining machine learning with domain expertise, his research advances the discovery of novel molecules and biological patterns, with broad applications in natural product discovery, metabolomics, and the life sciences. He also plays a leading role in the computational mass spectrometry community through active contributions to international consortia, open-source software initiatives, and standardization efforts.

Robin Schmid

Robin Schmid is Chief Scientific Officer at mzio GmbH in Bremen, Germany, where he shapes scientific strategy and community engagement for the open-source mzmine platform. With a PhD in analytical chemistry and a background in food chemistry, he specializes in computational mass spectrometry and metabolomics. Following postdoctoral work on host–microbiome interactions in Pieter Dorrestein's group at UC San Diego and research on plant specialized metabolism with Tomáš Pluskal at IOCB Prague, he founded mzio. His work focuses on developing algorithms and mzmine workflows that seamlessly integrate data analysis of mass spectrometry, ion mobility, chromatography, and imaging technologies to advance non-targeted discovery of small molecules.

Olivier Cailloux

Olivier Cailloux is a Maitre de Conférences in Computer Science at the LAMSADE laboratory of Paris-Dauphine University. He specializes in decision theory and social choice, with particular expertise in preference modeling and multicriteria decision analysis. His research interests focus on establishing the legitimacy of recommendation systems, preference aggregation in multicriteria contexts, and developing explanations for recommendations that adapt to individual user subjectivity. Drawing from formal argumentation theory, he has extended preference elicitation methodologies to voting rule selection and developed frameworks for explaining voting outcomes through concrete applications.

1. Introduction

Natural products (NPs), from different biological sources have played a key role in drug discovery.[1] Collections of solvent-derived extracts from diverse organisms, known as natural extracts libraries (NELs), are central to the systematic exploration of bioactive specialized metabolites. These libraries, typically formatted in well plates and comprising hundreds of extracts, enable high-throughput screening for novel therapeutic leads.[2] Despite the historical importance of NPs and their multiple successful drug discovery examples, *i.e.*, statins for cardiovascular diseases and taxanes for cancer, pharmaceutical companies have become increasingly reluctant to invest in NP-based drug discovery programs due to challenges such as the frequent rediscovery of known bioactives. As a result, some are more willing to share their NELs with academic institutions, which are better positioned to undertake the exploratory risks associated with NP research. Recently, the issue of rediscovery is being increasingly addressed by bioinformatics tools that efficiently process and analyze mass spectrometry (MS) and nuclear magnetic resonance (NMR) data acquired from complex extracts and assist the process of biological and structural dereplication. This led to a renewed interest in NP-inspired omics-based drug discovery. For the sake of focus and depth, this review centers on MS-based methods, though we acknowledge that NMR spectroscopy remains a powerful complementary tool for NP research, particularly valued for its reproducibility and quantification capabilities.

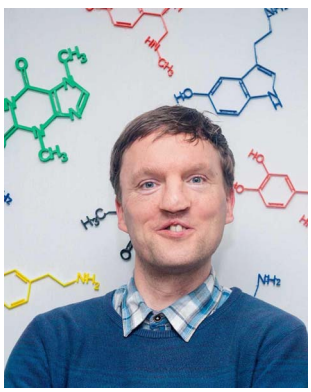
1.1 The importance of scalability in NELs exploration

As NELs datasets increase in size and complexity, they grow in power to find novel bioactives; consequently, developing approaches that support their efficient and comprehensive mining becomes a necessity to unlock their potential. Scalable NELs exploration aims to:

- Map biochemical diversity more effectively: by handling larger and more diverse libraries, scalable approaches can reveal the true breadth of biochemical diversity present in nature, uncovering rare or previously overlooked metabolites.
- Prioritize the most relevant samples and features: efficient data processing and intelligent prioritization strategies ensure that resources are focused on the most promising extracts and molecular features, accelerating the path from screening to discovery.
- Assess chemical novelty with greater precision: advanced computational tools allow for the robust assessment of chemical novelty directly from MS data, helping to identify truly novel compounds and avoid redundant efforts.

Traditionally, NP chemists have made use of available tools by adapting them to suit their research needs.[3] In practice, this has meant relying on general-purpose software or workflows that were not originally developed to address the specific complexities of exploring NELs.[4,5] While these tools have provided valuable information, they may no longer be sufficient as the scale of NELs studies increases. As the complexity and size of the datasets expand, more tailored solutions are needed: solutions that not only handle large volumes of data, but also integrate seamlessly with the workflows of NP chemists.

This review highlights the multifaceted concept of scalability in NELs exploration. We focus on the scalability of MS-based metabolomics, from data acquisition onward, assuming the availability of well-assembled extracts libraries. Readers interested in extracts library creation are referred to recent reviews and articles.[2,6–8] We compare the unique challenges and opportunities presented by NELs to those encountered in classical large-scale metabolomics studies, such as studies using human or yeast samples. Finally, we discuss the principal benefits and limitations of current computational



Justin J. J. van der Hooft

Justin J. J. van der Hooft is an Assistant Professor in Computational Metabolomics in the Bioinformatics Group at Wageningen University, NL, and an author of over 100 peer-reviewed articles in the metabolomics field. Justin is very fascinated by the ingenuity of nature in creating marvelous chemical structures. After his MSc (Molecular Sciences, 2007, WUR), and his PhD (2012) at the Biochemistry and Bioscience groups in Wageningen, he worked as a PostDoc in Glasgow, UK, and Wageningen. Since 2020, his team has been developing computational metabolomics methodologies to boost structural annotation power and to find novel bioactive metabolites and infer their source and function.

Mehdi A. Beniddir

Mehdi A. Beniddir is a Full Professor of natural products chemistry at the Faculty of Pharmacy of Paris-Saclay University. He graduated in pharmacy and received his MSc degree from Paris-Sud University in 2009. He obtained his PhD at the Institut de Chimie des Substances Naturelles (ICSN-CNRS) in 2012. He was subsequently a postdoctoral fellow at Paris-Saclay University. His research interests include the streamlined discovery of intricate natural substances from plants, marine invertebrates, and micro-organisms using prioritization strategies integrating the principles of decision theory to mimic the chemist's intuition in targeting natural substances.

metabolomics tools and strategies, specifically for processing, annotation, querying, and prioritization, and offer perspectives on future developments poised to enhance knowledge generation from NEL-based drug discovery.

1.2 Field-specific considerations: natural products *versus* biomedical and environmental metabolomics

Scalability is a term whose meaning varies depending on the field of application. In biomedical research, scalability often refers to population size: how effectively metabolomics methods can accommodate data sets derived from extensive human cohorts, where thousands of biological samples are analyzed to uncover biomarkers or metabolic signatures of diseases.[9] Although the chemical diversity in human samples may be lower compared to samples of environmental or natural sources, the complexity of the data is still considerable, influenced by factors such as circadian rhythms, diet or disease states, which contribute to temporal and concentration variations. Despite these variations, the matrix of ions across samples in biomedical studies tends to be smaller, making the scalability of such analyses more manageable compared to the vast diversity of metabolites found in NP or environmental studies.

In contrast, within environmental sciences, scalability highlights the capacity to capture both the molecular diversity and the breadth of coverage across intricate environmental matrices. At the same time, it emphasizes on the ability to handle large sample sizes, detect a broad range of metabolites, and maintain data quality despite sample variability. These divergent interpretations underscore the multifaceted nature of scalability, shaped by technical and conceptual demands.

For NELs, scalability should incorporate lessons from these disciplines but tailor its focus to unique challenges. Specifically, scalability in NELs exploration involves handling a large series of extracts while maximizing the quality and breadth of the metabolome coverage. The purpose is to facilitate data-driven decisions that prioritize the samples and metabolites to discover novel or bioactive compounds. The field of NELs research can transition similarly to more comprehensive and high-throughput strategies by using large-scale metabolomics approaches, which have proven effective in biomedical and environmental applications.

1.3 Defining scalability for NELs exploration: practical dimensions

In the context of NELs-derived metabolomics, scalability is best defined as the ability of analytical and computational tools to accommodate the analysis of large libraries of extracts and metabolites without compromising data quality or interpretability. True scalability goes beyond mere throughput; it requires tools that facilitate actionable decision-making by automating key steps, providing intuitive visualizations, and efficiently summarizing complex datasets.

Key practical dimensions of scalability in NELs exploration include:

- Time: the speed at which tools can process large datasets while maintaining computational efficiency. As the number of samples and metabolite features increases, scalable solutions must ensure timely data analysis to avoid bottlenecks.
- Quality: the extent and reliability of metabolome coverage and annotation. Scalable tools must maintain high standards for data integrity and metabolite annotation, even as sample numbers grow.
- Data retrieval: the accessibility and interpretability of processed data. Scalable systems should enable seamless querying, visualization, and exploration of large datasets, empowering chemists to prioritize extracts or metabolites for further investigation without impeding the discovery process.[10]

2. Scalability of MS-based metabolomics approaches applied to NELs

In this section, we will discuss how MS-based metabolomics could face scalability issues associated with NELs (Fig. 1), and how these could potentially be circumvented or solved.

2.1 MS data acquisition

Scaling MS-based metabolomics for the exploration of NELs requires overcoming significant challenges in throughput, data complexity, and the need to futureproof the acquisition process. Although the democratization of computational MS has made it more accessible, the sheer volume of data generated in large-scale studies requires a robust acquisition strategy to ensure data quality, metabolome coverage, and processing efficiency. Fundamentally, the scalability of NELs studies depends on obtaining high-quality data at the outset; no matter how advanced post-acquisition tools become, starting with low-quality data compromises the utility of even the most sophisticated computational approaches. Acquisition strategies (Fig. 2) must evolve to meet the needs of large and diverse datasets. Key factors in scaling MS-based metabolomics for NELs include optimizing data throughput without sacrificing quality, ensuring broad metabolome coverage, and anticipating the demands of future technological advancements. Achieving these goals requires not only improved instrumentation but also more efficient workflows capable of processing large datasets with high fidelity. One bottleneck in this context is the acquisition speed of MS data for large NELs. To address this, Linington *et al.* developed MultiplexMS, a dual-grid orthogonal multiplexing strategy that increases the throughput of untargeted MS analyses by pooling rows and columns of extract grids and computationally deconvoluting the pooled MS data into individual feature lists. While this method significantly accelerates data collection, it may introduce trade-offs such as increased computational complexity or potential ambiguity in feature assignment, particularly in highly complex mixtures. The discovery of bioactive NPs is well-suited to a pool/deconvolute approach since individual NP structures are sparsely distributed across large NELs.[11]

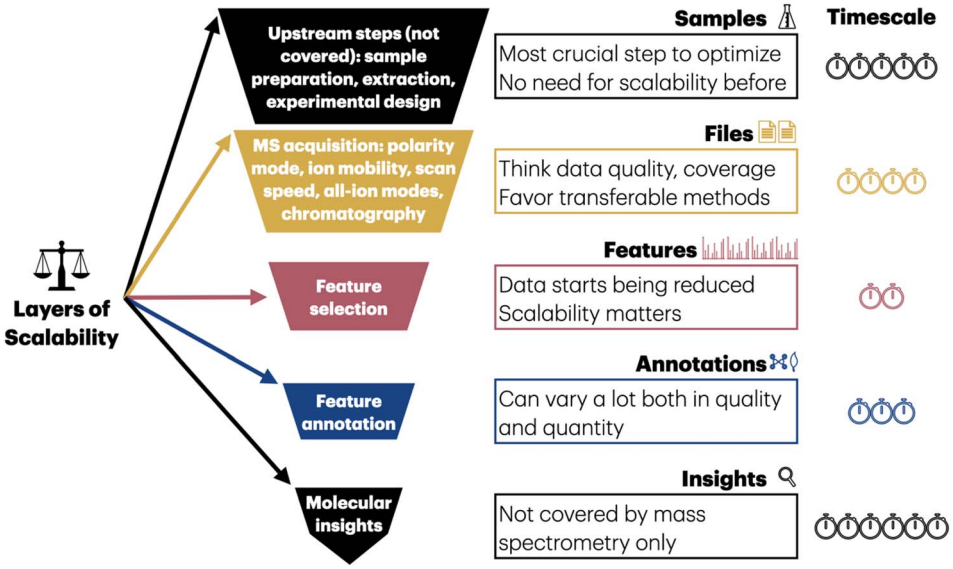


Fig. 1 An overview of key issues in scalability. Principal steps are illustrated, together with the estimated times, related objects, and their size. The experimental design and sample generation steps are illustrated as crucial, but are not covered in this article. The data generation step is time-consuming and generates hundreds to thousands of files. The feature generation step is the most efficient. It generates thousands to millions of features in a relatively short time. The annotation step is more time-intensive and the final number of annotated features is smaller. Finally, knowledge generation can take years and cover only a few molecules.

2.1.1 Resolution. In NPs metabolomics, separation techniques are essential for resolving the complex mixtures typical of natural extracts. The most used separation technique for this purpose is liquid chromatography (LC). Occurring before ionization—typically electrospray ionization (ESI) in the context of NPs—this step is essential for reducing ions co-elution,

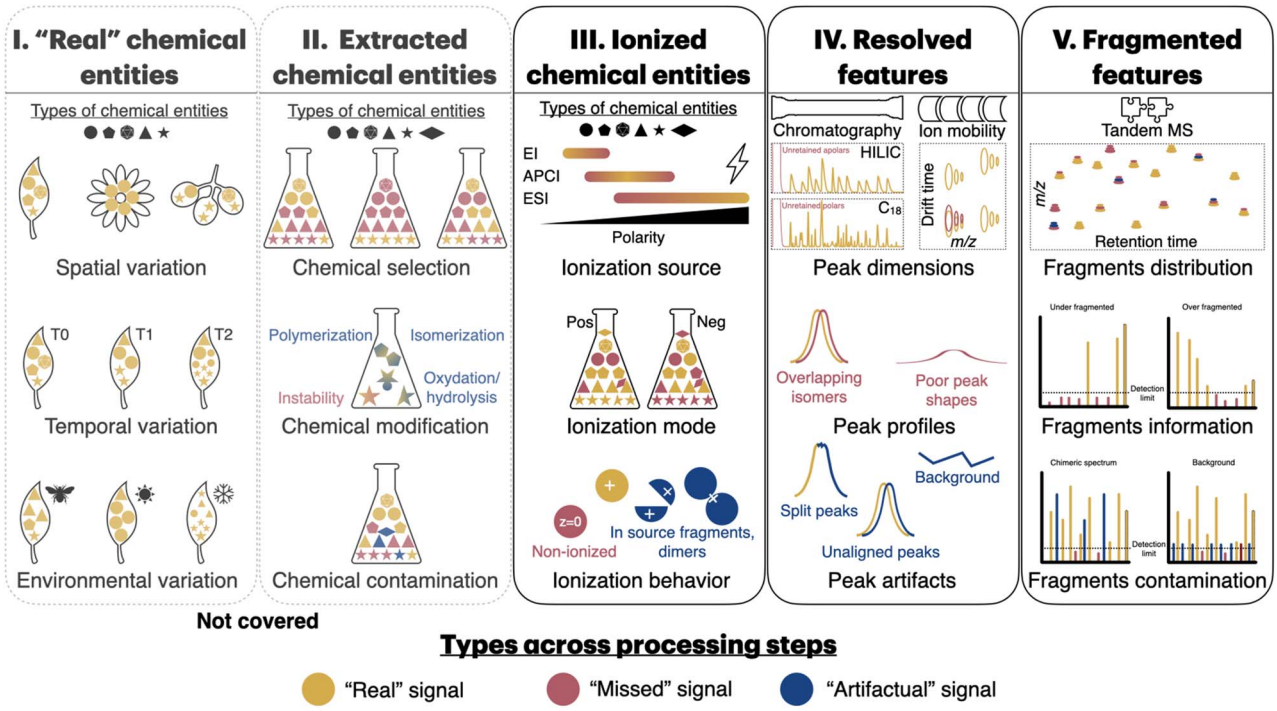


Fig. 2 Focus on mass spectrometry data acquisition steps. The selection of raw material and its extraction are described as previous steps but these are out of the scope of this article. Point (III) starts with the ionization of the molecules, where artifactual signals can be generated, and some molecules missed. Ionization is followed by resolution (IV), which can be chromatographic or ion mobility-based separation. Again, care should be taken with missed/artifactual features. Finally, the fragmentation step (V) is also depicted with some of its limitations.

mitigating ion suppression, allowing the detection of isomers, and improving the sensitivity toward low-abundance metabolites. Following ionization, ion mobility spectrometry (IMS) can be employed as an additional separation step, further distinguishing ions based on their size, shape, and charge.

2.1.1.1 Chromatography. Chromatography remains a cornerstone of metabolomics workflows, enabling the separation of metabolites from complex mixtures prior to MS analysis. Balancing high-throughput capabilities with the need for high resolution is a major challenge in studies involving NELs. Although omitting chromatography altogether may provide the highest throughput, this comes at the expense of resolution and often leads to data congestion, where an overwhelming number of unresolved metabolites complicate downstream analysis. In addition, without chromatographic separation, ion suppression effects become more pronounced, as co-eluting metabolites may interfere with the detection and quantification of target compounds. Furthermore, without proper separation, the fragmentation of isomers or structurally similar compounds becomes nearly impossible, limiting the ability to differentiate and accurately identify metabolites with similar mass-to-charge ratios.

Alternatives such as capillary electrophoresis (CE)[12] and supercritical fluid chromatography (SFC)[13] offer promising solutions to improve separation efficiency but have not been widely adopted due to limited infrastructure and the scalability challenges they present. In addition, nanoflow[14] or microflow chromatography, which operates at lower flow rates, can offer improved sensitivity to detect metabolites at low concentrations. However, its application in NELs studies remains limited by throughput constraints and the need for specialized instrumentation.

For NELs exploration to truly scale, the future of chromatography must balance speed and resolution, maximizing the number of metabolites detected per unit time, while minimizing data overlap that hinders interpretability. Novel chromatography technologies, including automated high-throughput platforms and advanced column chemistries, such as stationary phases that allow mixed mode separation,[15] will be critical in addressing current limitations. Although mixed-mode separation has already been applied to allow the separation of small polar compounds in (bio)pharmaceutical analysis,[16] or in PFAS,[17] reports on NPs are still very few.[18] Importantly, future chromatography solutions must be adaptable to a variety of natural extracts, including often disregarded polars, ensuring that the method used can handle various chemical spaces encountered in NELs studies.

2.1.1.2 Ion mobility. MS represents an important development in MS, adding a third dimension to the data by separating ions according to their shape and size, in addition to their mass-to-charge ratio. This offers a richer, more comprehensive view of the metabolome, especially for structurally similar metabolites that might otherwise be indistinguishable. However, IMS increases data complexity and processing demands, as current workflows for integrating IMS data with traditional MS are not yet standardized.

Although advances in IMS technology show great promise, IMS is still underutilized in NELs studies due to the challenges of widespread adoption and the heavy data processing burden it creates. For IMS to scale effectively in metabolomics, there is a pressing need for software tools that can process, analyze, and integrate IMS data seamlessly with conventional MS workflows. In addition, improvements in the IMS instrument design could further reduce acquisition times while maintaining the necessary resolution to differentiate structurally similar metabolites.

Future research should focus on developing IMS platforms that are faster and compatible with high-throughput NELs workflows, ensuring the incorporation of this valuable analytical dimension into large-scale studies without a significant loss of efficiency.

2.1.2 Ionization. This review focuses exclusively on ESI, as it remains the predominant ionization technique in NP metabolomics. While alternative ionization methods such as atmospheric pressure chemical ionization (APCI) or matrix-assisted laser desorption/ionization (MALDI) offer unique advantages in other contexts, their application in natural extracts analysis is currently limited and thus considered out of scope here.

Polarity selection plays a critical role in MS acquisition, directly impacting the depth of the metabolome covered. The positive ion mode remains the most widely used, offering better sensitivity for certain metabolite classes and enabling the generation of more adducts.[19] However, it also introduces challenges, such as increased noise and in-source fragmentation (ISF),[20] which can complicate the interpretation of the data. The negative ion mode, while less commonly employed, can capture a complementary range of metabolites, offering vital insights into the complete chemical space. Although polarity switching, which alternates between positive and negative modes within a single run, holds promise for improving coverage, it introduces practical issues such as reduced scan rates and potential data misalignment. No consensus has yet emerged on the best polarity strategy for NELs analysis, highlighting the need for advanced acquisition methods that could dynamically optimize polarity on a per-sample basis. Future strategies may implement adaptive polarity switching based on sample characteristics or integrate multi-modal data computationally to achieve more comprehensive metabolome profiles.

2.1.3 Fragmentation. A major challenge in LC-ESI-MS-based metabolomics is the lack of standardized collision energy settings for fragmentation, particularly when using collision-induced dissociation (CID). This variability in energy settings across laboratories and instruments leads to inconsistent fragmentation spectra, complicating both metabolite annotation and the effective use of public or commercial spectral libraries.[21]

Moreover, CID often provides only partial structural information. This limitation can result in insufficient spectral data for confident compound annotation, leaving many features unannotated or ambiguously assigned. The lack of comprehensive fragmentation information is a significant bottleneck for advancing large-scale, high-throughput NP metabolomics.

To address these challenges, there is growing interest in adopting multimodal fragmentation strategies, such as

electron-activated dissociation (EAD) and ultraviolet photodissociation (UVPD), alongside traditional CID.[22] These alternative methods can generate complementary and richer fragmentation patterns, improving structural elucidation and metabolite coverage. Integrating multimodal fragmentation into LC-ESI-MS workflows promises to enhance the depth and reliability of spectral information, ultimately facilitating more robust and scalable metabolite annotation in NELs.

2.1.3.1 DDA and DIA. The choice of data acquisition strategy is crucial in scaling MS-based metabolomics for NELs exploration. Data-dependent acquisition (DDA) has long been the standard approach, allowing for detailed MS/MS analysis of the most abundant ions. However, DDA inherently biases the data toward more abundant metabolites, often leaving low-abundance compounds underrepresented.

In contrast, data-independent acquisition (DIA) aims to capture comprehensive MS/MS data for all detectable ions, providing a more complete metabolome profile. One key challenge with DIA (or too wide window DDA) is the production of chimeric spectra, where fragments from multiple co-eluting precursor ions are combined within a single MS/MS scan. This spectral overlap complicates downstream data analysis and can hinder confident metabolite identification, especially in complex natural extracts.[23] Although DIA offers significant advantages in terms of data breadth, it comes at the cost of increased data density, placing additional strain on computational resources, and requiring advanced algorithms for spectral deconvolution and feature extraction. Remarkably, there are still no NELs exploration studies using DIA to date. As the demand grows to cover more low-abundance metabolites, acquisition strategies must evolve ensuring optimal metabolome coverage while maintaining computational feasibility.

2.1.4 Detection. The choice between time-of-flight (ToF) and Orbitrap mass analyzers has significant implications for LC-ESI-MS-based metabolomics, particularly in terms of resolution, dynamic range, scan speed, and mass accuracy. ToF analyzers provide full-spectral data with excellent mass accuracy and isotopic fidelity. Modern ToF instruments offer fast acquisition speeds (often exceeding 100 Hz) and a broad dynamic range, making them well-suited for applications requiring rapid scanning and robust quantitation, such as large-scale biomedical metabolomics and high-throughput screening.

In contrast, Orbitrap analyzers operate by trapping ions in an electrostatic field and measuring their oscillation frequencies, which are then converted to m/z values *via* Fourier transformation. Orbitrap are renowned for their ultra-high mass resolution (often surpassing 240 000 at m/z 200), and their sub-ppm mass accuracy. However, their resolution is inversely related to scan speed; achieving the highest resolution requires longer transient acquisition times, which can limit the number of data points acquired across narrow chromatographic peaks. This trade-off can be a limitation in high-throughput workflows but is less problematic when structural elucidation and confident annotation are prioritized, as in NP metabolomics.

Ultimately, the choice of instrument not only shapes the acquisition phase but also dictates the computational strategies, annotation pipelines, and even the biological questions that can be addressed in each field.

2.2 MS data processing

2.2.1 MS data formats, parsers, user libraries. Efficient and scalable management of MS data requires standardized, open, and lossless formats that facilitate interoperability, long-term storage, and compatibility with downstream tools. Formats such as mzML,[24] mzTab-M,[25] mzML2ISA[25,26] and mzSpecLib[27] exemplify the growing suite of open standards developed to address these needs. Among them, mzML is widely regarded as the cornerstone of MS data representation, offering a comprehensive XML-based structure to capture raw data, metadata, and instrument settings. Complementary formats, such as mzTab-M, are tailored for tabular outputs and facilitate the exchange of processed data, while mzSpecLib focuses on spectral libraries and their integration into workflows.

Looking ahead, the adoption of these formats must emphasize adaptability to emerging technologies such as artificial intelligence (AI) and machine learning (ML). AI-driven analyses demand consistent high-quality datasets, often requiring data to be structured and annotated in ways conducive to advanced algorithms. Ensuring that current formats are extensible and compatible with future standards will allow seamless integration of AI into NELs workflows. For example, standardized ontologies and metadata annotations could enhance interpretability while reducing preprocessing demands.

Finally, software libraries that facilitate the processing steps for the final users, such as pyOpenMS,[28] matchms,[29] spectrum_utils,[30] MSnbase,[31] and others serve as critical tools for interfacing with these formats. Most of these tools have formed or are forming user and developer communities to foster their uptake and maintenance. Each package has its own core functionalities, usually inspired by the reasons for conceptualizing the package. For example, the origin of matchms lies in making the comparison of mass fragmentation easier. Over time, additional functionality has been added, *i.e.* reading in various MS data types and curation of MS library metadata.

Hence, by simplifying data parsing, visualization, and manipulation, these libraries lower the barrier to entry for researchers seeking to develop custom workflows. Efforts to ensure that these tools remain open-source and broadly compatible will foster collaboration and innovation across the field, making MS data management a scalable and future-proof endeavor. Developers can further lower barriers by ensuring documentation is accessible to non-experts in chemistry, within reasonable limits.

There are many open-source tools available that are used for MS data processing in NP research, such as mzmine (Java),[32] XCMS[33] (R), Metaboanalyst (R),[34] MS-DIAL (C#),[35] and OpenMS (C++, Python).[36] We are aware that these and other metabolomics tools have been undergoing developments over the last years to support the analysis of larger datasets. The major steps in feature detection and project-centric feature alignment are similar in all these tools with varying algorithm options. The

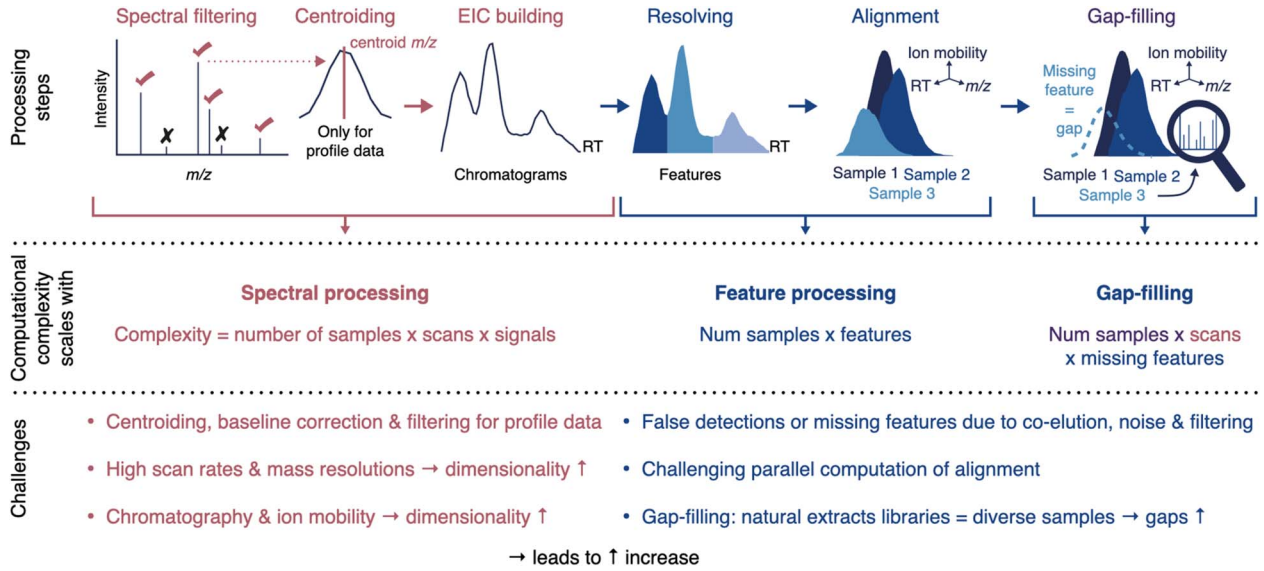


Fig. 3 Overview of the main feature detection and alignment steps and corresponding complexity and challenges. Feature detection starts with mass spectral processing like spectral filtering and centroiding of profile mode spectra. Performance for these steps is defined by the number of spectral signals, scan rate, and the number of samples. Chromatogram building connects similar *m*/*z* signals across the retention time dimension and feature resolving applies peak detection. Then, in a project-centric approach, features are aligned based on *m*/*z* and other separation dimensions and potential missed features (gaps) are filled in an optional step. The performance depends on the number of features detected across all samples and in the case of gap-filling on access to the spectral data. Higher feature quality filters may lower the number of noisy features, therefore increasing the alignment performance.

next section exemplifies a workflow for LC-MS data processing in mzmine and how this tool solved some of the scalability challenges.

2.2.2 Feature detection. Feature detection typically describes a complex multistep approach that aims to extract *m*/*z* signals of interest and their abundance across samples (Fig. 3). Most tools require centroid mass spectra by applying mass detection, *i.e.*, peak detection algorithms, on each scan. If spectra were already acquired or converted as centroids, this step applies only a noise threshold to reduce the memory requirements and speed up processing. The data processing workflow comprises multiple spectral processing and feature processing steps, as exemplified for LC-MS data processing in Fig. 3, and may vary for other hyphenated sample introduction systems. For example, mzmine provides an easy configuration through the mzwizard for LC-MS, GC-MS, and MALDI-MS with options to include ion mobility separation. First, raw data import supports open MS data formats, *e.g.*, mzML or imzML, and various vendor-specific formats. For chromatography, extracted ion chromatograms (EICs) are built by connecting *m*/*z* signals across the entire retention time (RT) range. Peak detection algorithms then resolve each EIC into separate features, *i.e.*, chromatographic peaks with specific *m*/*z*, RT, and other optional identifiers. For LC-IMS-MS measurements, features are then extended into the ion mobility dimension by building and resolving ion mobilograms.

All feature detection processing steps can run in parallel using one thread per sample and thread pooling. Performance implications for feature detection arise from the number of samples and the complexity of the MS data. A faster scanning mass spectrometer with higher resolution, like those often used for IMS-MS, produces more *m*/*z* signals resolved in a multidimensional space. This amount of data quickly outgrows the available random access memory (RAM). Modern MS data processing tools use memory mapping to offload data onto disk, creating a direct connection between computer memory and storage, allowing access and manipulation of large datasets. In the case of mzmine, spectra, chromatograms, mobilograms, and most of the feature tables are memory mapped, increasing the number of samples that can be processed in parallel. In the case of timsTOF IMS-MS data processing in mzmine, the memory mapping reduced the memory requirements by >90%. A recent update to the internal feature list data model in mzmine 3.7.8 improved the RAM requirements 10×, enabling the processing of larger studies. Another optimization for large datasets applies centroiding and filtering to spectra during import, allowing parts of the chromatogram's beginning and end to be removed.

2.2.3 Feature alignment. Feature alignment is another cornerstone of MS-based metabolomics workflows, ensuring that ion signals corresponding to the same compound are correctly matched across runs. Without reliable alignment, downstream quantitative and comparative analyses become error-prone, especially as study size grows. Broadly, alignment algorithms follow one of two paradigms—project-centric and sample-centric. Each approach offers distinct advantages and challenges depending on the scale and nature of the study, which will be discussed in the following subsections.

2.2.3.1 Project-centric approach. Feature alignment takes all sample-specific feature lists and aligns features based on their

m/*z*, RT, and tandem mass spectra, ion mobility or collisional cross section (CCS) if available. The aligned feature list usually contains gaps that originate from misalignment, maybe due to RT shifts, and feature detection issues during the chromatogram resolving. For example, if a double peak was split in one sample but not in the other, or if a chromatographic peak fails the user-defined peak shape constraints. Therefore, a gap-filling step is often added as a secondary feature detection that is informed by the initial results. Gap filling automates the typical manual process of going back to the raw data and extracting the intensity of the missing *m*/*z* ranges. The final feature list is then used for downstream statistical analysis, compound annotation, and exploration of the chemical space. Integration of results from multiple downstream tools is possible through the feature ID, sometimes called row ID. There are many optional filters and steps that are described in more detail in a recent protocol.[37]

Most steps until feature alignment can run in parallel as one task for each sample. However, algorithms that align results or that run on a single aligned feature list need special attention during their design to split the work for modern multicore hardware. The join aligner and gap-filling in mzmine are good examples of steps that used to be bottlenecks for large-scale MS analysis. Processing thousands of LC-MS runs would take multiple days to complete in old mzmine 2 versions,[38] but were completely redesigned for concurrency reducing the time required to minutes or seconds. In a benchmark, mzmine 3 processed 1920 diverse plant extracts in less than 40 min and 8270 human plasma extracts in less than 50 min on a data processing computer.[32] Generally, most feature detection tools scale hardware vertically on a single computer or server with more CPU, RAM, and fast (SSD) storage to map memory and offload data from RAM.

Still, one of the most important performance deciders is the size of the dataset and workflow parameter optimization. For large-scale analysis, higher noise levels, stricter feature shape constraints, and other filters can reduce the number of noisy features that will otherwise increase the processing time.

Project-centric alignment offers a unified reference frame that ensures consistent *m*/*z* and retention-time coordinates across all samples, minimizing cumulative drift and simplifying batch-effect detection and correction. By generating a single consensus feature map, it improves management of missing value and streamlines manual curation. Compared to sample-centric alignment, the project-centric approach also includes features without tandem mass spectra and those with poor fragmentation patterns in its statistical analysis and feature prioritization workflows.

2.2.3.2 Sample-centric approaches and knowledge graphs: a scalable paradigm for metabolomics. The sample-centric approach proposed by Gaudry *et al.*, through the introduction of MEMO, represents a novel approach in metabolomics data organization, particularly when viewed through the lens of scalability.[39] Unlike traditional project-centric frameworks, which structure data around specific studies or experiments, the sample-centric approach focuses on individual samples as fundamental units. This allows for more flexible integration, aggregation, and re-analysis of data across diverse projects and contexts. Knowledge graphs[40,41] are central to this methodology, linking samples to metadata, experimental results, and chemical or biological annotations in a highly interconnected network. Such graphs facilitate the exploration of relationships between datasets, offering scalability by enabling efficient querying, visualization, and hypothesis generation.

In contrast, project-centric approaches often silo data within specific experimental scopes, limiting reuse, and requiring additional preprocessing to integrate results across studies. This fragmentation becomes a bottleneck when scaling to large datasets, as it inhibits the aggregation of information, essential for comprehensive analyses. The sample-centric model, on the contrary, inherently supports scalability by allowing data from new samples or studies to be seamlessly incorporated into existing knowledge frameworks.

Furthermore, by leveraging knowledge graphs, researchers can apply advanced computational tools such as machine learning and artificial intelligence to identify patterns or prioritize samples and features at a scale that project-centric approaches struggle to achieve. By reconceptualizing metabolomics workflows around the sample-centric model and harnessing the power of knowledge graphs, the field is better equipped to address the challenges of NELs exploration at scale.

2.2.4 Feature grouping. The grouping of various ion adducts, in-source fragments, and multimers of the same compound is often done on a feature grouping level using the *m*/*z* and other separation dimensions as identifiers. The mzmine workflow provides metaCorrelate to group features that may originate from the same compound. Depending on the study size, the grouping can be based on just RT windows, a feature height Pearson correlation across all samples, or a more comprehensive feature shape Pearson correlation between feature pairs within the same sample. These three options increase in computational complexity, and feature shape comparison may reduce the throughput significantly for large studies with many coeluting features. If the ionization conditions are similar across all samples, the feature height correlation will provide a significant performance increase. In a second step, called ion identity networking in mzmine,[42] ion adducts, in source fragments and multimers are annotated by searching for specific *m*/*z* differences between grouped features that correspond to pairs from an ion library. Overall, the initial feature grouping reduces the number of metabolite features and, as such, it decreases the number of comparisons required in the final ion annotation. Still, the size of the ion library and the number of coeluting features increase the computational complexity and render the feature grouping step one of the time-limiting steps. Very large studies may decide to skip this step. Furthermore, typically, NELs metabolomics data files contain a relatively large portion of less abundant features. To streamline exploration in tools and interactive dashboards, the resulting peak quantification table can be subsequently filtered based on feature abundance by removing lower-intensity features. Another filter that could be applied is the minimum number of samples in which the feature should be present across the entire dataset or within a sample group. Note that

removing (relatively) unique features is likely to reduce the number of novel features as they tend to be more unique, so it is important to consider this well in NELs explorations targeting novel specialized metabolites.

2.3 MS data annotation

Before going into the details of the tools used in this important task, two fundamental notions (*i.e.* structural and spectral similarities) must be defined to understand key aspects of scalability in MS data annotation.

2.3.1 Structural similarity. Structural similarity is a key concept in cheminformatics and computational MS, enabling the comparison of molecular structures to identify related compounds. One of the most widely used approaches for this purpose is the Tanimoto similarity, which calculates the Jaccard index between molecular fingerprints (*i.e.* the size of the intersection of two molecular fingerprints divided by the size of their union).[43] Molecular fingerprints are representations of molecules as bit vectors, where each bit corresponds to the presence or absence of specific structural fragments within the molecule. These fingerprints are generated by applying a kernel to the molecule, which encodes structural features such as atom connectivity and substructures. Extended-connectivity fingerprints (ECFPs), which capture circular substructures around atoms, are among the most popular fingerprint types due to their ability to capture a wide range of molecular features.[44] Recently, new fingerprints have emerged, also designed to specifically handle biomolecules and natural products, such as MAP4.[45] This increase in specificity, however, due to more calculations, also comes with increased computation time. The Tanimoto similarity is computationally efficient, especially when fingerprints are precomputed, as it relies on simple bitwise operations. Consequently, Tanimoto similarity is a practical choice for large-scale applications requiring many molecular comparisons. Despite its efficiency, Tanimoto similarity has limitations. Because encoding molecules as binary vectors is based on local substructures rather than the full molecular structures, this simplified representation can lead to counterintuitive similarity values, such as low similarity scores for molecules that are visually and chemically similar or unexpectedly high scores for structurally distinct molecules with coincidental overlaps in their fingerprints.[46]

To address these shortcomings, alternative methods based on graph representations of molecules have gained traction. In these approaches, molecules are modeled as graphs, with atoms as nodes and bonds as edges. Structural similarity is then assessed using metrics such as the maximum common subgraph (MCS) and maximum common edge subgraph (MCES). Intuitively, MCS determines the largest subgraph that two molecular graphs share, focusing on the common structural framework, while MCES extends this by also considering the largest subset of edges (bonds) shared between the graphs. These metrics are more interpretable than Tanimoto similarity, as they directly reflect shared structural features and provide insights into how molecules are related in terms of their core scaffolds and bond arrangements.

However, the computational complexity of MCS and MCES calculations presents a significant challenge. Both problems are nondeterministic polynomial-time complete, meaning that there is no known algorithm that can solve them efficiently for all cases in polynomial time. As a result, comparing large numbers of molecular pairs using these metrics can be prohibitively time-consuming, especially for large datasets. To overcome this bottleneck, heuristic-based approaches have been developed to approximate MCS and MCES calculations. One recent advancement is the introduction of the myopic-MCES Python package, which uses heuristic optimization methods to estimate the MCES distance for molecular pairs above a defined similarity threshold.[47] By focusing on high-similarity pairs and approximating results for others, myopic-MCES achieves significant speed improvements. While this approach sacrifices some accuracy for less similar pairs, it offers a practical trade-off, enabling the efficient analysis of large molecular datasets. Although still too new to have been used in NELs exploration, GESim[48] offers the best of both worlds—combining the speed of fingerprints with the structural richness of graph-based similarity, approaching the computational efficiency of fingerprints.

2.3.2 Spectral similarity. The comparison of mass spectra lies at the heart of most MS data analysis tools (Fig. 4). For example, library matching, analogue search, and mass spectral networking are amongst the computational metabolomics strategies that require mass spectral comparisons. With a few spectra at hand, such comparisons can be performed manually, assessing whether the fragmented molecules are structurally similar or whether they share scaffolds or building blocks. However, to handle large amounts of spectra efficiently, MS similarity scores are used as numeric representations of the spectral similarity. Such scores allow researchers to rank library matching results and to apply thresholds when creating mass

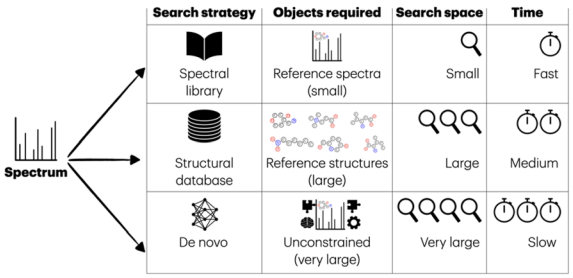


Fig. 4 Overview of different search strategies, the objects required, the search space they need to consider, and the computational efficiency of similarity calculations. Spectral library searching only needs to consider a small amount of reference spectra, thus a small search space. Structural database retrieval requires a large training set, and employs a larger search space. Finally, *de novo* search requires a very large amount of training data and essentially considers an unconstrained search space. Spectrum similarity (using cosine-based approaches) can be calculated relatively efficiently, and structural similarity (using the Tanimoto similarity) can be computed very efficiently, while *de novo* molecule generation requires advanced algorithmic and deep learning approaches that are more computationally intensive.

spectral networks. Traditionally, a few spectral similarity scores are used with the cosine score as the main go-to score; however, recently a range of new flavors have been developed that have shown to outperform the cosine score on specific tasks using various benchmarking datasets. In this section, we will highlight various strategies that build on the MS similarity score.

2.3.2.1 Exact search. Exact search attempts to find identical molecules by matching experimental spectra with reference measurements in a spectral library. One of the most widely used methods for spectral annotation is cosine similarity, also referred to as the (normalized) dot product. The cosine similarity between two spectra measures the degree of alignment between their fragment intensities, rewarding overlapping peaks and penalizing those that are mismatched. The mathematical formula for cosine similarity is as follows:

$$\text{Cosine similarity} = \frac{\sum_{(i,j)\in M} I_{1,i} \times I_{2,j}}{\sqrt{\sum_i {I_{1,i}}^2} \times \sqrt{\sum_j {I_{2,j}}^2}}$$

Here, $I_{1,i}$ and $I_{2,j}$ are the intensities of the i-th peak in the first spectrum and the j-th peak in the second spectrum, respectively. M is the set of all pairs of peaks (i, j) where the m/z difference between the peaks is within a specified tolerance of the fragment, *i.e.*, $|m/z_1{,}i - m/z_2{,}j| \leq$ tolerance. For increased computational efficiency, the intensities of the fragment can be normalized to unit length, thus turning the denominator to a constant.

Cosine similarity is intuitive: it captures the proportion of matching intensity between two spectra while penalizing unmatched intensities. Although it has been a foundational method used for decades,[49] it remains the default approach in many bioinformatics pipelines due to its simplicity and robustness. However, implementing cosine similarity efficiently requires careful consideration. One common approach involves binning the spectra according to the fragment mass tolerance and using standard vector operations. Although straightforward, this method is prone to edge effects, where peaks near bin boundaries are incorrectly assigned to neighboring bins, leading to missed matches. A more precise approach is a peak-by-peak comparison, iterating through all peaks in both spectra and matching those within the fragment mass tolerance.[29] This avoids binning artifacts and ensures accurate comparisons.

In terms of computational scalability, a key consideration is handling duplicate peak matches. For example, a single peak in one spectrum might match multiple peaks in other spectra due to similar m/z values. Including all possible matches would artificially increase the similarity score. To address this, only the most relevant peak matches should be considered. A common heuristic involves a greedy search that iteratively selects the pair of peaks with the highest contribution to the explained intensity, ensuring that the peaks are not reused.[29,50] This approach is computationally efficient, with a time complexity of $O(n \log n)$, where n is the number of peaks in a spectrum. Alternatively, the Hungarian algorithm for combinatorial optimization can provide an optimal solution to the peak matching problem, but its $O(n^3)$ time complexity makes it impractical for large-scale spectral comparisons. The heuristic approach is typically sufficient with minimal deviations from the optimal solution in most cases. Furthermore, this approach can be efficiently implemented using graphics processing units (GPUs),[51] providing orders of magnitude speedup compared to central processing unit (CPU)-based implementations. This computational efficiency, combined with its intuitive approach and competitive results,[49] makes cosine similarity an essential baseline method for large-scale spectral comparisons.

2.3.2.2 Modified cosine similarity. An extension of cosine similarity, is commonly used as a spectral similarity metric during analogue searching. Modified cosine similarity is capable of capturing not only directly matching peaks but also peaks shifted by the pairwise precursor mass difference and thus accounts for fragments differing by a modification inferred from the precursor mass difference. During exact searches, where only spectra within a small precursor mass tolerance are compared, the modified cosine similarity reduces to the standard cosine similarity.[52] However, in analogue searches and molecular networking, where larger precursor mass differences are allowed, the modified cosine similarity is able to capture both identical fragments and those resulting from neutral losses or structural variations.[53] Despite these advantages, the modified cosine similarity also has limitations. It assumes that the precursor mass difference corresponds to a single modification, which restricts its accuracy for molecules differing by multiple modifications. This challenge arises because the algorithm cannot partition the precursor mass difference between multiple potential shifts, limiting its ability to accurately score spectra of more complexly related molecules.

2.3.2.3 Analogue search. Reference spectral libraries can be further leveraged through an analogue search. This seeks to determine whether the reference library contains similar or related structures as to those that were measured in the experiment. There is no exact definition of relatedness, but analogues are typically defined as "sharing a main scaffold" or "belonging to the same compound family". Indeed, analogue search can provide structural annotation guidance, as the found analogue may help in structural annotation of overlapping parts of the queried mass spectrum. Moreover, the mass difference between the query and the database hit may also contain relevant clues, especially when they represent values corresponding to hydroxylation, methylation, glycosylation, or other known biotransformations.

However, analogue searching poses a substantial challenge to computational workflows. Any naive approach for large-scale analogue searching is unfeasible, regardless of the similarity measure used, due to the quadratic number of spectrum comparisons that need to be performed. Although exact matching benefits from selecting possible library entries based on recorded precursor masses, thereby avoiding a considerable amount of mass spectral comparisons, analogue search, in principle, needs to consider the entire library (although, in practice, most tools usually still restrict to subparts of the library). The increasing size of (public) mass spectral libraries is beneficial, but it also imposes an increased challenge to mine them, especially considering that experimental datasets are also

becoming more information-rich, resulting in many-*versus*-many comparisons. To mitigate this challenge, at least partly, different tools use different approaches. In the following, we will highlight several analogue searching tools and discuss how they handle this challenge.

2.3.2.4 GNPS analogue search. Within the GNPS ecosystem, the mass spectral library annotation allows for both exact matching and analogue search. Users can indicate thresholds for the similarity score used to filter possible library hits. By default, analogue search is switched off. If toggled on, the default maximum mass difference is 1999 Da that, if unaltered, does substantially increase the analysis time. Hence, to make this analogue search more scalable, the maximum mass shift can be set by the user. In the NPs context, a maximum of 200 Da covers a nice amount of biotransformations that could occur, and it will reduce the analysis time considerably. However, this comes with the cost that some relevant analogues may no longer be found. The use of the modified cosine score is helpful to find analogues that differ in one building block, consequently shifting mass fragments with the mass of the building block.

2.3.2.5 Fragment ion indexing. Recent advancements in computational efficiency have leveraged fragment ion indexing to handle the increasingly large datasets generated by modern MS experiments. This approach can significantly speed up the calculation of spectrum similarities, including cosine similarity and spectral entropy-based scores, improving scalability. Fragment ion indexing, originally popularized by the MSFragger tool for open modification searches in proteomics,[54] operates by creating an efficient representation of peak presence across spectra. The process begins by binning spectra based on fragment m/z values, with bin widths corresponding to the fragment mass tolerance. For each spectrum, the presence or absence of peaks within these bins is then encoded in efficient data structures that facilitate rapid access and comparison.

Querying with a fragment ion index involves comparing a query spectrum against the indexed database to identify spectra with a defined minimum number of matching peaks. This comparison is computationally efficient because it directly retrieves the matching fragments from the index without having to consider irrelevant spectrum pairs. The result is a drastic reduction in the time required to filter and rank spectra, especially for large-scale datasets.

The fragment ion indexing strategy has been applied to tasks such as analogue searching, matching against spectral repositories, and for molecular networking.[55] By replacing more computationally intensive pairwise similarity calculations with efficient indexing-based queries, these applications achieve better scalability without sacrificing accuracy.

A notable extension of fragment ion indexing is its application to spectral entropy-based similarity calculations. Spectral entropy provides an alternative to cosine similarity by measuring the disorder of the fragment ion intensity distribution.[56] In the context of analogue searches, where query spectra are compared to spectra of related but non-identical molecules, fragment ion indexing accelerates entropy-based similarity calculations. This approach, termed flash entropy, uses a similar indexed framework to efficiently process large datasets, enabling rapid identification of spectra with high entropy-based similarity.[57]

It is important to note, however, that because fragment ion indexing relies on peak binning, edge effects can occur, as discussed previously. While the method remains computationally efficient and broadly applicable, users should be aware of these limitations and consider alternative approaches, such as peak-by-peak comparisons, for tasks requiring the highest precision.

2.3.2.6 Suspect library. The nearest neighbor suspect spectral library is a recent approach to leveraging repository-scale MS data for the discovery of structural analogs.[58] This method capitalizes on the growing availability of open data repositories and scalable computational tools to interpret unannotated spectra that are linked to annotated spectra corresponding to known molecular structures. By deriving insights from hundreds of millions of MS/MS spectra across thousands of datasets, this strategy perfectly exemplifies the power of large-scale MS data analysis. The creation of the nearest neighbor suspect spectral library involved analyzing 521 million MS/MS spectra derived from 1335 public projects hosted on repositories such as GNPS/MassIVE,[59] MetaboLights,[60] and Metabolomics Workbench.[61] Using molecular networking with the modified cosine similarity, "suspects" were extracted, which are unannotated spectra that were connected in the molecular network to annotated spectra *via* spectral matches. Next, by propagating molecular annotations and analyzing the observed precursor mass differences compared to a curated list of potential modifications, structural hypotheses for these suspects were generated. This approach underscores the major potential of repository-scale analyses. By co-analyzing datasets from thousands of studies, it becomes possible to uncover relationships and patterns that are undetectable within the confines of individual datasets. The outcome of this effort was the creation of the nearest neighbor suspect spectral library comprising 87 916 unique MS/MS spectra. This spectral library is freely accessible on the GNPS platform, where researchers can use it to identify novel structural analogs that are absent from traditional reference spectral libraries. Across various applications, integrating the nearest neighbor suspect spectral library has been shown to on average double the number of annotations during spectral library searches, significantly enhancing the scope of molecular discovery.

2.3.2.7 Machine and deep learning-based similarities. The implementation of the cosine score has inspired the creation of various flavors that use different weighings to influence the contribution of smaller or large m/z values or intensities.[56] Furthermore, several machine/deep learning-based similarity scores have been proposed and evaluated in their exact matching performance.[62] Spec2Vec is the firstly introduced unsupervised machine learning-based score,[63] learning fragmental and neutral loss relationship of mass from large amounts of mass spectra. A key asset is its tolerance to multiple minor modifications: the authors demonstrate how, unlike signal alignment-based scores, Spec2Vec still results in higher similarity scores even when two molecules differ in more than two structural modifications, and as a result, their mass spectra have little signal overlap. In addition, it showed encouraging

performance in library searching. Another example of a supervised deep learning-based similarity score is MS2DeepScore.[64,65] Here, the scoring is trained using MS libraries as input, with the Tanimoto score between structure pairs as a goal to approach with the spectral score. Again, the score turned out to be more modification-tolerant than the signal alignment-based scores, and, even though not specifically trained, had encouraging performance in library matching. Overall, the scalability of such machine-/deep-learning-based similarity scores is promising, as one needs to compute an embedding for each mass spectrum only once, after which the cosine similarity can be applied on the matrix of embeddings.

We note that current approaches mostly rely on binning to determine if mass signals should be considered the same across spectra, similar to how signal alignment-based scores do. The bin size is a key factor in scalability: smaller bins will result in more accurate discrimination between isobaric mass fragments, but in larger computational times. Furthermore, using smaller bin sizes (*i.e.*, 0.005 Da) relies on high-resolution MS data. To accommodate a larger section of the public data, wider bin sizes (*i.e.*, 0.01 or 0.1 Da) are typically used, thus sacrificing some accuracy over scalability and coverage. We note how also a combination of such scores can be used for tasks such as library matching or analogue search (see previous Section 2.3.2). Supervised scores such as MS2DeepScore are reliant on the availability of curated comparable library spectra, and recent initiatives such as MassSpecGym are valuable resources as they specifically aim to have machine learning-ready data.[66] Furthermore, having a large public dataset available for learning or training and testing/validation, will also help to compare new scores more effectively. Overall, on the basis of the above developments and considerations, we expect more mass-spectral similarity scores to be developed in the near future, and we hope that the community will further adopt and test them for their specific use cases. For example, a recent study compared the discriminative power of various similarity scores for monoterpene indole alkaloids.[67] We anticipate that with more such studies, it will become increasingly clear what kind of score to use for which task and chemical compound class.

2.3.2.8 MS2Query. An alternative route to find structural analogues based on mass spectral similarity was proposed in 2023 with MS2Query.[68] Building on machine learning-based mass spectral similarity scores, the unsupervised Spec2Vec[63] and the supervised MS2DeepScore[64,65] (see de Jonge & Mildau *et al.* for further information[62]), MS2Query uses an overarching machine learning model that takes the input from the above scores and other spectral information to rank potential analogues and exact matches in a provided mass spectral library. As the tool is not dependent on mass fragmental overlap (as all the cosine score flavours are), MS2Query can account for multiple modifications that typically result in low or sometimes even no fragmental overlap between the two analogue mass spectra. After training the machine learning models, the retrieval of ranking scores is relatively fast, thereby removing the need to cap hits based on a maximum mass shift – instead, the entire library can be considered. When doing analogue search with approximately 6000 query spectra in an approximately 300 000-sized mass spectral library, MS2Query was almost an order of magnitude faster than the modified cosine-based analogue search as implemented in matchms at the time of publication in 2023 (*i.e.*, approximately 80 spectra per min *versus* approximately 10 spectra per min), without maximum mass shift for MS2Query, and using 100 Da for the modified cosine-based search (thus heavily restricting the number of candidate analogues). Altogether, MS2Query offers an alternative route to analogue searching, and when capped using precursor mass information of the query spectrum, can also be used to search for exact matches. As it is built on the matchms package (see also Section 2.2.1), it fits neatly in Python-based metabolomics annotation workflows used for large-scale annotations, for example as done by Simone *et al.*[69]

2.3.3 Annotation using spectral libraries. In this section, we will discuss a non-comprehensive selection of spectral and structural MS-based annotation tools in the context of large-scale NELs metabolomics studies.[70–72]

Spectral library matching remains a cornerstone for MS-based annotation in large-scale NELs metabolomics studies. Several open-access spectral libraries are widely used for dereplication and identification of NPs. MassBank[73] is one of the earliest and most comprehensive public repositories, providing high-quality reference spectra for a broad range of metabolites. GNPS has become a central platform, not only aggregating MS/MS spectra relevant to NPs from diverse sources but also enabling community-driven curation and containing more specialized sub libraries such as PhytoChemical Library, the NIH Natural Products Libraries, the Lichen Database,[74] the MIADB,[75] or Annonaceous Metabolites Database.[76] Here we would like to emphasize how much the size covered by the chosen library will impact later scalability. Other openly available spectral libraries specialized libraries include the multiple ones shared by Brungs *et al.*[77] or MassBank of North America.

Despite the increasing size and diversity of these libraries, coverage of the vast chemical space of NPs remains incomplete, and matching rates in large-scale untargeted studies are still low. This limitation has driven the development of hybrid approaches that bridge spectral and structural annotation. The first one was MetFrag,[78] followed by CFM-ID,[79] and more recently FIORA.[80] CFM-ID was used to fragment 1 million compounds of interest originating from Wikidata (or its LOTUS subset) and shared as an ISDB[81] (*In Silico* DataBase). These integrative strategies are increasingly important for NELs studies, where the diversity of metabolites often exceeds the coverage of any single spectral library.

2.3.4 Annotation using structural libraries

2.3.4.1 SIRIUS. SIRIUS has emerged as the leading suite for molecular structure annotation from MS/MS data. At its core, SIRIUS determines the molecular formula of precursor ions and their fragments using high-resolution MS, constructing fragmentation trees that model the breakdown of molecules during MS/MS analysis. To simplify use, the platform's modular design now includes several specialized subtools: CSI:FingerID[82] predicts molecular fingerprints and enables structure identification by searching large molecular databases; CANOPUS[83] assigns compound classes also to unknown metabolites;

ZODIAC[84] improves molecular formula ranking by leveraging relationships across spectra; COSMIC[85] provides confidence scoring for structural annotations; and MSNovelist[86] supports *de novo* structure generation. Parts of these subtools will be presented here and others in later subsections.

Sirius' compute times are important, but the most intensive steps are outsourced to external servers. While those times might look prohibitive, the problems it answers are much more complex. Keeping in mind the computation time increases exponentially with m/z, SIRIUS's scalability has been demonstrated on small molecules through its application in large-scale metabolomics studies. For example, in a recent study,[85] its ability to process 20 080 LC-MS/MS datasets, including a human dataset of 2666 runs completed in 4 days and an Orbitrap dataset of 17 414 runs requiring 21 days, both on a 96-core compute node. In a study oriented towards multi-omics of Earth's microbiomes,[38] Sirius annotated fragmentation spectra in 880 environmental samples, though specific computation times for SIRIUS were not detailed, it was specified that the Sirius v. 4.4.25, headless, Linux was used for this task. The term "headless" indicates that this version is designed to run without a graphical user interface (GUI), often used for servers or virtual machines where graphical interfaces are unnecessary, allowing for more efficient use of system resources. These instances highlight SIRIUS's capacity for repository-scale analysis, managing databases like PubChem (77 190 484 unique structures) and a biomolecule structure database (391 855 unique structures), indicating robust scalability with high-performance computing resources.

2.3.5 Annotation of substructures. Substructure annotation is another way to get structural insights, when full structural annotation is not possible. Several computational tools have been developed to address this challenge: CFM-ID, mentioned earlier for *in silico* generation of spectra from structures, has substructure annotation at its core. While it slows its execution down, the fragments generated by CFM-ID are matched to substructures of the original structure, using RDKit to generate possible combinations. SIRIUS now also offers substructure annotation capabilities. With the release of SIRIUS 5, the Epimetheus module enables direct visualization and assignment of substructures to MS/MS peaks, allowing users to inspect and validate substructure annotations alongside candidate structures. This development further strengthens SIRIUS's position as a comprehensive suite for both structure and substructure annotation in metabolomics workflows.

2.3.5.1 MS2LDA. Introduced in 2016, MS2LDA was the first tool to identify substructure-related mass spectral patterns in an unsupervised manner.[87] By applying a topic modeling algorithm to MS/MS spectra, MS2LDA groups frequently co-occurring mass fragments and neutral losses into patterns termed Mass2Motifs. Ideally, these Mass2Motifs serve as "substructure footprints" that can be linked to underlying (bio) chemical structures. The number of these unsupervised "free" Mass2Motifs to be discovered must be specified by the user—a task that becomes increasingly challenging with complex mixtures and large datasets. As a result, MS2LDA outputs typically contain both meaningful and noise-driven patterns. To address this, a dedicated web application was developed, offering visualization tools such as feature histograms and color-coded spectra to support expert review. However, manual interpretation of Mass2Motifs remains a bottleneck in large-scale analyses, prompting the development of automated tools for prioritizing plausible motifs for downstream annotation.

2.3.5.2 MotifDB. Recognizing the value of capturing and sharing validated substructure patterns, the MotifDB repository was established shortly after MS2LDA's launch.[88] It now hosts several curated MotifSets derived from plant, microbial, urine, and reference standard datasets. Ten of these MotifSets contain at least 10 manually annotated and validated Mass2Motifs—ranging from 10 motifs derived from monoindole alkaloid standards to 134 motifs from experimental urine data. In total, nearly 500 annotated Mass2Motifs are available. These can be integrated directly into MS2LDA analyses to screen experimental data for known motifs, streamlining substructure identification and, in some cases, eliminating the need for further structural annotations.

2.3.5.3 MESSAR. An alternative supervised substructure discovery approach introduced in 2020.[89] This tool was trained to connect mass fragments and mass differences to specific substructures to perform substructure recommendations motivated by association rule mining. The substructure annotation takes place in two steps: first, the mass fragments and differences are annotated with structures, and secondly, these annotations are clustered to find the maximum common substructures that best fit with the provided information.

2.3.5.4 Large-scale substructure mining. Both the MS2LDA-MotifDB and MESSAR route may work well for relatively small datasets, with the number of mass fragment and neutral loss features combined for all mass spectra as the key factor,[90] but their applicability for larger-scale datasets is limited. One aspect that contributes to this is the memory requirements of these tools: when run with a large amount (over 100 000) combined mass fragments and neutral losses, a typical server easily runs out of memory. We should add here that running MS2LDA with mainly annotated Mass2Motfis is substantially faster and less computationally demanding than running MS2LDA with predominantly "free" Mass2Motifs, as part of the topic modelling distributions are fixed for annotated preset Mass2Motifs and no longer need to be learnt from the data. Another aspect is the required human expert knowledge and time to analyze and compare MESSAR results (of step 2) to come to plausible and reliable substructure information, and to validate the MS2LDA-MotifDB matches. For instance, is the annotated Mass2Motif genuinely associated with the experimental mass spectrum, or could it be a misleading result caused by the overlap of one or two prominent features that make up the Mass2Motif substructure pattern? Current initiatives are focused on addressing these challenges in MS2LDA, with the goal of developing a community-driven substructure annotation platform that is modular, scalable, and that makes it easier to interpret the mass spectral patterns. Very recently, MS2LDA 2.0 has been launched as a stepping stone toward such an

ecosystem, including Mass2Motif Annotation Guidance to facilitate Mass2Motif structural annotation and a completely redesigned MS2LDAViz application to visualize and interpret Mass2Motifs interactively.[91]

2.3.6 Annotation of unknown structures (*de novo*). Currently, complex extracts contain more unknown than known structures, and these include completely novel chemistry.[20] Thus, computational metabolomics tools that are able to aid in the elucidation of novel chemistry are indispensable when analyzing NELs.

2.3.6.1 MSNovelist. Exploring the unknown chemical space remains one of the greatest challenges in metabolomics, where many spectra do not match known structures or analogues. MSNovelist[86] offers a cutting-edge solution by predicting putative molecular scaffolds or structural features for unknown compounds. By employing generative models trained on large chemical libraries, MSNovelist suggests possible structural hypotheses based on MS/MS data, even for compounds entirely outside the boundaries of known chemical space. This capability is crucial for scaling metabolomics to NELs, where the chemical diversity vastly exceeds the capacity of traditional structural annotation tools. Diving into the unknown not only expands our understanding of chemical diversity but also highlights potential leads for further experimental investigation. By combining generative tools like MSNovelist with repository-wide non-structural annotation strategies, metabolomics can achieve a scalable and integrative framework for addressing both the known and unknown realms of NELs.

2.3.7 Non-structural annotation

2.3.7.1 Repository scale. Non-structural annotation provides critical insights into metabolomics datasets without requiring full structural elucidation, offering a scalable approach to handle the vast chemical diversity present in NELs. Kruve *et al.*[92,93] demonstrated how activities can be predicted based on empirical data without resolving chemical structures, using tools that infer properties such as toxicity or bioactivity through machine learning models trained on physicochemical and spectral features. This approach is particularly valuable for prioritizing compounds with potential biological relevance when structural elucidation is infeasible due to data limitations or complexity. Similarly, Capecchi *et al.*[94] showcased how biosource annotation—linking metabolites to their biological origins—can be predicted. Other trials to predict the missing data points are ongoing.[95] By using statistical models to correlate spectral features with biosource metadata, they enabled the prediction of probable biological origins of unknown compounds. This is exemplified by workflows such as MASST repositories,[96] which integrate spectral similarity to link known spectra to biological sources or previously studied NPs. On a broader scale, repository-wide analyses of spectral databases provide a powerful means of non-structural annotation, leveraging the vast known spectral space to contextualize unknowns based on their spectral neighborhoods.

2.3.7.2 Biological source. Incorporating taxonomic and biological source information into metabolomics workflows is a crucial step toward improving annotation accuracy and scalability in NELs studies. Rutz *et al.*[97] demonstrated that Taxonomically Informed Metabolite Annotation (TIMA) significantly enhances the confidence of NPs annotations by integrating taxonomic metadata directly into computational workflows. This approach enables researchers to prioritize annotations that align with known biosynthetic capacities of organisms, reducing false positives and improving the overall reliability of metabolite annotation. Building upon this foundation, the LOTUS initiative[98] presents a scalable, community-driven framework for associating molecules with their biological origins. By structuring NPs data with rich taxonomic metadata, LOTUS fosters interoperability and knowledge-sharing across research disciplines, breaking traditional silos in metabolomics. This initiative not only improves annotation quality but also expands the reach of repository-scale analyses, enabling global collaborations in NPs discovery.

The impact of taxonomically informed workflows is further exemplified by foodMASST,[99] microbeMASST,[96] and plantMASST,[100] which leverage repository-wide spectral similarity searches to link unknown metabolites with known biological sources. By integrating taxonomy-driven approaches with spectral matching, researchers can uncover novel chemical relationships, refine annotation confidence, and scale metabolomics analyses to unprecedented levels. These methods collectively demonstrate that incorporating biological context into computational tools is not just an enhancement—it is essential for the future of scalable, high-confidence metabolomics.

2.3.7.3 Color-coded MN and bioactivity correlations approaches. Integrating biological activity data with large-scale untargeted MS-based metabolomics data involved several innovative approaches. Initially, the NP community employed color-coded molecular networks to represent the biological activities of active fractions. This visual method allowed researchers to manually prioritize MS features based on their bioactivity. A larger-scale example of this approach is the work by Olivon *et al.*,[4] who analyzed 292 extracts from various Euphorbiaceae species. The authors successfully integrated metabolomic, taxonomic, and bioactivity data into a single data matrix, facilitating the prioritization of bioactive compounds. Another significant advancement in this field is bioactivity-based molecular networking, proposed by Nothias *et al.*[101] This method predicts the bioactivity of each MS feature by calculating the Pearson correlation between activity profiles and intensity profiles for each feature in the sample set. The approach utilizes a combination of open-source tools and custom R scripts to achieve this integration. Despite its potential, no studies have yet applied this tool for NELs bioactivity integration. However, these methods are labor-intensive and require tailored workflows to be applicable to NELs. At last, an updated version of this approach has been adapted by McCall *et al.*[102] by applying Spearman correlation instead of a Pearson correlation which assumes linear relationships and normally distributed data.

While our focus here has been on correlation-based approaches due to their widespread use and accessibility in the field, alternative methods also exist. These include supervised machine learning algorithms, multivariate statistical

models (*e.g.*, PLS-DA), and network-based or topology-informed strategies, which can capture more complex, non-linear relationships between chemical features and biological activity. Although these approaches are promising and increasingly explored, correlation methods remain the most commonly applied in current NELs workflows due to their interpretability and ease of implementation.

2.3.7.4 From compound activity mapping to NP analyst. The quest for the development of new strategies for the prioritization of lead compounds with unique structural and/or biological properties from large scale untargeted metabolomics data led Linington *et al.* to design Compound Activity Mapping platform.[103] In this study, the authors combined high-content screening and untargeted MS-based metabolomics of 234 natural extracts, combining 10 977 MS features with 58 032 biological measurements to identify 13 clusters of fractions containing 11 known compound families and four new compounds. Seven years later, the same authors developed NP Analyst,[104] a stand-alone platform for data integration that includes both data analysis and data visualization components. NP Analyst accepts bioassay data of almost any type and is compatible with MS data from major instrument manufacturers, making it a versatile tool for generating global network views of biologically active chemical space for large extracts libraries. NP analyst analyzed a set of 925 pre fractions from an in-house marine actinobacterial strain library for bioactive compound discovery, using biological data from BioMAP antibacterial profiling against 15 bacterial pathogens.

2.4 Querying, prioritization, and decision-making

2.4.1 Querying metabolomics data. Efficient querying is a pivotal aspect of scaling MS-based metabolomics, enabling researchers to extract meaningful insights from increasingly complex datasets (Fig. 5). As metabolomics workflows grow to incorporate large-scale NELs studies, the development of versatile and scalable query languages tailored for MS data is essential. This section highlights two key approaches to querying MS data: MassQL, designed specifically for MS data, and SPARQL, a semantic query language for knowledge graphs.

MassQL[105] is a specialized query language tailored for metabolomics data, offering an intuitive syntax to perform advanced queries on MS datasets. Built to accommodate the complexity of MS, MassQL allows researchers to query raw and processed data by specifying parameters such as RTs, mass-to-charge ratios, and intensity thresholds. This granularity makes it an excellent tool for identifying features of interest within large datasets or pinpointing specific metabolites across samples.

What sets MassQL apart is its accessibility to non-programmers, as its syntax closely resembles natural language. This democratizes data exploration, enabling a broader range of researchers to engage with complex datasets without requiring extensive computational expertise. As metabolomics scales, MassQL's ability to handle vast datasets efficiently will become increasingly important, particularly for high-throughput studies where rapid and targeted data extraction is critical.

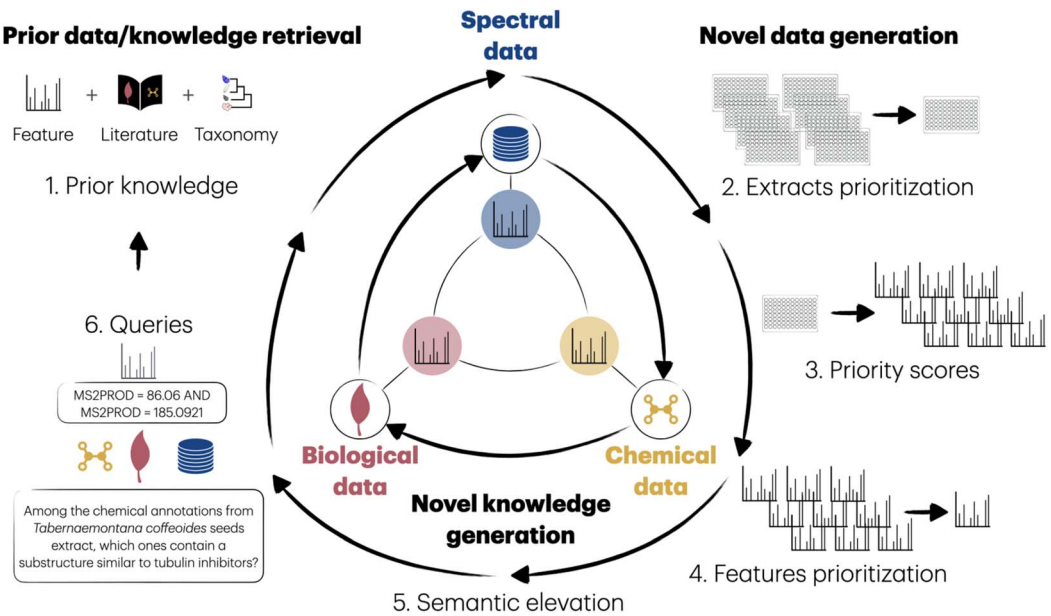


Fig. 5 Scalable workflow for the exploration of NELs. This diagram illustrates an iterative framework that integrates prior knowledge retrieval with novel data generation to enable scalable exploration of natural extracts libraries. Mass spectrometry data serve as a central component, linking biological and chemical domains as well as spectral data themselves. (1) Existing data—including spectral features, literature, and taxonomy—are incorporated as prior knowledge. (2) Extracts are prioritized, and novel spectral data are generated. (3) These data are used to further refine extract prioritization based on spectral signatures. (4) Individual features within prioritized extracts are then ranked. (5) Semantic elevation combines biological, chemical, and spectral information to enable more meaningful annotation and interpretation. (6) The resulting knowledge enables complex, hypothesis-driven queries—such as identifying compounds with substructures similar to known bioactive molecules—and informs subsequent iterations, creating a continuous, virtuous cycle of discovery.

SPARQL (SPARQL Protocol and RDF Query Language) offers a complementary approach to querying, particularly in the context of knowledge graphs and linked data. By leveraging the semantic structure of knowledge graphs, SPARQL enables complex and relational queries that go beyond the scope of raw MS data. For small molecules, the Integrated Database of Small Molecules (IDSM[106]) offers an incredible gateway to different small-molecules datasets also covering biological assays. It also comes with fast structural[107] and spectral[108] similarity extensions. For instance, researchers can query not only spectral data but also metadata linking metabolites to biological sources, experimental conditions, or known activities (see examples in **https://idsm.elixir-czech.cz/sparql**).

SPARQL's flexibility is particularly advantageous in large-scale NELs studies, where integrating diverse datasets is crucial for generating actionable insights. For example, a query could retrieve all metabolites with similar spectral features to a compound of interest, produced by a specific biosource, and associated with a given activity. As knowledge graphs become more widely adopted in metabolomics, SPARQL will be an indispensable tool for navigating the interconnected data landscape, facilitating both hypothesis-driven and exploratory analyses.

2.4.1.1 Toward scalable query frameworks. The scalability of querying in metabolomics depends not only on the robustness of tools like MassQL and SPARQL but also on their integration into broader workflows. Combining the specificity of MassQL with the relational power of SPARQL could provide a hybrid approach, leveraging the strengths of each language to address both raw data exploration and metadata-rich queries. Together, these tools will be critical for advancing metabolomics into the era of large-scale NELs studies, enabling researchers to manage and analyze the ever-expanding volumes of data with precision and efficiency.

2.4.2 Prioritization and decision-making. Inventa[109] is a computational workflow that enables an optimal decision making for extracts prioritization by pinpointing the structural novelty through the calculation of a priority score. The latter sums four components, including a feature component (that measures the feature specificity and annotation), a literature component (that reflects the number of compounds reported in the literature), the class component (that indicates whether a chemical class is detected), and the similarity component (that exploits multiple outlier-based machine learning algorithms to highlight dissimilarity) together with their modulating factors. As a proof of concept, Inventa was applied to a collection of 76 taxonomically related extracts of the Celastraceae family and resulted in the prioritization of an extract and the subsequent description of thirteen new beta-agarofuran sesquiterpenes.

2.4.2.1 Rational library minimization. In a recent study McCall *et al.* developed an innovative approach to rationally minimize NELs by directly addressing cross-organismal redundancy in NP production.[102] This method leverages a custom R code that processes the node table generated through GNPS classical molecular networking workflow. To select extracts to be added to the rational library, the algorithm aggregates the features by scaffold (molecular family), then chooses the extract that contains the most scaffolds. Then, those scaffolds are deleted from the data set. After that, the extract with the most scaffolds not already accounted for is added to the rational library. This process iterates until a desired percent of maximum diversity is reached or maximum diversity. The authors applied their approach on a NELs of 1439 fungal extracts and reduced the size of the library to 216 extracts (6.6 fold reduction) without decreasing bioactive hit rates. Remarkably, this approach enables dramatic cost reduction across all subsequent high throughput screening projects using NELs.

2.4.2.2 FERMO. A recently redesigned and updated tool that supports reproducible prioritization of samples and metabolite features based on bioactivity or other types of phenotypic information is FERMO.[110] The main contribution of FERMO is bringing together the heterogeneous data types required to jointly assess during the prioritization process. By eliminating metabolite features that are unlikely to be relevant for your study, *i.e.*, because they are of known structures or unlikely to be correlated to bioactivity, a reduced set of metabolite features remains for further validation. Currently, FERMO uses filters based on novelty, uniqueness, specificity, and phenotype-association. By design, scalability was accounted for by separating the FERMO core code for data integration and metric calculation from the subsequent visualization through a dashboard that is also embedded in an online web application (**https://fermo.bionformatics.nl**). Doing so, a user could run FERMO on a larger set of metabolomics profiles with associated metadata and activity data, without the direct need to visualize all data in a dashboard. For example, a selection of the metabolite features could be extracted from the resulting tables based on a set of filters or thresholds, for example on feature abundance (see also Sections 2.2.3 and 2.2.4). Another limitation on the running time of FERMO is the size of the MS library used for annotations: with the current matchms-based implementation of performing mass spectral comparisons, adding annotations from the large GNPS public library (containing half a million reference spectra) may take a while (several hours) to complete; hence, a smaller (approximately hundreds to a few thousand of spectra) focussed (*i.e.*, with relevant molecules for the samples type) library is currently recommended. As library matches form an integral part of the FERMO novelty score, future work will aim to tackle this scalability issue by implementing novel solutions as also described in this review. Finally, FERMO's output can be stored and shared with the scientific community in a reusable format as a starting point for future exploratory data analysis studies.

2.4.2.3 MS2DECIDE. MS2DECIDE[111] is a white-box recommendation tool that prioritizes candidate spectra by integrating complementary annotations from three engines—GNPS, SIRIUS, and ISDB-LOTUS—*via* a decision-theoretic function that integrates expert knowledge in NPs. For each spectrum, the three engines each propose a best-match structure with an associated similarity/confidence score; pairwise Tanimoto coefficients among these structures add three more metrics. These six values feed into the expert-tuned function to produce

a "knownness" score, which ranks candidates by their likelihood of novelty. Although its model currently relies on a single expert input (potentially introducing bias), MS2DECIDE offers transparent, customizable reasoning and lays the groundwork for hybrid approaches that combine multiple expert insights with automated optimization.

2.4.2.4 msFeaST. The experimental design of a metabolomics study typically gives rise to several relevant comparisons between sample groups to answer the research question(s) at hand for which various statistical approaches are available and used in the metabolomics field. In NELs exploration, these are typically related to bioactivity and structural novelty. Current metabolomics discovery workflows mostly perform statistical analysis as one of the final steps to prioritize the relevant metabolite features. A recent tool, however, integrates statistical analyses and metabolite grouping, the results of which can be loaded in an interactive dashboard. As such, msFeaST offers a complementary workflow to molecular networking and FERMO.[112] In the overview, metabolite features are displayed as nodes on a chemical map. In that map, drawn based on mass spectral embeddings computed based on a mass spectral similarity score of choice (*i.e.*, modified cosine[52] or MS2DeepScore[64,65]), the size of the nodes is proportional to statistical input: *i.e.*, the log 2-fold change between two specified groups based on the provided metadata, or the *p*-value that is associated with the feature-set testing, a group-based statistical approach. The larger the node, the more relevant it may be for the differentiation between the two groups, giving the user a clue on which parts of the chemical map to focus on for further analyses. In NELs exploration, areas in the chemical landscape with larger nodes could correspond to metabolite feature groups with bioactivity using metadata that contains the outcome of performed bioassays. What makes msFeaST unique, is its novel approach of grouping the features using the similarity matrix, and its integrated networking capabilities by interactively showing which metabolite features would be connected in a molecular network if one would be constructed. However, such approaches come with limitations in scalability, due to several reasons related to humans and computers. For example, if the chemical landscape in the form of a mass spectral embedding goes beyond 5–10 thousand features, it is very difficult for humans to keep the oversight, and for interactive dashboards to remain responsive when the user queries it. This has to do with the visual limits of intake of humans on the one hand, and the requirement that all the information and sub panels are lined on the other hand.[113] Therefore, dedicated choices in sample set selection and feature thresholding (as discussed in Sections 2.2.3 and 2.2.4), *i.e.*, to reduce noise, are recommended before using tools such as FERMO and msFeaST.

3. Concluding remarks/outlook

The exploration of NELs sits at the intersection of multiple disciplines—from analytical chemistry and computational biology to pharmacognosy, data science, and ecology. Realizing the full potential of these rich biochemical resources requires a deeply multidisciplinary approach, supported by a commitment to open science and community-driven knowledge sharing. Initiatives like ENPKG[3] exemplify this vision: creating interoperable, reusable, and openly accessible knowledge infrastructures that allow diverse communities to collaboratively build and refine understanding. This strategy will be the way forward for unlocking the full potential of NP research in the data-driven era and we encourage developments in this direction. In the same spirit, the future of MS-based metabolomics for NELs exploration must embrace open, extensible platforms where data, tools, and annotations are treated as shareable, evolving assets—catalyzing reproducibility, collective intelligence, and global collaboration.

As datasets grow in volume and complexity, significant challenges remain in scaling both data acquisition and computational interpretation. Despite the expansion of NP-oriented spectral repositories like GNPS, microbeMASST, and plantMASST, or structural ones like COCONUT,[114] LOTUS,[98] or NPAtlas,[115] these resources still represent only a narrow window into the immense chemodiversity of NPs.

To address these limitations, future methodologies must integrate scalable computational strategies with high-quality experimental design. Scalable NELs analysis offers transformative potential: enabling the construction of global chemical atlases, revealing biosynthetic patterns[116] across ecosystems, and accelerating the discovery of molecules with pharmaceutical, agricultural, or ecological relevance. The integration of metabolomics with genomics,[117] biosynthetic pathway prediction,[118] and phenotypic profiling[104,110] will further strengthen efforts to assign function to the molecular dark matter of nature.

The strive for robust, high-resolution data acquisition strategies that maximize spectral quality and coverage will be critical. Limitations in chromatography—such as peak capacity and scan speed—must be addressed alongside rigorous preprocessing, to ensure downstream tools operate on meaningful signals. In large-scale studies, where sample numbers and molecular features multiply rapidly, these foundational considerations are critical for ensuring scalable and reproducible insights.

On the computational side, continued innovation in spectral matching, annotation algorithms, and interactive data summarization will be key. Community-driven efforts like GNPS, ReDU,[119] as well as more centrally built platforms like the Molecules Gateway[69] are already demonstrating how open infrastructure and automated workflows can scale annotation and analysis across thousands of samples. At the same time, spectral libraries must grow not only in size, but in sophistication—supporting rich metadata integration and enabling annotation inheritance across molecular families. Mass spectral embeddings or foundational molecular networks represent promising directions for organizing and contextualizing this knowledge more effectively. This will both help the community to better map biochemical diversity, even when more and larger datasets are being generated, and to better assess novelty across samples and features. The latter will be essential to guide our attention to the most relevant and promising NPs for further research.

Once a few thousand metabolite features have been selected to sieve through, interactive visualization becomes feasible, and an (online) dashboard can function as a central place to interact with all the heterogeneous data that is available for those features. This allows for, *i.e.*, on-the-fly adjustment of filter settings, and when all panels of the dashboard are linked and updated, this will facilitate the selection of NELs metabolite features for further validation. Especially with the multi-facetted datasets to be considered, visualization is key during the decision-making process to ensure that parameter settings, filters, and thresholds have the desired effects whilst minimizing any unwanted side-effects.

Finally, the software infrastructure that powers metabolomics must evolve with the same attention to scalability and openness. Modular design, transparent documentation, and community-maintained pipelines are vital for long-term sustainability and innovation. These principles not only support technical robustness, but also echo the broader goals of projects like Wikifunctions (**https://www.wikifunctions.org/wiki/Wikifunctions:Main_Page**): creating a world where functional knowledge is openly available, composable, and interoperable across contexts.

In summary, building a scalable and collaborative future for NELs metabolomics will require coordinated progress across data generation, analysis, software development, and community governance. By embracing open science, multi-disciplinarity, and shared infrastructure, we can transform how we explore, understand, prioritize, and utilize the extraordinary chemical diversity encoded in the natural world.

4. Author contributions

Adriano Rutz: conceptualization, writing – original draft, review & editing, validation, visualization. Wout Bittremieux: writing – original draft, review & editing, validation, visualization. Robin Schmid: writing – original draft, review & editing, visualization. Olivier Cailloux: writing – review & editing. Justin J. J. van der Hooft: conceptualization, writing – original draft, review & editing, validation. Mehdi A. Beniddir: conceptualization, writing – original draft, review & editing, validation, visualization.

5. Conflicts of interest

J. J. J. vdH. is member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy, and consults for Corteva Agriscience, Indianapolis, IN, USA. R. S. is a co-founder and employee of mzio GmbH (Bremen, Germany), which develops the mzmine software for MS data processing. All other authors declare to have no competing interests.

6. Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

7. Acknowledgements

M. A. B. was supported by the National French Agency (ANR Grant 15-CE29-0001) and CNRS PRIME 80 MITI. W. B. acknowledges support by the Research Foundation–Flanders (FWO grant G0AHY25N). R. S. and the development of mzmine is funded by the European Union, the BAB – Funding Bank for Bremen and Bremerhaven, and the Senator of Economics, Ports and Transformation Bremen (65002459).

8. Notes and references

1 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, International Natural Product Sciences Taskforce and C. T. Supuran, *Nat. Rev. Drug Discovery*, 2021, **20**, 200–216.
2 B. A. P. Wilson, C. C. Thornburg, C. J. Henrich, T. Grkovic and B. R. O'Keefe, *Nat. Prod. Rep.*, 2020, **37**, 893–918.
3 P.-M. Allard, A. Gaudry, L.-M. Quirós-Guerrero, A. Rutz, M. Dounoue-Kubo, T. W. N. Walker, E. Defossez, C. Long, A. Grondin, B. David and J.-L. Wolfender, *Gigascience*, 2023, **12**, giac124.
4 F. Olivon, P.-M. Allard, A. Koval, D. Righi, G. Genta-Jouve, J. Neyts, C. Apel, C. Pannecouque, L.-F. Nothias, X. Cachet, L. Marcourt, F. Roussi, V. L. Katanaev, D. Touboul, J.-L. Wolfender and M. Litaudon, *ACS Chem. Biol.*, 2017, **12**, 2644–2651.
5 D. J. Floros, P. R. Jensen, P. C. Dorrestein and N. Koyama, *Metabolomics*, 2016, **12**, 145.
6 M. A. Cook, D. Pallant, L. Ejim, A. D. Sutherland, X. Wang, J. W. Johnson, S. McCusker, X. Chen, M. George, S. Chou, K. Koteva, W. Wang, C. Hobson, D. Hackenberger, N. Waglechner, O. Ejim, T. Campbell, R. Medina, L. T. MacNeil and G. D. Wright, *J. Ind. Microbiol. Biotechnol.*, 2023, **50**, kuad042, DOI: **10.1093/jimb/kuad042**.
7 G. G. Conrado, R. da Rosa, R. D. Reis and L. R. Pessa, *Rev. Bras. Farmacogn.*, 2024, **34**, 706–721.
8 J.-L. Wolfender, M. Litaudon, D. Touboul and E. F. Queiroz, *Nat. Prod. Rep.*, 2019, **36**, 855–868.
9 G. Hajjar, M. C. Barros Santos, J. Bertrand-Michel, C. Canlet, F. Castelli, N. Creusot, S. Dechaumet, B. Diémé, F. Giacomoni, P. Giraudeau, Y. Guitton, E. Thévenot, M. Tremblay-Franco, C. Junot, F. Jourdan, F. Fenaille, B. Comte, P. Pétriacq and E. Pujos-Guillot, *TrAC, Trends Anal. Chem.*, 2023, **167**, 117225.
10 S. R. Johnson and B. M. Lange, *Front. Bioeng. Biotechnol.*, 2015, **3**, 22.
11 M. J. J. Recchia, T. U. H. Baumeister, D. Y. Liu and R. G. Linington, *Anal. Chem.*, 2023, **95**, 11908–11917.
12 N. Drouin, M. van Mever, W. Zhang, E. Tobolkina, S. Ferre, A.-C. Servais, M.-J. Gou, L. Nyssen, M. Fillet, G. S. M. Lageveen-Kammeijer, J. Nouta, A. J. Chetwynd, I. Lynch, J. A. Thorn, J. Meixner, C. Lößner, M. Taverna, S. Liu, N. T. Tran, Y. Francois, A. Lechner, R. Nehmé, G. Al Hamoui Dit Banni, R. Nasreddine, C. Colas, H. H. Lindner, K. Faserl, C. Neusüß, M. Nelke, S. Lämmerer, C. Perrin, C. Bich-Muracciole, C. Barbas, Á. L. Gonzálvez, A. Guttman, M. Szigeti, P. Britz-

McKibbin, Z. Kroezen, M. Shanmuganathan, P. Nemes, E. P. Portero, T. Hankemeier, S. Codesido, V. González-Ruiz, S. Rudaz and R. Ramautar, *Anal. Chem.*, 2020, **92**, 14103–14112.
13 B. van de Velde, D. Guillarme and I. Kohler, *J. Chromatogr. B:Anal. Technol. Biomed. Life Sci.*, 2020, **1161**, 122444.
14 S. Girel, M. Galmiche, M. Fiault, V. Mieville, P. Nowak-Sliwinska, S. Rudaz and I. Meister, *Anal. Chem.*, 2025, **97**, 5109–5117.
15 B. Wei, L. Dai and K. Zhang, *J. Chromatogr. A*, 2025, **1739**, 465524.
16 K. Zhang and X. Liu, *J. Pharm. Biomed. Anal.*, 2016, **128**, 73–88.
17 R. Montes, R. Rodil, L. Placer, J. M. Wilms, R. Cela and J. B. Quintana, *Anal. Bioanal. Chem.*, 2020, **412**, 4849–4856.
18 D. I. Falev, D. S. Kosyakov, N. V. Ul'yanovskii and D. V. Ovchinnikov, *J. Chromatogr. A*, 2020, **1609**, 460458.
19 W. J. Nash, J. B. Ngere, L. Najdekr and W. B. Dunn, *Anal. Chem.*, 2024, **96**, 10935–10942.
20 Y. El Abiead, A. Rutz, S. Zuffa, B. Amer, S. Xing, C. Brungs, R. Schmid, M. S. P. Correia, A. M. Caraballo-Rodriguez, A. Zarrinpar, H. Mannochio-Russo, M. Witting, I. Mohanty, T. Pluskal, W. Bittremieux, R. Knight, A. D. Patterson, J. J. J. van der Hooft, S. Böcker, W. B. Dunn, R. G. Linington, D. S. Wishart, J.-L. Wolfender, O. Fiehn, N. Zamboni and P. C. Dorrestein, *Nat. Metab.*, 2025, **7**, 435–437.
21 C. D. Broeckling, A. Ganna, M. Layer, K. Brown, B. Sutton, E. Ingelsson, G. Peers and J. E. Prenni, *Anal. Chem.*, 2016, **88**, 9226–9234.
22 F. Kong, U. Keshet, T. Shen, E. Rodriguez and O. Fiehn, *Anal. Chem.*, 2023, **95**, 16810–16818.
23 J. Wandy, R. McBride, S. Rogers, N. Terzis, S. Weidt, J. J. J. van der Hooft, K. Bryson, R. Daly and V. Davies, *Front. Mol. Biosci.*, 2023, **10**, 1130781.
24 L. Martens, M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Römpp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P.-A. Binz and E. W. Deutsch, *Mol. Cell. Proteomics*, 2011, **10**, R110.000133.
25 N. Hoffmann, J. Rein, T. Sachsenberg, J. Hartler, K. Haug, G. Mayer, O. Alka, S. Dayalan, J. T. M. Pearce, P. Rocca-Serra, D. Qi, M. Eisenacher, Y. Perez-Riverol, J. A. Vizcaíno, R. M. Salek, S. Neumann and A. R. Jones, *Anal. Chem.*, 2019, **91**, 3302–3310.
26 M. Larralde, T. N. Lawson, R. J. M. Weber, P. Moreno, K. Haug, P. Rocca-Serra, M. R. Viant, C. Steinbeck and R. M. Salek, *Bioinformatics*, 2017, **33**, 2598–2600.
27 J. Klein, H. Lam, T. D. Mak, W. Bittremieux, Y. Perez-Riverol, R. Gabriels, J. Shofstahl, H. Hecht, P.-A. Binz, S. Kawano, T. Van Den Bossche, J. Carver, B. A. Neely, L. Mendoza, T. Suomi, T. Claeys, T. Payne, D. Schulte, Z. Sun, N. Hoffmann, Y. Zhu, S. Neumann, A. R. Jones, N. Bandeira, J. A. Vizcaíno and E. W. Deutsch, *Anal. Chem.*, 2024, **96**, 18491–18501.
28 H. L. Röst, U. Schmitt, R. Aebersold and L. Malmström, *Proteomics*, 2014, **14**, 74–77.
29 F. Huber, S. Verhoeven, C. Meijer, H. Spreeuw, E. Castilla, C. Geng, J. van der Hooft, S. Rogers, A. Belloum, F. Diblen and J. Spaaks, *J. Open Source Softw.*, 2020, **5**, 2411.
30 W. Bittremieux, *Anal. Chem.*, 2020, **92**, 659–661.
31 L. Gatto and K. S. Lilley, *Bioinformatics*, 2012, **28**, 288–289.
32 R. Schmid, S. Heuckeroth, A. Korf, A. Smirnov, O. Myers, T. S. Dyrlund, R. Bushuiev, K. J. Murray, N. Hoffmann, M. Lu, A. Sarvepalli, Z. Zhang, M. Fleischauer, K. Dührkop, M. Wesner, S. J. Hoogstra, E. Rudt, O. Mokshyna, C. Brungs, K. Ponomarov, L. Mutabdžija, T. Damiani, C. J. Pudney, M. Earll, P. O. Helmer, T. R. Fallon, T. Schulze, A. Rivas-Ubach, A. Bilbao, H. Richter, L.-F. Nothias, M. Wang, M. Orešič, J.-K. Weng, S. Böcker, A. Jeibmann, H. Hayen, U. Karst, P. C. Dorrestein, D. Petras, X. Du and T. Pluskal, *Nat. Biotechnol.*, 2023, **41**, 447–449.
33 X. Domingo-Almenara and G. Siuzdak, *Methods Mol. Biol.*, 2020, **2104**, 11–24.
34 Z. Pang, Y. Lu, G. Zhou, F. Hui, L. Xu, C. Viau, A. F. Spigelman, P. E. MacDonald, D. S. Wishart, S. Li and J. Xia, *Nucleic Acids Res.*, 2024, **52**, W398–W406.
35 H. Takeda, Y. Matsuzawa, M. Takeuchi, M. Takahashi, K. Nishida, T. Harayama, Y. Todoroki, K. Shimizu, N. Sakamoto, T. Oka, M. Maekawa, M. H. Chung, Y. Kurizaki, S. Kiuchi, K. Tokiyoshi, B. Buyantogtokh, M. Kurata, A. Kvasnička, U. Takeda, H. Uchino, M. Hasegawa, J. Miyamoto, K. Tanabe, S. Takeda, T. Mori, R. Kumakubo, T. Tanaka, T. Yoshino, M. Okamoto, H. Takahashi, M. Arita and H. Tsugawa, *Nat. Commun.*, 2024, **15**, 9903.
36 J. Pfeuffer, C. Bielow, S. Wein, K. Jeong, E. Netz, A. Walter, O. Alka, L. Nilse, P. D. Colaianni, D. McCloskey, J. Kim, G. Rosenberger, L. Bichmann, M. Walzer, J. Veit, B. Boudaud, M. Bernt, N. Patikas, M. Pilz, M. P. Startek, S. Kutuzova, L. Heumos, J. Charkow, J. C. Sing, A. Feroz, A. Siraj, H. Weisser, T. M. H. Dijkstra, Y. Perez-Riverol, H. Röst, O. Kohlbacher and T. Sachsenberg, *Nat. Methods*, 2024, **21**, 365–367.
37 S. Heuckeroth, T. Damiani, A. Smirnov, O. Mokshyna, C. Brungs, A. Korf, J. D. Smith, P. Stincone, N. Dreolin, L.-F. Nothias, T. Hyötyläinen, M. Orešič, U. Karst, P. C. Dorrestein, D. Petras, X. Du, J. J. J. van der Hooft, R. Schmid and T. Pluskal, *Nat. Protoc.*, 2024, **19**, 2597–2641.
38 J. P. Shaffer, L.-F. Nothias, L. R. Thompson, J. G. Sanders, R. A. Salido, S. P. Couvillion, A. D. Brejnrod, F. Lejzerowicz, N. Haiminen, S. Huang, H. L. Lutz, Q. Zhu, C. Martino, J. T. Morton, S. Karthikeyan, M. Nothias-Esposito, K. Dührkop, S. Böcker, H. W. Kim, A. A. Aksenov, W. Bittremieux, J. J. Minich, C. Marotz, M. M. Bryant, K. Sanders, T. Schwartz, G. Humphrey, Y. Vásquez-Baeza, A. Tripathi, L. Parida, A. P. Carrieri, K. L. Beck, P. Das, A. González, D. McDonald, J. Ladau, S. M. Karst, M. Albertsen, G. Ackermann, J. DeReus, T. Thomas, D. Petras, A. Shade, J. Stegen, S. J. Song, T. O. Metz, A. D. Swafford, P. C. Dorrestein, J. K. Jansson,

J. A. Gilbert, R. Knight and Earth Microbiome Project 500 (EMP500) Consortium, *Nat. Microbiol.*, 2022, **7**, 2128–2150.
39 A. Gaudry, F. Huber, L.-F. Nothias, S. Cretton, M. Kaiser, J.-L. Wolfender and P.-M. Allard, *Front. Bioinform.*, 2022, **2**, 842964.
40 A. Gaudry, M. Pagni, F. Mehl, S. Moretti, L.-M. Quiros-Guerrero, L. Cappelletti, A. Rutz, M. Kaiser, L. Marcourt, E. F. Queiroz, J.-R. Ioset, A. Grondin, B. David, J.-L. Wolfender and P.-M. Allard, *ACS Cent. Sci.*, 2024, **10**, 494–510.
41 D. Meijer, M. A. Beniddir, C. W. Coley, Y. M. Mejri, M. Öztürk, J. J. J. van der Hooft, M. H. Medema and A. Skiredj, *Nat. Prod. Rep.*, 2025, **42**, 654–662.
42 R. Schmid, D. Petras, L.-F. Nothias, M. Wang, A. T. Aron, A. Jagels, H. Tsugawa, J. Rainer, M. Garcia-Aloy, K. Dührkop, A. Korf, T. Pluskal, Z. Kameník, A. K. Jarmusch, A. M. Caraballo-Rodríguez, K. C. Weldon, M. Nothias-Esposito, A. A. Aksenov, A. Bauermeister, A. Albarracin Orio, C. O. Grundmann, F. Vargas, I. Koester, J. M. Gauglitz, E. C. Gentry, Y. Hövelmann, S. A. Kalinina, M. A. Pendergraft, M. Panitchpakdi, R. Tehan, A. Le Gouellec, G. Aleti, H. Mannochio Russo, B. Arndt, F. Hübner, H. Hayen, H. Zhi, M. Raffatellu, K. A. Prather, L. I. Aluwihare, S. Böcker, K. L. McPhail, H.-U. Humpf, U. Karst and P. C. Dorrestein, *Nat. Commun.*, 2021, **12**, 3832.
43 T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*, 1958.
44 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.
45 A. Capecchi, D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, **12**, 43.
46 D. R. Flower, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 379–386.
47 F. Kretschmer, J. Seipp, M. Ludwig, G. W. Klau and S. Boecker, *Nat. Commun.*, 2025, **16**, 554.
48 H. Shiokawa, S. Ishida and K. Terayama, *J. Cheminf.*, 2025, **17**, 57.
49 S. E. Stein and D. R. Scott, *J. Am. Soc. Mass Spectrom.*, 1994, **5**, 859–866.
50 W. Bittremieux, P. Meysman, W. S. Noble and K. Laukens, *J. Proteome Res.*, 2018, **17**, 3463–3474.
51 T. Onoprishvili, J.-H. Yuan, K. Petrov, V. Ingalalli, L. Khederlarian, N. Leuchtenmuller, S. Chandra, A. Duarte, A. Bender and Y. Gloaguen, *Bioinformatics*, 2025, **41**, btaf081.
52 W. Bittremieux, R. Schmid, F. Huber, J. J. J. van der Hooft, M. Wang and P. C. Dorrestein, *J. Am. Soc. Mass Spectrom.*, 2022, **33**, 1733–1744.
53 J. Watrous, P. Roach, T. Alexandrov, B. S. Heath, J. Y. Yang, R. D. Kersten, M. van der Voort, K. Pogliano, H. Gross, J. M. Raaijmakers, B. S. Moore, J. Laskin, N. Bandeira and P. C. Dorrestein, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, E1743–E1752.
54 A. T. Kong, F. V. Leprevost, D. M. Avtonomov, D. Mellacheruvu and A. I. Nesvizhskii, *Nat. Methods*, 2017, **14**, 513–520.
55 M. Mongia, T. M. Yasaka, Y. Liu, M. Guler, L. Lu, A. Bhagwat, B. Behsaz, M. Wang, P. C. Dorrestein and H. Mohimani, *Nat. Biotechnol.*, 2024, **42**, 1672–1677.
56 Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta and O. Fiehn, *Nat. Methods*, 2021, **18**, 1524–1531.
57 Y. Li and O. Fiehn, *Nat. Methods*, 2023, **20**, 1475–1478.
58 W. Bittremieux, N. E. Avalon, S. P. Thomas, S. A. Kakhkhorov, A. A. Aksenov, P. W. P. Gomes, C. M. Aceves, A. M. Caraballo-Rodríguez, J. M. Gauglitz, W. H. Gerwick, T. Huan, A. K. Jarmusch, R. F. Kaddurah-Daouk, K. B. Kang, H. W. Kim, T. Kondić, H. Mannochio-Russo, M. J. Meehan, A. V. Melnik, L.-F. Nothias, C. O'Donovan, M. Panitchpakdi, D. Petras, R. Schmid, E. L. Schymanski, J. J. J. van der Hooft, K. C. Weldon, H. Yang, S. Xing, J. Zemlin, M. Wang and P. C. Dorrestein, *Nat. Commun.*, 2023, **14**, 8488.
59 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. B. P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. O. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.
60 K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. de Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin and C. Steinbeck, *Nucleic Acids Res.*, 2013, **41**, D781–D786.
61 M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K. S. Nair, S. Sumner and S. Subramaniam, *Nucleic Acids Res.*, 2016, **44**, D463–D470.

62 N. F. de Jonge, K. Mildau, D. Meijer, J. J. R. Louwen, C. Bueschl, F. Huber and J. J. J. van der Hooft, *Metabolomics*, 2022, **18**, 103.
63 F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers and J. J. J. van der Hooft, *PLoS Comput. Biol.*, 2021, **17**, e1008724.
64 F. Huber, S. van der Burg, J. J. J. van der Hooft and L. Ridder, *J. Cheminf.*, 2021, **13**, 84.
65 N. F. de Jonge, E. Chekmeneva, R. Schmid, D. Joas, L.-J. Truong, J. J. J. van der Hooft and F. Huber, *bioRxiv*, 2025, preprint, DOI: **10.1101/2024.03.25.586580**.
66 R. Bushuiev, A. Bushuiev, N. F. de Jonge, A. Young, F. Kretschmer, R. Samusevich, J. Heirman, F. Wang, L. Zhang, K. Dührkop, M. Ludwig, N. A. Haupt, A. Kalia, C. Brungs, R. Schmid, R. Greiner, B. Wang, D. S. Wishart, L.-P. Liu, J. Rousu, W. Bittremieux, H. Rost, T. D. Mak, S. Hassoun, F. Huber, J. J. J. van der Hooft, M. A. Stravs, S. Böcker, J. Sivic and T. Pluskal, *Adv. Neural Inf. Process. Syst.*, 2025, **37**, 110010–110027.
67 S. Szwarc, A. Rutz, K. Lee, Y. Mejri, O. Bonnet, H. Hazni, A. Jagora, R. B. Mbeng Obame, J. K. Noh, E. Otogo N'Nang, S. C. Alaribe, K. Awang, G. Bernadat, Y. H. Choi, V. Courdavault, M. Frederich, T. Gaslonde, F. Huber, T.-S. Kam, Y. Y. Low, E. Poupon, J. J. J. van der Hooft, K. B. Kang, P. Le Pogam and M. A. Beniddir, *J. Cheminf.*, 2025, **17**, 62.
68 N. F. de Jonge, J. J. R. Louwen, E. Chekmeneva, S. Camuzeaux, F. J. Vermeir, R. S. Jansen, F. Huber and J. J. J. van der Hooft, *Nat. Commun.*, 2023, **14**, 1752.
69 M. Simone, M. Iorio, P. Monciardini, M. Santini, N. Cantù, A. Tocchetti, S. Serina, C. Brunati, T. Vernay, A. Gentile, M. Aracne, M. Cozzi, J. J. J. van der Hooft, M. Sosio, S. Donadio and S. I. Maffioli, *J. Nat. Prod.*, 2024, **87**, 2615–2628.
70 L. Cao, M. Guler, A. Tagirdzhanov, Y.-Y. Lee, A. Gurevich and H. Mohimani, *Nat. Commun.*, 2021, **12**, 3718.
71 R. R. da Silva, M. Wang, L.-F. Nothias, J. J. J. van der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes and P. C. Dorrestein, *PLoS Comput. Biol.*, 2018, **14**, e1006089.
72 Z. Zhou, M. Luo, H. Zhang, Y. Yin, Y. Cai and Z.-J. Zhu, *Nat. Commun.*, 2022, **13**, 6656.
73 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.
74 D. Olivier-Jimenez, M. Chollet-Krugler, D. Rondeau, M. A. Beniddir, S. Ferron, T. Delhaye, P.-M. Allard, J.-L. Wolfender, H. J. M. Sipman, R. Lücking, J. Boustie and P. Le Pogam, *Sci. Data*, 2019, **6**, 294.
75 A. E. Fox Ramos, P. Le Pogam, C. Fox Alcover, E. Otogo N'Nang, G. Cauchie, H. Hazni, K. Awang, D. Bréard, A. M. Echavarren, M. Frédérich, T. Gaslonde, M. Girardot, R. Grougnet, M. S. Kirillova, M. Kritsanida, C. Lémus, A.-M. Le Ray, G. Lewin, M. Litaudon, L. Mambu, S. Michel, F. M. Miloserdov, M. E. Muratore, P. Richomme-Peniguel, F. Roussi, L. Evanno, E. Poupon, P. Champy and M. A. Beniddir, *Sci. Data*, 2019, **6**, 15.
76 S. A. Agnès, T. Okpekon, Y. A. Kouadio, A. Jagora, D. Bréard, E. V. Costa, F. M. A. da Silva, H. H. F. Koolen, A.-M. Le Ray-Richomme, P. Richomme, P. Champy, M. A. Beniddir and P. Le Pogam, *Sci. Data*, 2022, **9**, 270.
77 C. Brungs, R. Schmid, S. Heuckeroth, A. Mazumdar, M. Drexler, P. Šácha, P. C. Dorrestein, D. Petras, L.-F. Nothias, V. Veverka, R. Nencka, Z. Kameník and T. Pluskal, *ChemRxiv*, 2025, preprint, DOI: **10.26434/chemrxiv-2024-l1tgh-v3**.
78 C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender and S. Neumann, *J. Cheminf.*, 2016, **8**, 3.
79 F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner and D. S. Wishart, *Anal. Chem.*, 2021, **93**, 11692–11700.
80 Y. Nowatzky, F. F. Russo, J. Lisec, A. Kister, K. Reinert, T. Muth and P. Benner, *Nat. Commun.*, 2025, **16**, 2298.
81 P.-M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, V. C. Pham, F. Roussi, M. Litaudon and J.-L. Wolfender, *Anal. Chem.*, 2016, **88**, 3317–3323.
82 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12580–12585.
83 K. Dührkop, L.-F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein and S. Böcker, *Nat. Biotechnol.*, 2021, **39**, 462–471.
84 M. Ludwig, L.-F. Nothias, K. Dührkop, I. Koester, M. Fleischauer, M. A. Hoffmann, D. Petras, F. Vargas, M. Morsy, L. Aluwihare, P. C. Dorrestein and S. Böcker, *Nat. Mach. Intell.*, 2020, **2**, 629–641.
85 M. A. Hoffmann, L.-F. Nothias, M. Ludwig, M. Fleischauer, E. C. Gentry, M. Witting, P. C. Dorrestein, K. Dührkop and S. Böcker, *Nat. Biotechnol.*, 2022, **40**, 411–421.
86 M. A. Stravs, K. Dührkop, S. Böcker and N. Zamboni, *Nat. Methods*, 2022, **19**, 865–870.
87 J. J. J. van der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess and S. Rogers, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 13738–13743.
88 S. Rogers, C. W. Ong, J. Wandy, M. Ernst, L. Ridder and J. J. J. van der Hooft, *Faraday Discuss.*, 2019, **218**, 284–302.
89 Y. Liu, A. Mrzic, P. Meysman, T. De Vijlder, E. P. Romijn, D. Valkenborg, W. Bittremieux and K. Laukens, *PLoS One*, 2020, **15**, e0226770.
90 J. J. J. van der Hooft, J. Wandy, F. Young, S. Padmanabhan, K. Gerasimidis, K. E. V. Burgess, M. P. Barrett and S. Rogers, *Anal. Chem.*, 2017, **89**, 7569–7577.
91 L. R. Torres Ortega, J. Dietrich, J. Wandy, H. Mol and J. J. J. van der Hooft, *bioRxiv*, 2025, preprint, DOI: **10.1101/2025.06.19.659491**.
92 H. Sepman, L. Malm, P. Peets, M. MacLeod, J. Martin, M. Breitholtz and A. Kruve, *Anal. Chem.*, 2023, **95**, 12329–12338.

93 N. Meekel, A. Kruve, M. H. Lamoree and F. M. Been, *Environ. Sci. Technol.*, 2025, **59**, 5056–5065, DOI: 10.1021/acs.est.4c10498.
94 A. Capecchi and J.-L. Reymond, *J. Cheminf.*, 2021, **13**, 82.
95 M. Visani, *ARPHA Preprints*, 2023, preprint, DOI: 10.3897/arphapreprints.e116230.
96 S. Zuffa, R. Schmid, A. Bauermeister, P. W. P Gomes, A. M. Caraballo-Rodriguez, Y. El Abiead, A. T. Aron, E. C. Gentry, J. Zemlin, M. J. Meehan, N. E. Avalon, R. H. Cichewicz, E. Buzun, M. C. Terrazas, C.-Y. Hsu, R. Oles, A. V. Ayala, J. Zhao, H. Chu, M. C. M. Kuijpers, S. L. Jackrel, F. Tugizimana, L. P. Nephali, I. A. Dubery, N. E. Madala, E. A. Moreira, L. V. Costa-Lotufo, N. P. Lopes, P. Rezende-Teixeira, P. C. Jimenez, B. Rimal, A. D. Patterson, M. F. Traxler, R. de C. Pessotti, D. Alvarado-Villalobos, G. Tamayo-Castillo, P. Chaverri, E. Escudero-Leyva, L.-M. Quiros-Guerrero, A. J. Bory, J. Joubert, A. Rutz, J.-L. Wolfender, P.-M. Allard, A. Sichert, S. Pontrelli, B. S. Pullman, N. Bandeira, W. H. Gerwick, K. Gindro, J. Massana-Codina, B. C. Wagner, K. Forchhammer, D. Petras, N. Aiosa, N. Garg, M. Liebeke, P. Bourceau, K. B. Kang, H. Gadhavi, L. P. S. de Carvalho, M. Silva Dos Santos, A. I. Pérez-Lorente, C. Molina-Santiago, D. Romero, R. Franke, M. Brönstrup, A. Vera Ponce de León, P. B. Pope, S. L. La Rosa, G. La Barbera, H. M. Roager, M. F. Laursen, F. Hammerle, B. Siewert, U. Peintner, C. Licona-Cassani, L. Rodriguez-Orduña, E. Rampler, F. Hildebrand, G. Koellensperger, H. Schoeny, K. Hohenwallner, L. Panzenboeck, R. Gregor, E. C. O'Neill, E. T. Roxborough, J. Odoi, N. J. Bale, S. Ding, J. S. Sinninghe Damsté, X. L. Guan, J. J. Cui, K.-S. Ju, D. B. Silva, F. M. R. Silva, G. F. da Silva, H. H. F. Koolen, C. Grundmann, J. A. Clement, H. Mohimani, K. Broders, K. L. McPhail, S. E. Ober-Singleton, C. M. Rath, D. McDonald, R. Knight, M. Wang and P. C. Dorrestein, *Nat. Microbiol.*, 2024, **9**, 336–345.
97 A. Rutz, M. Dounoue-Kubo, S. Ollivier, J. Bisson, M. Bagheri, T. Saesong, S. N. Ebrahimi, K. Ingkaninan, J.-L. Wolfender and P.-M. Allard, *Front. Plant Sci.*, 2019, **10**, 1329.
98 A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson and P.-M. Allard, *Elife*, 2022, **11**, e70780, DOI: 10.7554/eLife.70780.
99 K. A. West, R. Schmid, J. M. Gauglitz, M. Wang and P. C. Dorrestein, *npj Sci. Food*, 2022, **6**, 22.
100 P. W. P. Gomes, H. Mannochio-Russo, R. Schmid, S. Zuffa, T. Damiani, L.-M. Quiros-Guerrero, A. M. Caraballo-Rodríguez, H. N. Zhao, H. Yang, S. Xing, V. Charron-Lamoureux, D. N. Chigumba, B. E. Sedio, J. A. Myers, P.-M. Allard, T. V. Harwood, G. Tamayo-Castillo, K. B. Kang, E. Defossez, H. H. F. Koolen, M. N. da Silva, C. Y. Y. E Silva, S. Rasmann, T. W. N. Walker, G. Glauser, J. M. Chaves-Fallas, B. David, H. Kim, K. H. Lee, M. J. Kim, W. J. Choi, Y.-S. Keum, E. J. S. P. de Lima, L. S. de Medeiros, G. A. Bataglion, E. V. Costa, F. M. A. da Silva, A. R. V. Carvalho, J. D. E. Reis, S. Pamplona, E. Jeong, K. Lee, G. J. Kim, Y.-S. Kil, J.-W. Nam, H. Choi, Y. K. Han, S. Y. Park, K. Y. Lee, C. Hu, Y. Dong, S. Sang, C. R. Morrison, R. M. Borges, A. M. Teixeira, S. Y. Lee, B. S. Lee, S. Y. Jeong, K. H. Kim, A. Rutz, A. Gaudry, E. Bruelhart, I. F. Kappers, R. Karlova, M. Meisenburg, R. Berdaguer, J. S. Tello, D. Henderson, L. Cayola, S. J. Wright, D. N. Allen, K. J. Anderson-Teixeira, J. L. Baltzer, J. A. Lutz, S. M. McMahon, G. G. Parker, J. D. Parker, T. R. Northen, B. P. Bowen, T. Pluskal, J. J. J. van der Hooft, J. J. Carver, N. Bandeira, B. S. Pullman, J.-L. Wolfender, R. D. Kersten, M. Wang and P. C. Dorrestein, *bioRxiv*, 2024, preprint, DOI: 10.1101/2024.05.13.593988.
101 L.-F. Nothias, M. Nothias-Esposito, R. da Silva, M. Wang, I. Protsyuk, Z. Zhang, A. Sarvepalli, P. Leyssen, D. Touboul, J. Costa, J. Paolini, T. Alexandrov, M. Litaudon and P. C. Dorrestein, *J. Nat. Prod.*, 2018, **81**, 758–767.
102 M. Ness, T. Peramuna, K. L. Wendt, J. E. Collins, J. B. King, R. Paes, N. M. Santos, C. Okeke, C. R. Miller, D. Chakrabarti, R. H. Cichewicz and L.-I. McCall, *mSystems*, 2025, **10**, e0084424.
103 K. L. Kurita, E. Glassey and R. G. Linington, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 11999–12004.
104 S. Lee, J. A. van Santen, N. Farzaneh, D. Y. Liu, C. R. Pye, T. U. H. Baumeister, W. R. Wong and R. G. Linington, *ACS Cent. Sci.*, 2022, **8**, 223–234.
105 T. Damiani, A. K. Jarmusch, A. T. Aron, D. Petras, V. V. Phelan, W. Bittremieux, D. D. Acharya, M. M. A. Ahmed, A. Bauermeister, M. J. Bertin, P. D. Boudreau, R. M. Borges, B. P. Bowen, C. J. Brown, F. O. Chagas, K. D. Clevenger, M. S. P. Correia, W. J. Crandall, M. Crüsemann, E. Fahy, O. Fiehn, N. Garg, W. H. Gerwick, J. R. Gilbert, D. Globisch, P. W. P. Gomes, S. Heuckeroth, C. A. James, S. A. Jarmusch, S. A. Kakhkhorov, K. B. Kang, N. Kessler, R. D. Kersten, H. Kim, R. D. Kirk, O. Kohlbacher, E. E. Kontou, K. Liu, I. Lizama-Chamu, G. T. Luu, T. L. Knaan, H. Mannochio-Russo, M. T. Marty, Y. Matsuzawa, A. C. McAvoy, L.-I. McCall, O. G. Mohamed, O. Nahor, H. Neuweger, T. H. J. Niedermeyer, K. Nishida, T. R. Northen, K. E. Overdahl, J. Rainer, R. Reher, E. Rodriguez, T. T. Sachsenberg, L. M. Sanchez, R. Schmid, C. Stevens, S. Subramaniam, Z. Tian, A. Tripathi, H. Tsugawa, J. J. J. van der Hooft, A. Vicini, A. Walter, T. Weber, Q. Xiong, T. Xu, T. Pluskal, P. C. Dorrestein and M. Wang, *Nat. Methods*, 2025, **22**, 1247–1254.
106 J. Galgonek and J. Vondrášek, *J. Cheminf.*, 2021, **13**, 38.
107 M. Kratochvíl, J. Vondrášek and J. Galgonek, *J. Cheminf.*, 2018, **10**, 27.
108 J. Galgonek and J. Vondrášek, *Bioinformatics*, 2024, **40**, btae174, DOI: 10.1093/bioinformatics/btae174.

109 L.-M. Quiros-Guerrero, L.-F. Nothias, A. Gaudry, L. Marcourt, P.-M. Allard, A. Rutz, B. David, E. F. Queiroz and J.-L. Wolfender, *Front. Mol. Biosci.*, 2022, **9**, 1028334.

110 M. M. Zdouc, L. M. Bayona Maldonado, H. E. Augustijn, S. Soldatou, N. de Jonge, M. Jaspars, G. P. van Wezel and J. J. J. van der Hooft, *bioRxiv*, 2025, preprint, DOI: 10.1101/2022.12.21.521422.

111 Y. Mejri, O. Cailloux, E. Otogo N'Nang, B. Séon-Méniel, J.-F. Gallard, P. Le Pogam, M. Öztürk-Escoffier and M. A. Beniddir, *Chem.:Methods*, 2025, e202400088.

112 K. Mildau, C. Büschl, J. Zanghellini and J. J. J. van der Hooft, *Bioinformatics*, 2024, **40**, btae584, DOI: 10.1093/bioinformatics/btae584.

113 K. Mildau, H. Ehlers, M. Meisenburg, E. Del Pup, R. A. Koetsier, L. R. Torres Ortega, N. F. de Jonge, K. S. Singh, D. Ferreira, K. Othibeng, F. Tugizimana, F. Huber and J. J. J. van der Hooft, *Nat. Prod. Rep.*, 2025, **42**, 982–1019.

114 V. Chandrasekhar, K. Rajan, S. R. S. Kanakam, N. Sharma, V. Weißenborn, J. Schaub and C. Steinbeck, *Nucleic Acids Res.*, 2025, **53**, D634–D643.

115 E. F. Poynton, J. A. van Santen, M. Pin, M. M. Contreras, E. McMann, J. Parra, B. Showalter, L. Zaroubi, K. R. Duncan and R. G. Linington, *Nucleic Acids Res.*, 2025, **53**, D691–D699.

116 C. Bağcı, M. Nuhamunada, H. Goyat, C. Ladanyi, L. Sehnal, K. Blin, S. A. Kautsar, A. Tagirdzhanov, A. Gurevich, S. Mantri, C. von Mering, D. Udwary, M. H. Medema, T. Weber and N. Ziemert, *Nucleic Acids Res.*, 2025, **53**, D618–D624.

117 F. C. Wolters, E. Del Pup, K. S. Singh, K. Bouwmeester, M. E. Schranz, J. J. J. van der Hooft and M. H. Medema, *Curr. Opin. Plant Biol.*, 2024, **82**, 102657.

118 J. J. J. van der Hooft, H. Mohimani, A. Bauermeister, P. C. Dorrestein, K. R. Duncan and M. H. Medema, *Chem. Soc. Rev.*, 2020, **49**, 3297–3314.

119 A. K. Jarmusch, M. Wang, C. M. Aceves, R. S. Advani, S. Aguirre, A. A. Aksenov, G. Aleti, A. T. Aron, A. Bauermeister, S. Bolleddu, A. Bouslimani, A. M. Caraballo Rodriguez, R. Chaar, R. Coras, E. O. Elijah, M. Ernst, J. M. Gauglitz, E. C. Gentry, M. Husband, S. A. Jarmusch, K. L. Jones II, Z. Kamenik, A. Le Gouellec, A. Lu, L.-I. McCall, K. L. McPhail, M. J. Meehan, A. V. Melnik, R. C. Menezes, Y. A. Montoya Giraldo, N. H. Nguyen, L. F. Nothias, M. Nothias-Esposito, M. Panitchpakdi, D. Petras, R. A. Quinn, N. Sikora, J. J. J. van der Hooft, F. Vargas, A. Vrbanac, K. C. Weldon, R. Knight, N. Bandeira and P. C. Dorrestein, *Nat. Methods*, 2020, **17**, 901–904.