Chinese Journal of Natural Medicines

Mass spectral database–based methodologies for the annotation and discovery of natural products

Fengyao Yang, Zeyuan Liang, Haoran Zhao, Jiayi Zheng, Lifang Liu, Huipeng Song, Guizhong Xin

Citation: Fengyao Yang, Zeyuan Liang, Haoran Zhao, Jiayi Zheng, Lifang Liu, Huipeng Song, Guizhong Xin, Mass spectral database–based methodologies for the annotation and discovery of natural products, Chinese Journal of Natural Medicines, 2025, 23(4), 410–420. doi: 10.1016/S1875–5364(25)60852–1.

View online: https://doi.org/10.1016/S1875–5364(25)60852–1

Chinese Journal of Natural Medicines 23 (2025) 410-420

Contents lists available at ScienceDirect

Chinese Journal of Natural Medicines

journal homepage: www.cjnmcpu.com/

Review

Mass spectral database-based methodologies for the annotation and discovery of natural products

Fengyao Yang[a], Zeyuan Liang[a], Haoran Zhao[a], Jiayi Zheng[a], Lifang Liu[a], Huipeng Song[b,*], Guizhong Xin[a,*]

[a] State Key Laboratory of Natural Medicines, Department of Chinese Medicines Analysis, School of Traditional Chinese Pharmacy, China Pharmaceutical University, Nanjing 210009, China

[b] College of Pharmacy, Liaoning University of Traditional Chinese Medicine, Dalian 116600, China

ARTICLE INFO

Article history:
Received 14 September 2024
Revised 6 November 2024
Accepted 15 November 2024
Available online 20 April 2025

Keywords:
Mass spectrometry
Natural products
Annotation
Databases

ABSTRACT

Natural products (NPs) have long held a significant position in various fields such as medicine, food, agriculture, and materials. The chemical space covered by NPs is extensive but often underexplored. Therefore, high-throughput and efficient methodologies for the annotation and discovery of NPs are desired to address the complexity and diversity of NP-based systems. Mass spectrometry (MS) has emerged as a powerful platform for the annotation and discovery of NPs. MS databases provide vital support for the structural characterization of NPs by integrating extensive mass spectral data and sample information. Additionally, the released annotation methodologies, based on a variety of informatics tools, continuously improve the ability to annotate the structure and properties of compounds. This review examines the current mainstream databases and annotation methodologies, focusing on their advantages and limitations. Prospects for future technological advancements are then discussed in terms of novel applications and research objectives. Through a systematic overview, this review aims to provide valuable insights and a reference for MS-based NPs annotation, thereby promoting the discovery of novel natural entities.

1. Introduction

Natural products (NPs), which are chemical substances naturally produced by living organisms, such as animals, plants, microorganisms, and marine organisms, exhibit diverse structures and, thus, different biological activities and physiological functions [1, 2]. These innate activities have made NPs a crucial source of inspiration for clinical drug discovery [3-6]. NPs offer better biocompatibility, structural modification potential, and safety compared to synthetic compounds [7]. Statistics indicate that more than half of the approved drugs worldwide are derived from NPs and their derivatives, which play a key role in the treatment of cancer, infectious diseases, and cardiovascular diseases [8-12]. Additionally, the discovery and characterization of NPs are essential for understanding the structural composition and gene function of organisms [13, 14]. Evidence also suggests that interactions between NPs and gut microorganisms can impact physiological states [15, 16]. However, the traditional extraction, isolation, and characterization processes are challenging, as most NPs are naturally low-content, making it difficult to extract and purify them directly [17]. Traditional methods have only explored a small portion of the NPs, and numerous compounds have been overlooked due to their interactions within complex mixtures or have been missed and degraded by extraction and separation processes [18, 19]. Therefore, there is a need to establish efficient, high-coverage, and high-sensitivity identification methodologies to discover the countless valuable NPs that remain in abundant natural resources.

Mass spectrometry (MS) is currently one of the most widely used platforms for structural characterization, offering high resolution and sensitivity. It is often coupled with chromatography to achieve robust separation, allowing the acquisition of mass and structural information from complex samples through relatively simple sample processing [1, 20]. However, the generation of large amounts of information also poses challenges for data mining, so it is critical to ensure the accuracy of the conversion of raw MS data into molecular structures. Several studies have proposed distinct levels of identification to indicate the confidence level of the annotation results [21, 22], with most discussions based on the levels established by the Metabolomics Standards Initiative (MSI) [23]. Most identifications fall within levels 2 and 3, meaning that structures and classes are inferred from physicochemical properties and spectral similarities obtained from literature or spectral databases. This highlights the importance of a thorough understanding of available databases and their proper utilization for accurate annotation. Simply using databases for molecular annotation is not enough, and manual annotation is time-consuming, laborious, and demands a high level of analytical expertise. To address this challenge, informatics tools have been developed to assist in structural analysis, thereby enabling a more efficient and extensive annotation process [24, 25].

This review examines the research progress in MS database-based methodologies for the annotation and discovery of NPs.

* Corresponding author.
E-mail addresses: songhuipeng15@163.com (H. Song); xingz@cpu.edu.cn (G. Xin)

https://doi.org/10.1016/S1875-5364(25)60852-1

First, the most commonly used MS databases in recent years are reviewed, and their usage, characteristics, and limitations are described separately to provide a reference for selecting a suitable database for identification purposes. Subsequently, MS-related annotation methodologies, which include *in-silico* spectral prediction, molecular fingerprinting, and molecular networking (MN), are also presented to compensate for the shortcomings of the databases or to extend their applications. Finally, the review concludes with a discussion of the issues facing the field and insights into the future, intending to provide readers with guidance and ideas for discovering and annotating new NPs.

2. Mass Spectral Databases for NPs Annotation

The expanding application of MS in analytical chemistry has led to the emergence of mass spectral databases over the past two decades. These databases are dispersed, with no single database encompassing all known compounds. The coverage and characteristics of the databases vary depending on the specific research objectives and application scenarios. For instance, there are databases dedicated to identifying components in traditional Chinese medicine (TCM), as well as those tailored for metabolomics or lipidomics research. In general, mass spectral databases begin with the determination data of authentic chemical standards and represent a compilation of MS features of known constituents. The core elements include exact mass, fragmentation pattern, and retention time, which serve as the three-dimensional criteria for molecular identification [26, 27]. Leveraging these attributes, it becomes feasible to swiftly generate a list of candidate compounds best suited to the target. Regrettably, several databases have ceased updates after their release, limiting their utility. This paper presents an overview of the most commonly used mass spectral databases in recent years (Table 1). The emphasis is on liquid chromatograph-mass spectrometry (LC-MS) databases, which are categorized as public, commercial, and self-built databases (Fig. 1).

2.1. LC-MS databases

2.1.1. Public databases

The Global Natural Products Social Molecular Networking (GNPS) is an ecosystem of MS that integrates various data resources and analytical tools. Notably, it is the sole public MS platform that supports MN, enabling spectrum visualization and facilitating annotation of unknown components. Furthermore, MN combines with its unique dereplication tool, allowing GNPS to perform continuous and regular automatic re-analysis, thereby establishing a virtuous cycle of iterative molecular annotation [30]. MassBank is the first public repository of shared MS data for small molecule compounds in the life sciences. Its most remarkable feature is the diversity of ionization techniques and instrument types, which can serve as filtering criteria for searching. However, MS data analyzed by different experimental methods may vary. Therefore, MassBank assists users in selecting the most relevant information for data analysis [31]. The Human Metabolome Database (HMDB) provides comprehensive reference information on nearly all known metabolites in the human body. Users can conveniently access corresponding information in other databases (KEGG, PubChem, UniProt, etc.) through hyperlinks. While HMDB also contains a portion of exogenous human substances (food, drugs, environmental exposure), this segment requires cautious consideration. In fact, compounds with spectral data cover less than 5% of HMDB. Consequently, competitive fragmentation modeling (CFM)-ID 4.0 for spectral prediction is employed to estimate the MS/MS data, though the accuracy of the predicted values should be viewed with appropriate scrutiny [32]. LIPID MAPS provides openly accessible lipid data, tools, and training resources, playing a pivotal role in the development of lipidomics [33]. A series of analytical tools form a complete workflow for lipidomics analysis. However, similar to HMDB, only a small fraction of lipids have MS information, which is displayed according to classification. Furthermore, only five lipid classes can be searched by precursor ion mass. Another drawback is that the data is presented in image form, rendering it less convenient to browse and use [34, 35]. Usually, public databases originate from diverse sources, and anyone can upload, browse, and download data freely. As a result, there is a wide range and wealth of information. However, data of varying quality are prone to inaccurate identification and require high analytical skills.

2.1.2. Commercial databases

mzCloud is a high-quality tandem mass spectrometry (MS/MS) database provided by Thermo. It is distinctive in arranging mass spectral data from different levels of MS^n into a spectral tree. Each node represents product ion spectra of the same mass-to-charge ratio (m/z) precursor ion obtained under varying conditions. The mass spectral tree enhances its searchability while providing additional information about unknown spectra. The substructure information of compounds not present in the database can be obtained byy comparing the product ion spectra of structurally related compounds through the precursor ion fingerprinting (PIF) technique. However, for NPs, this database is limited in scope [36]. Originally a freely accessible high-resolution MS/MS database, METLIN is now commercially available and requires a login for use. If identification results are not obtained by MS or MS/MS search, metabolite annotation without spectral reference can be accomplished by fragment search and neutral loss search [37]. In recent years, isoMETLIN has been developed as a novel database of isotopically labeled metabolites. It can complement the metabolites included in METLIN that lack MS/MS profiles and facilitate the identification of metabolites not present in METLIN [38, 39]. The NIST Tandem Mass Spectral Library is part of the NIST Mass Spectral Library established by the National Institute of Standards and Technology (NIST). The corresponding MS tools include MS Interpreter for linking peaks to their possible chemical structure and Hybrid Search for identifying compounds that are not in the library. However, they have to be downloaded and installed rather than being used directly from the web interface, as is the case with other databases. Wiley Registry of Tandem Mass Spectral Data--MSforID enables the sensitivity and specificity of small molecule identification. In addition, Wiley has assembled the Wiley Registry, KnowItAll Mass Spectrometry Database, NIST Mass Spectrometry Database, and other important mass spectral databases to form the world's largest high-quality MS database for LC-MS and gas chromatography-mass spectrometry (GC-MS). Overall, commercial databases are of high quality, with data from commercially available or synthetic standards assayed under tightly controlled conditions. Matched software and high compatibility make them user-friendly. However, their high price is a disadvantage.

2.1.3. Self-built databases

Both public and commercial databases serve as extensive repositories with broad coverage and rich information. However, their primary limitation is a lack of specificity for researchers with targeted objectives. In practical research, scholars often focus on specific compound groups, such as endogenous metabolites in the human body or the chemical constituents of particular TCM or medicinal plants. Directly utilizing these extensive databases can be impractical or inefficient, as it necessitates additional screening efforts. Consequently, some research teams opt to develop self-built or in-house databases tailored to their specific research needs. For example, Liu et al. constructed an in-house

Table 1 Summary of mainstream mass spectral databases

Database	Compound Source	Compound Statistics*	Spectrum Statistics*	Spectrum Type	Advantages	Disadvantages	Website
GNPS	General	29,589	591,778	LC-MS	1. The largest mass spectral database for NPs 2. Open and community-driven data sharing 3. Capable of establishing molecular networks 4. Automatic deduplication 5. Cover data from multiple types of instruments	1. Still contain unannotated spectral data 2. Mostly consist of positive ion mode data 3. Lack spectral cleaning/noise reduction	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp
Massbank	General	16,278	120,184	LC-MS, GC-MS	1. MS data from various MS settings and types 2. Detailed descriptions of collection conditions 3. All are experimental data with strong reference value 4. Free to download in different formats	1. Lack detailed descriptions of compounds	https://massbank.eu/MassBank/
HMDB	Animals	253,245	2,491,279	LC-MS, GC-MS	1. Provide detailed descriptions and properties along with external links 2. Multiple collision energies and types of instruments 3. Support various search methods 4. Downloadable	1. Predominantly predicted spectra with fewer experimental spectra 2. Contain contaminants from exogenous substances	https://hmdb.ca/
LipidMaps	Lipids	48,493	2,232	LC-MS	1. Feature the most comprehensive lipid structures 2. Allow viewing of experimental conditions 3. Provide plotting and analysis tools 4. Downloadable 5. Offer literature support	1. Data can only be displayed in image form 2. Search is limited by classification condition	https://lipidmaps.org/
mzCloud	General	32,330	16,531,567	LC-MS	1. Recalibrated experimental data 2. Multistage, multidimensional MS^n spectral data 3. Presented in the form of mass spectral trees	1. Require plugin installation to use 2. Limited number of NPs 3. Direct download is not offered 4. Collection conditions limited to Orbitrap	https://www.mzcloud.org/
ReSpect [28]	Plants	4,000	9,000	LC-MS	1. Downloadable (from the PRIMe website) 2. Feature MS/MS fragmentation association rules	1. Fragment search does not support precursor ion retrieval 2. Only Q-TOF and QqQ supported 3. Website service discontinued as of May 2024	http://spectra.psc.riken.jp/
LipidBlast	Lipids	119,341	212,685	LC-MS	1. Downloadable 2. Strict database validation for predicted spectra 3. *In-silico* spectra generated based on rules	1. Primarily developed using ion trap tandem mass spectrometry 2. Batch search for precursor ions is not allowed	https://fiehnlab.ucdavis.edu/projects/lipidblast
GMD [29]	Plants	10,546	26,590	GC-MS	1. Downloadable 2. Include mass spectral tags (MST: observed but unidentified spectra) 3. Provide various search functions including functional groups and plant parts	1. Limited to GC-MS platform 2. Updates are not timely	http://gmd.mpimp-golm.mpg.de/
METLIN	General	over 1,000,000	over 4,000,000	LC-MS	1. Broad coverage and comprehensive types 2. Support online search and both single and batch retrieval in various forms 3. isoMETLIN supplements missing spectra	1.Require payment to access and use 2.Specifically developed for Q-TOF instruments 3.MS^2 spectra are not downloadable	https://metlin.scripps.edu/landing_page.php?pgcontent=mainPage
NIST Tandem Mass Spectral Library 2023	General	51,501	2,374,064	LC-MS	1. High-resolution and high-precision MS data 2. A wide variety of precursor ion types and cleavage conditions 3. Provide accompanying analysis software and tools 4. Compatible with multiple suppliers 5. NIST/EPA/NIH EI-MS LIBRARY provides GC-MS data	1. Requires payment for access and use 2. Online search is not supported	https://www.nist.gov/programs-projects/tandem-mass-spectral-library
Wiley MSforID 2023	General	1,163	12,048	LC-MS	1. Together with other libraries forms the largest mass spectral database 2. Multiple search options and accurate search algorithms 3. Audited high-quality data 4. Broad compatibility 5. Wiley Registry provides GC-MS data	1. Require payment for access and use 2. Online search is not supported	https://msforid.com/

*Statistics are from websites of various databases as June 2024. Note that not all compounds have corresponding annotated spectral reference, such as GNPS, HMDB, LIPID MAPS.

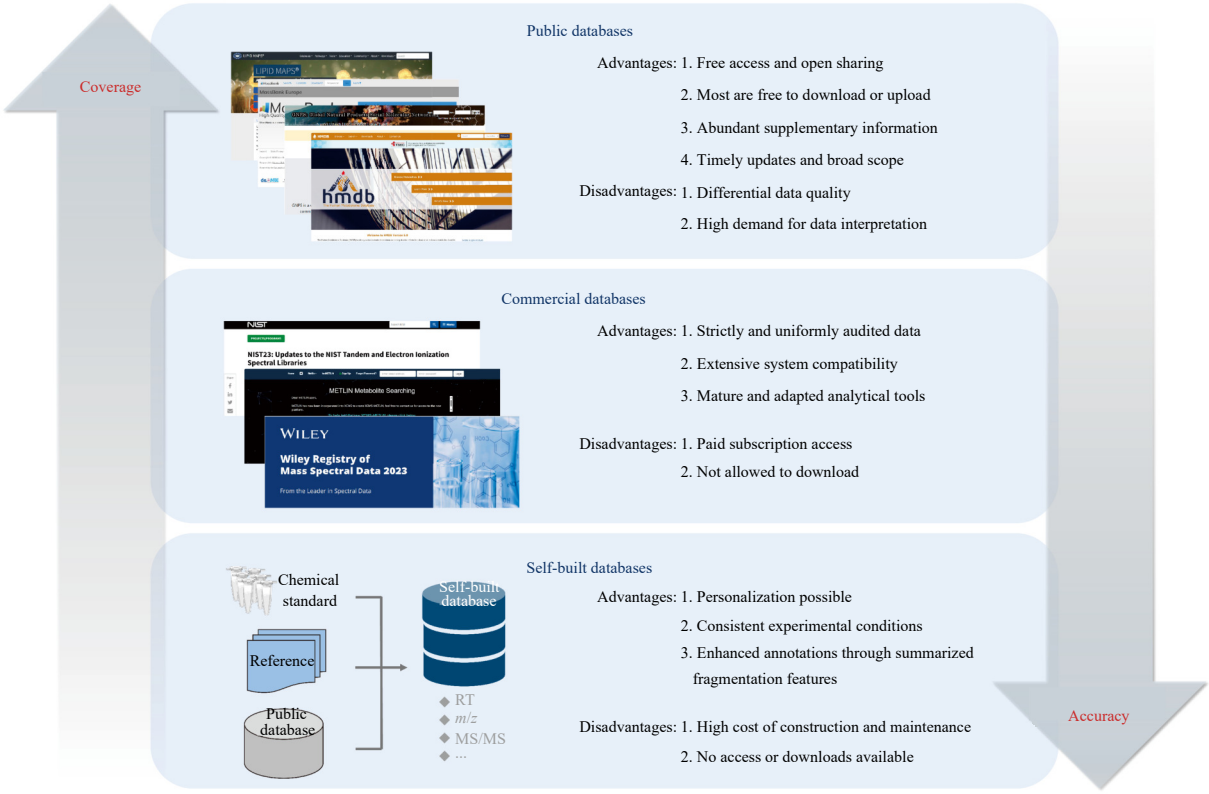

Fig. 1 Characteristics of the different categories of LC-MS databases.

mass spectral database by combining literature and open databases, aiming to elucidate the chemical composition of Chinese medicinal formulas. Focusing on Gegen-Qinlian Decoction (GQD), they recorded 741 compounds in their database and eventually identified 381 compounds from GQD using a chemical diagnostic characteristics algorithm [22]. Similarly, Huang et al. established a mass spectral database with 4,196 plant-derived saponins, providing free access and retrieval, but without the ability to browse and download the data [40]. These specialized databases have enabled researchers to narrow down their search hits and improve research efficiency. Furthermore, due to the inconsistency of experimental conditions, retention times are often difficult to use as reference information in extensive databases. Self-built databases can compensate for this deficiency and add an additional criterion for compound identification. Popov et al. provided an in-house database with retention time and spectral information for 191 types of sea cucumber triterpene saponins. Their analysis of fragmentation information revealed structure-related fragmentation patterns, and the retention times served as an important reference for analyzing chromatographic behavior [41].

In contrast to conventional database construction methods, Chen et al. proposed a novel strategy based on combinatorial principles. The researchers systematically divided steroid structures into four distinct structural components and, through permutation and combination, generated a database comprising 1,080 possible steroid structures (Fig. 2). This combinatorial approach established a robust foundation for subsequent structural characterization using neutral loss and fragment ion analysis. The key advantage of this method lies in its broad structural coverage while maintaining high specificity [42]. Similarly, Wang et al. constructed a theoretical precursor ion library (TPIL) and a characteristic fragment ion library (CFIL) for flavonoids based on free enumeration and fragmentation pattern analysis. This approach enables more precise structural annotation by facilitating the identification of specific substituents within flavonoid structures.

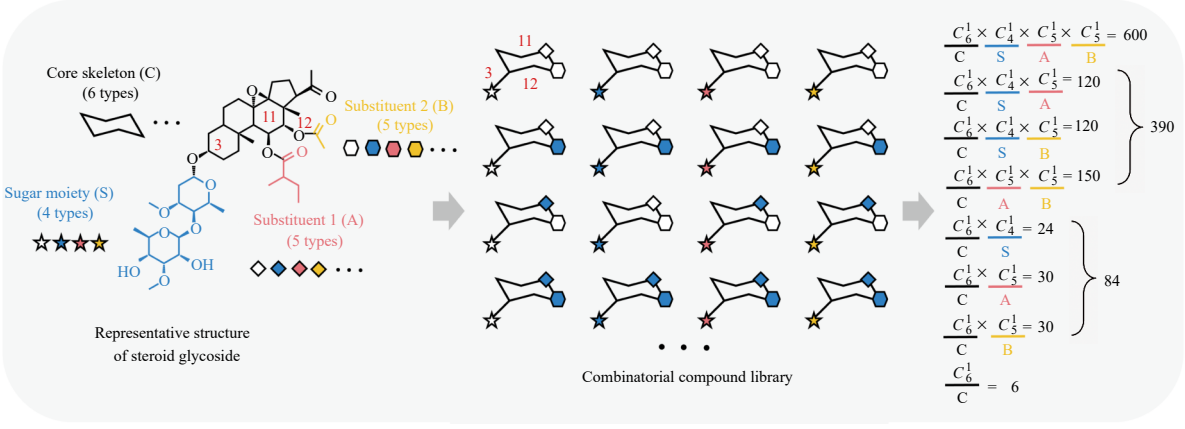

Fig. 2 Schematic illustration of the generation of a combinatorics-based compound library.

Additionally, the researchers introduced self-built software, AnnoSM, for automated annotation of substituent patterns in flavonoids and other compound classes, further enhancing structural characterization and data processing efficiency [43].

Compared to other databases, self-built databases demonstrate higher accuracy. This is due to two primary factors. Firstly, self-built databases are tailored to the specific research object or closely related species, resulting in more precise and targeted matching. Secondly, the unique data sources utilized in self-built databases enable enhanced breadth and depth within the defined scope. These data sources include information gathered from the scientific literature, as well as deductions derived from the properties of the analyzed compounds. This allows for the incorporation of elements that may not be covered by more extensive databases relying on commercially available standards. Furthermore, the experimental conditions that align with the study can improve the reliability of the data and reduce errors in annotation. However, the disadvantage of self-built databases is that they are typically non-public and require significant investment to construct and maintain.

2.2. GC-MS databases

GC-MS is distinguished from LC-MS by its maturity as an analytical technique. Ionization in GC-MS is primarily achieved through electron ionization (EI), a robust ionization method that generates numerous fragment ions with high reproducibility. This consistency across different laboratories allows for the construction of comprehensive spectral libraries, as the same compounds can reliably produce repeated data. Consequently, GC-MS benefits from the availability of relatively extensive databases, which are briefly outlined in this study.

The NIST Mass Spectral Library was initially developed as an EI-MS database. Currently, one of the prominent databases is the NIST/EPA/NIH Mass Spectral Library, which is specifically an EI mass spectral library. The latest version of this library contains 394,054 EI spectra for 347,100 compounds, and it is available for most instruments with EI capability. Additionally, the library includes a GC Retention Index (RI) Database, providing retention indices for as many compounds in the library as possible to facilitate the exclusion of results below the mass index threshold. The Wiley Registry 2023 provides 873,300 spectra for 741,000 compounds and is also compatible with most MS data systems. Furthermore, the previously mentioned MassBank and HMDB include EI data references among their offerings.

3. Methodologies for NPs Annotation

3.1. *In-silico* spectra prediction-based methodology

A significant drawback of mass spectral databases is the limited coverage of compounds and the impracticality of obtaining standards for all discovered compounds to acquire their spectra. It is estimated that only a small fraction, less than 1%, of known compounds have associated spectral information [44]. This has prompted substantial research efforts in predicting *in-silico* MS/MS spectra [45]. By utilizing compound structures as a starting point, researchers have simulated MS using various informatics tools to establish virtual databases, with the aim of bridging the gap between mass spectral databases and structural databases [46] (Fig. 3A).

Quantum chemical methods perform theoretical simulations of MS based solely on the physical and chemical behavior of molecules in a mass spectrometer to predict cleavage patterns. The quantum chemistry electron ionization mass spectra (QCEIMS) approach is the most pioneering method, which can generate calculated EI spectra of any compound by combining ab initio molecular dynamics with statistical sampling [47, 48]. The primary advantage of this method is that spectral prediction of underrepresented compounds performs well without range limitations, as it does not require dependence on databases or training sets. Products derived from trimethylsilylation (TMS), which is widely used for GC-MS, can also be predicted for fragmentation by using QCEIMS, and this approach has demonstrated good predictive capabilities [49]. However, studies have shown that prediction accuracy varies across different structural classes, with notably lower accuracy for organic oxygenated compounds compared to other molecular types [50]. A major limitation of QCEIMS lies in its high computational demand, making the prediction of large molecules or multiple compounds highly time-consuming. Additionally, the application of quantum chemistry to predict electrospray ionization (ESI) spectra remains a significant challenge.

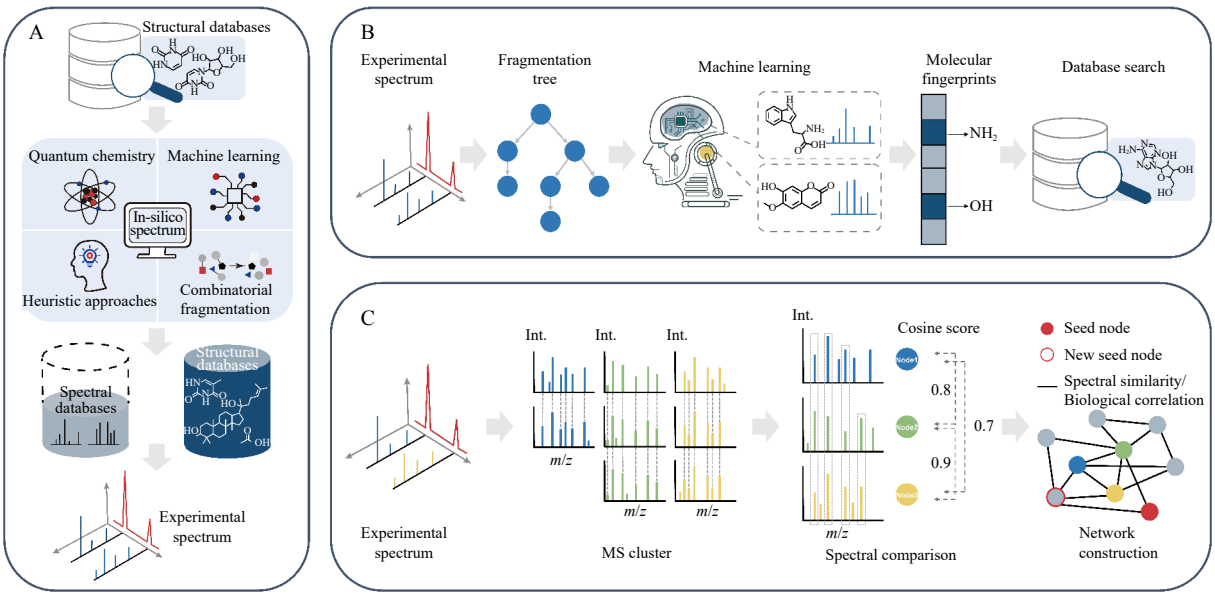

Fig. 3 Three methodologies for structural annotation of NPs. (A) Schematic illustration of *in-silico* spectra prediction-based methodology. (B) Schematic illustration of molecular fingerprint-based methodology. (C) Schematic illustration of MN-based methodology.

Machine learning-based approaches offer powerful computational models that can address the issue of slow computation. The CFM approach is a probabilistic generative model that learns fragmentation patterns and optimizes its parameters based on acquired data [51]. It demonstrates strong applicability for both ESI and EI spectral predictions. The latest version, CFM-ID 4.0, enhances predictive accuracy by learning parameters from the topological features of chemical structures represented by tensors. This advancement allows the model to generate more precise spectral predictions across a broader range of compounds [52, 53]. Beyond spectral prediction, CFM-ID 4.0 also ranks potential structures that correspond to a given MS/MS spectrum, making it one of the most effective tools for spectral annotation. Building on this foundation, Wang et al. fine-tuned the model parameters and implemented a deep learning framework using transfer learning to improve spectral predictions for both known and novel compounds. Their results demonstrated the potential of this approach to identify compounds without reference spectra and even recognize entirely new chemical entities [54].

Lipid identification benefits from the application of heuristic methodologies due to the high degree of structural similarity among recurring fragments. LipidBlast, a prominent tool in lipidomics research, utilizes this approach by inputting lipid structures from pathways such as LIPID MAPS and generating rule-based tandem mass spectra based on fragmentation analysis for each class of structures [55]. This has facilitated the updating and refinement of lipid databases [56]. However, this approach faces limitations in extending to molecules with dissimilar structures. In contrast, MetFrag employs a combinatorial fragmentation method that enumerates all possible fragments of candidate structures based on chemical bond breaking, and then compares and assigns peaks [57, 58]. Statistical methods have also been integrated into MetFrag 2.4.5 to optimize the scoring function for improved annotation [59]. Additionally, MS-FINDER predicts fragmentations and ranks candidate structures based on hydrogen rearrangement rules in bond cleavages at low-energy collision-induced dissociation (CID), achieving an interpretation rate of nearly 80% for fragmentation ions and successfully identifying two new compounds without standard spectra [60].

3.2. Molecular fingerprint-based methodology

Molecular fingerprinting refers to the identification of the chemical structural features or substructures (e.g., benzene rings, hydroxyl groups) of a compound, as revealed by its spectral data. In contrast to *in-silico* prediction, the molecular fingerprint-based methodology starts directly from MS/MS spectra. Initially, a series of chemical structure attributes, such as molecular fingerprints and molecular formulas, are extracted from the raw data. These features are then queried and compared against structure databases to obtain candidate structures for scoring and ranking (Fig. 3B). FingerID, an early approach in this field, is a method for predicting molecular characteristic fingerprints from tandem mass spectra using a support vector machine (SVM) algorithm. The fingerprint prediction model is trained on public databases, making this machine learning-based method compatible with spectra from various mass spectrometer types [61].

The introduction of the fragmentation tree (FT) data structure has led to the development of several new methodologies. The FT represents molecular fragmentation pathways based on the premise that all fragments originate from a precursor. The nodes in the FT correspond to possible molecular formulas of the compounds or fragments, while the edges represent the hypothesized fragmentation reactions and losses [62-64]. This approach provides a framework for *de novo* identification of compounds. For instance, CSI: FingerID converts spectra into fragmentation trees, which are then combined with other similarity metrics for multiple kernel learning to improve the performance of predicting molecular fingerprints [65]. Another technique, class assignment and ontology prediction using mass spectrometry (CANOPUS), integrates two machine learning approaches. It uses probabilistic fingerprints predicted by SVM as input to a deep neural network (DNN), which is trained on a large number of molecular structures to predict compound classes based on molecular fingerprints. Remarkably, CANOPUS can still achieve reliable predictions even for compound classes with limited or no spectra available for training [66].

SIRIUS 4.0 has integrated CSI: FingerID and CANOPUS. Initially, SIRIUS was developed as a tool for analyzing molecular formulae using isotopic patterns, including isotopic distribution and mean peak masses [67]. The FT analysis function was subsequently added. SIRIUS first considers all possible molecular formulae based on an integrated isotope pattern scoring model, without discarding atypical ones. It then computes the FT using a Maximum A Posteriori estimation, selecting the molecular formula that best explains the data [68]. Finally, CSI: FingerID searches the structural database or provides structural information for candidate scoring based on a Bayesian network. CSI: FingerID has been technically upgraded to quickly provide data-driven structural annotations for all detected MS features in the complete dataset [69]. Hoffmann et al. proposed a workflow called COSMIC that incorporates SIRIUS to assign confidence scores for annotation results, aiming to distinguish between correct and incorrect hits (Fig. 4). The application of COSMIC to various datasets has demonstrated that it not only outperforms spectral library matching in dereplication but also allows high-confidence annotation of unreported substances. Specifically, COSMIC was used to discover 12 additional amino acid conjugations of bile acids, all of which were validated by manual evaluation or synthetic standards [70]. MSNovelist is a workflow that combines SIRIUS with recurrent neural networks (RNN). Molecular formulae and molecular fingerprints predicted by SIRIUS and CSI: FingerID are fed into the RNN model to generate SMILES sequence output, thereby establishing a continuous link from spectra to fingerprints to structures. This process of generating candidate structures directly from spectra is not limited by structural databases. MSNovelist was applied to a bryophyte dataset, leading to the inference of seven novel compound structures [71]. These *de novo* annotation methodologies can serve as a valuable alternative for the discovery of new biological small molecules.

3.3. MN-based methodology

MN is a network-based approach that reveals potential associations between compounds within complex samples by visualizing tandem mass spectra. This approach is founded on the premise that structurally similar compounds exhibit comparable mass spectral features. In MN, the mass spectral features of each compound are represented as nodes, and the spectral similarities between compounds are depicted as edges, collectively forming a network structure. Subsequent data processing involves collapsing and grouping the spectra using clustering algorithms or other network analysis techniques. This enables the propagation of annotations based on the mass spectral database. Consequently, even if a node does not directly match the spectral database, it can be recognized through the reference spectra from neighboring nodes [30, 72-74] (Fig. 3C).

MN within the GNPS platform is a prominent example. In MN, spectral similarity is measured by the cosine score, which accounts for differences in fragment ion intensities and precursor ion m/z between spectral pairs. Higher cosine values, with 1 indicating identical spectra, typically have a threshold of 0.5–0.7 [75]. However, the MS-Cluster approach solely considers fragmentation patterns during clustering, disregarding the possibility of

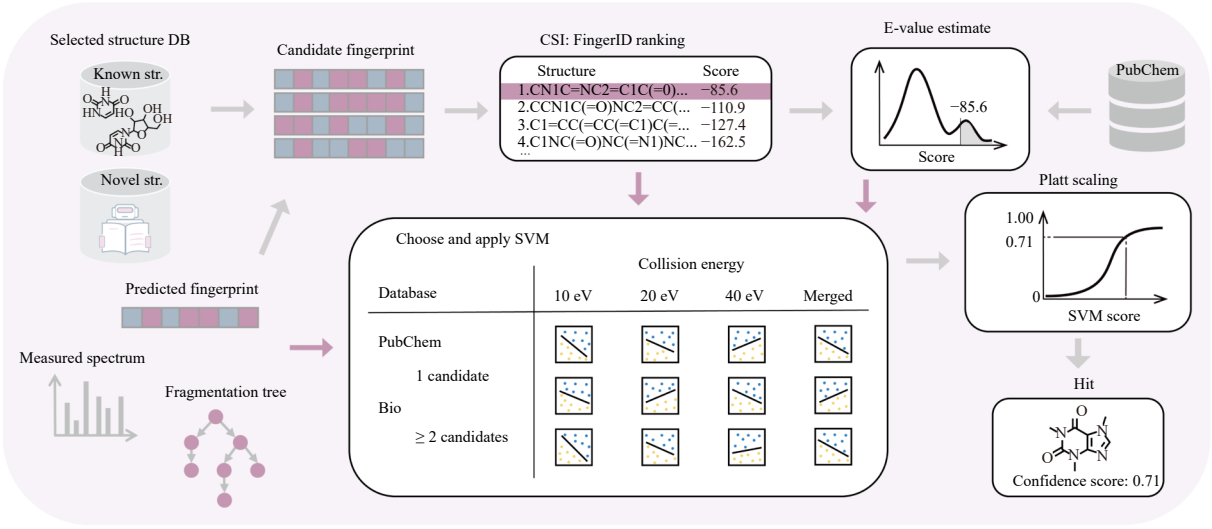

Fig. 4 The workflow of COSMIC: After selecting the structural database, the raw data is processed using SIRIUS. Only candidates with the highest CSI: FingerID rankings are considered for further analysis. To assign confidence scores, spectral features, E-value estimation, multiple SVMs, and Platt scaling are integrated into the evaluation process. DB, database; str, structure.

isomers [76]. Feature-based molecular networking (FBMN) incorporates more comprehensive information, including isotope pattern, retention time, and corresponding abundance, to perform data reduction based on feature lookup. This improved tool not only enables the discrimination of isomers based on retention time but also generates quantitative information through feature detection and incorporates it into the MN [77]. The subsequent development of ion identity molecular networking (IIMN) provides further complementarity by identifying and collapsing multiple ionic adduct types generated by ionization of the same molecule, thereby reducing redundant unconnected nodes and facilitating the interpretation of MN [78].

In addition to classic MN, other methodologies of network analysis exist. In 2008, Hao et al. proposed the family network approach, which first categorized component families based on searched diagnostic ions and then connected a coherent network according to bridging components, which are fragment ions common in different families. This approach enabled the identification of individual components and elucidation of the substructures represented by all diagnostic ions, thus efficiently narrowing down the database hits [79]. Wang et al. developed a structure-guided molecular network strategy (SGMNS) to construct a global connectivity molecular network (GCMN) through the similarity of molecular fingerprints of chemical structures in metabolomic databases, with a recursive algorithm used to propagate the network annotation [80]. The structural similarity network annotation platform for mass spectrometry (SNAP-MS) made full use of MS^1 data, screening and categorizing compounds based on the molecular formula of the core skeleton and then matching chemical similarity groupings with MS feature groupings in MN. An evaluation of the method's performance using complex mixtures and pure compound libraries showed that SNAP-MS can provide correct annotation of compound families for subnetworks, serving as a valuable complement to CANOPUS [81].

Given the rapid development of network analysis, there are a wide range of successful application cases in areas such as TCM research and metabolomics [82-84]. For example, Li et al. achieved targeted isolation and structural identification of new NPs with low abundance but high activity under the guidance of MN [85]. Li et al. analyzed NPs in herbal medicines through FBMN and dual ionization mode, which enabled high-precision identification and isomer differentiation [86]. In addition to using MN for understanding the material basis of TCM, Chen et al. established a semi-quantitative feature-based molecular networking (SQFBMN) approach to explore the correlation of chemical constituents in *Cannabis sativa* leaves with their origin and bioactivities [87]. Zhang et al. detected the prototypes and metabolites of Chinese herbal formulas by MN based on optimized data acquisition [88]. Quinn et al. revealed the chemical effects of the microbiome on different organs through MN-supported meta-mass-shift chemical profiling and reported microbe-mediated bile-acid conjugations for the first time [89].

3.4. Combined methodology

The various annotation strategies are not entirely distinct from one another. Some studies have ingeniously combined different concepts and techniques to produce integrated methodologies for compound annotation. Li et al. integrated quantum chemical calculations with hydrogen-deuterium exchange to elucidate the CID mechanism of lindenane sesquiterpenoids (LSs), with the goal of developing an automated target feature extraction program. This approach facilitated the discovery of 96 LSs, of which 37 were potentially novel compounds [90]. Liu et al. proposed an enhanced MN-based approach called DFMN-ISD, which combines diagnostic fragment, MN, and FT strategies. This method was successfully applied to the annotation of Fritillaria steroidal alkaloids and is expected to be extended to characterize the composition of other herbal medicines [91].

The introduction of biological knowledge has facilitated the development of novel annotation methodologies. These approaches leverage beyond traditional MS phenomena, such as adducts, isotopes, and fragments, to incorporate peak-to-peak relationships reflecting biotransformation. For instance, NetID employs a global optimization strategy *via* integer linear programming to fully utilize available information [92]. Metabolite annotation and dysregulated network analysis (MetDNA) defines reaction pairs connecting metabolic substrates and products, forming a metabolic reaction network (MRN). This annotation is then recursively expanded by treating annotated neighbor metabolites as new seeds [93]. However, these single-network methodologies can be enhanced by integrating multiple layers of information. The knowledge-guided multi-layer network (KGMN/MetDNA2) approach combines three distinct networks: one constructed by spectral library searching and enzymatic reactions, another incorporating metabolic biotransformation constraints, and a third recognizing different ionic forms to improve peak identification accuracy (Fig. 5). This integrated approach significantly increased the identification accuracy of known metabolites to over 95% while facilitating the efficient annotation of unknown com-

pounds [94]. These advancements in biological knowledge-driven annotation methodologies serve multiple purposes. First, they can facilitate the selection and isolation of bioactive substances. The NP^3 MS Workflow computes a bioactivity correlation coefficient (BCC) to rank consensus spectra that may represent bioactive NPs directly from complex mixtures [95]. Similarly, the bioactivity-based MN proposed by Nothias et al. maps bioactivity scores in MN to expose potentially bioactive NPs from fractionated extracts [96]. Second, these methodologies can be utilized for biomarker identification. For instance, Gong et al. depicted metabolic networks of complex systems based on chemicalome-to-metabolome matching, providing possible associations between prototypes and metabolites [97]. Additionally, Wang et al. developed a metabolic pathway extension (MPE) method that enables mapping from initial metabolites to the metabolome through summarized metabolic reactions [98]. Both approaches focus on matching molecular mass differences to metabolic reactions.

4. Concluding Remarks and Future Perspectives

In summary, the cohesive integration of informatics tools and accumulated experience has led to the development of well-established mass spectral databases and associated algorithmic methodologies over the past decade. While the breadth and richness of the database content have increased and are continuously being updated, the storage and accumulation of databases remain a common issue, as the content is still not entirely satisfactory. The future direction should focus on the storage of diversified data in a standardized form for efficient retrieval [99]. However, existing statistics indicate a low overlap between different databases [100], and combining them could enhance complementarity and alleviate the problem to some extent. There is a long-standing consensus on the need for open sharing of databases [25, 101], which has the potential to facilitate the mining of unknown components through re-analysis [102]. In this aspect, GNPS excels, with MassIVE, a public data repository containing sample information (metadata) and spectrum information [76]. The launched metadata capture system ReDU contributes to the curation and utilization of metadata [103]. The existence of metadata has spawned a series of MS analysis tools, such as reference data-driven (RDD) analysis, which enhances the understanding of complex samples based on metadata-annotated source data [104], and MicrobeMASST, a taxonomic search tool that associates microbe-derived metabolites with relative producers [105]. Furthermore, Mohanty et al. employed MassQL to mine and filter public data in MassIVE, creating a mass spectral database of modified bile acids and discovering previously unknown bile acids, thus enriching the diversity of bile acid modifications [106]. As can be seen, metadata can endow data with vitality and significance, offering enormous potential for data sharing and full utilization. The MS databases serve not only as a reference for the annotation of known compounds but also provide a foundation for the presumption of unknown compounds.

Advances in algorithmic methodologies have enabled structure elucidation to surpass the limitations of databases and even extend to unknown compounds. There is no single perfect method, as each approach possesses distinct characteristics. Researchers can maximize the value of these methods by selecting the most appropriate technique for their research objectives. *In silico* spectra prediction can be a valuable complement to MS databases. Molecular fingerprinting-based and MN-based methodologies may hypothesize unknown compounds, with the former requiring structural databases and the latter mass spectral databases. Additionally, the former annotates individually, while the latter utilizes the relationships between spectra. Approaches that incorporate biological knowledge offer further insights for structural annotation and facilitate correlations with biological activity. Beyond the algorithms, the overall research workflows are also crucial. Reverse metabolomics employs a reverse route, synthesizing compounds first and then mining metadata to reveal associated phenotypes, species, and sample types. This can be considered a universal approach for discovering other molecules from biological systems [107, 108]. Nie et al. combined click-chemistry-based enrichment with MN to identify several secondary bile

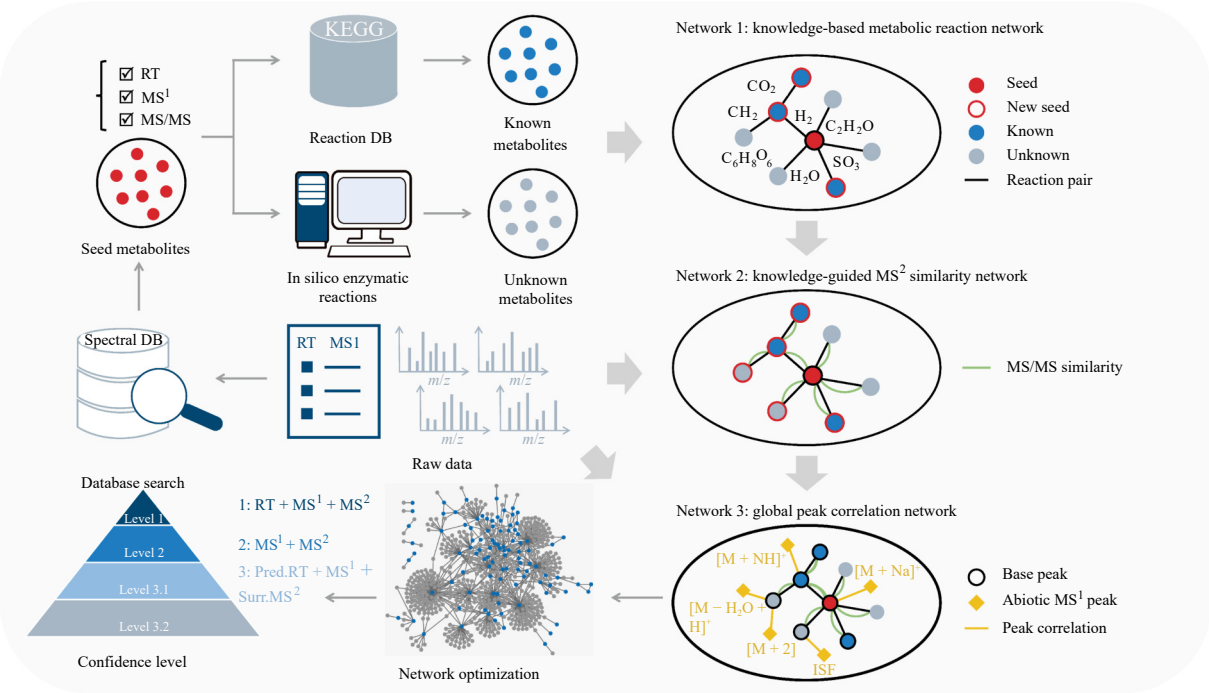

Fig. 5 The workflow of KGMN (MetDNA2): It begins with the identification of seed metabolites by matching experimental data against a mass spectral database. In Network 1, metabolites are connected based on known metabolic reactions. Network 2 is then constructed by integrating biological relationships and spectral similarities to refine annotations. Annotated metabolites serve as new seeds for iterative propagation. Network 3 further annotates different ion forms. Finally, network optimization and filtering are applied to generate metabolite annotations at varying confidence levels.

acids, including the previously uncharacterized 3-succinylated cholic acid [109]. FNICM is a research framework that uses targeted metabolomics and biological knowledge as seeds to propagate perturbations to undetected metabolites, providing potential functional metabolites and a basis for subsequent annotation [110]. These studies offer unique inspiration for NP analysis. While some algorithms and strategies were initially developed for specific applications, such as metabolomics, applying them to different scenarios may yield unexpected successes.

Comprehensive annotation with high accuracy has always been the primary objective in NPs analysis. This process cannot be accomplished solely through data analysis steps, but rather requires a holistic focus on the entire workflow. The initial stages of sample preparation and data acquisition are crucial in determining the coverage of the assay. NPs exhibit diverse polarity and solubility, necessitating careful extraction considerations to ensure target analytes are captured while eliminating interferences [111, 112]. The incorporation of MS probes has emerged as a popular trend, as it can extend the detection range by enhancing sensitivity or reducing polarity [113-115]. However, a lack of matching high-throughput data processing techniques remains a challenge. In contrast to typical chemical probes, Zhao et al. introduced an innovative enzyme probe approach to localize and distinguish isomers of 12α-hydroxylated bile acids using the enzyme conversion rate as an index [116]. Additionally, acquisition parameters and conditions, such as mobile phase, ionization mode, collision energy, and acquisition speed, significantly impact the quantity and quality of the obtained spectra [117-119]. For instance, the recently developed electron impact excitation of ions from organics (EIEIO) fragmentation technique can enrich ion fragmentation and enhance annotation [120]. Zhang et al. combined the comprehensive coverage of full-scan and the high-quality data-dependent acquisition (DDA) to establish a new data acquisition method, dpDDA, which can improve stability and coverage [121]. The setting of search conditions in the spectral library search process is crucial, as it determines the pool of potential candidates. Appropriate configuration of search conditions can significantly increase the likelihood of isolating correct results. Various data filtering strategies have been developed to remove interfering signals and extract mass spectral features for rapid identification [122, 123]. These strategies include diagnostic ion filtration (DIF) [124-126], neutral loss filtration (NLF) [127], mass deficit filtration (MDF) [128-132], and in-source fragment elimination [133]. It is essential to recognize that annotation should not be considered the ultimate goal, but rather the groundwork for functional annotation and the exploration of the activity and efficacy of NPs. A global perspective-based, multi-segment strategy may offer promising solutions for the future of NPs annotation and discovery.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 82274064, 82374026, and 82204591).

Declaration of Competing Interest

These authors have no conflict of interest to declare.

References

1 Dong SH, Duan ZK, Bai M, et al. Advanced technologies targeting isolation and characterization of natural products. *Trac-Trends Anal Chem*. 2024;175:117711. https://doi.org/10.1016/j.trac.2024.117711.
2 Katz L, Baltz RH. Natural product discovery: past, present, and future. *J Ind Microbiol Biotechnol*. 2016;43(2–3):155-176. https://doi.org/10.1007/s10295-015-1723-5.
3 Butler MS, Robertson AA, Cooper MA. Natural product and natural product derived drugs in clinical trials. *Nat Prod Rep*. 2014;31(11):1612-1661. https://doi.org/10.1039/C4NP00064A.
4 Atanasov AG, Zotchev SB, Dirsch VM, et al. Natural products in drug discovery: advances and opportunities. *Nat Rev Drug Discov*. 2021;20(3):200-216. https://doi.org/10.1038/s41573-020-00114-z.
5 Thomford NE, Senthebane DA, Rowe A, et al. Natural products for drug discovery in the 21st century: innovations for novel drug discovery. *Int J Mol Sci*. 2018;19(6):1578. https://doi.org/10.3390/ijms19061578.
6 Deng WY, Chen F, Zhao Y, et al. Anti-hepatitis B virus activities of natural products and their antiviral mechanisms. *Chin J Nat Med*. 2023;21(11):803-811. https://doi.org/10.1016/S1875-5364(23)60505-9.
7 Luo ZW, Yin FC, Wang XB, et al. Progress in approved drugs from natural product resources. *Chin J Nat Med*. 2024;22(3):195-211. https://doi.org/10.1016/S1875-5364(24)60582-0.
8 Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod*. 2020;83(3):770-803. https://doi.org/10.1021/acs.jnatprod.9b01285.
9 Zhang BY, Zheng YF, Zhao J, et al. Identification of multi-target anti-cancer agents from TCM formula by *in silico* prediction and *in vitro* validation. *Chin J Nat Med*. 2022;20(5):332-351. https://doi.org/10.1016/S1875-5364(22)60180-8.
10 Li S, Fan GF, Li XJY, et al. Modulation of type I interferon signaling by natural products in the treatment of immune-related diseases. *Chin J Nat Med*. 2023;21(1):3-18. https://doi.org/10.1016/S1875-5364(23)60381-4.
11 Chen JX, Ding ZQ. Natural products as potential drug treatments for acute promyelocytic leukemia. *Chin Med*. 2024;19(1):57. https://doi.org/10.1186/s13020-024-00928-8.
12 Chen JX, Ding ZQ. Advances in natural product anti-coronavirus research (2002–2022). *Chin Med*. 2023;18(1):13. https://doi.org/10.1186/s13020-023-00715-x.
13 Pye CR, Bertin MJ, Lokey RS, et al. Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci USA*. 2017;114(22):5601-5606. https://doi.org/10.1073/pnas.1614680114.
14 Chen G, Zhou D, Wang CM, et al. Advances in the role of natural products in human gene expression. *Chin J Nat Med*. 2022;20(1):1-8. https://doi.org/10.1016/S1875-5364(22)60147-X.
15 Zhang NN, Jiang ZM, Li SZ, et al. Evolving interplay between natural products and gut microbiota. *Eur J Pharmacol*. 2023;949:175557. https://doi.org/10.1016/j.ejphar.2023.175557.
16 Bauermeister A, Mannochio-Russo H, Costa-Lotufo LV, et al. Mass spectrometry-based metabolomics in microbiome investigations. *Nat Rev Microbiol*. 2022;20(3):143-160. https://doi.org/10.1038/s41579-021-00621-9.
17 Zhang QW, Lin LG, Ye WC. Techniques for extraction and isolation of natural products: a comprehensive review. *Chin Med*. 2018;13:20. https://doi.org/10.1186/s13020-018-0177-x.
18 Salem MA, Perez de Souza L, Serag A, et al. Metabolomics in the context of plant natural products research: from sample preparation to metabolite analysis. *Metabolites*. 2020;10(1):37. https://doi.org/10.3390/metabo10010037.
19 Atanasov AG, Waltenberger B, Pferschy-Wenzig EM, et al. Discovery and resupply of pharmacologically active plant-derived natural products: a review. *Biotechnol Adv*. 2015;33(8):1582-1614. https://doi.org/10.1016/j.biotechadv.2015.08.001.
20 Allard PM, Genta-Jouve G, Wolfender JL. Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification. *Curr Opin Chem Biol*. 2017;36:40-49. https://doi.org/10.1016/j.cbpa.2016.12.022.
21 Schymanski EL, Jeon J, Gulde R, et al. Identifying small molecules *via* high resolution mass spectrometry: communicating confidence. *Environ Sci Technol*. 2014;48(4):2097-2098. https://doi.org/10.1021/es5002105.
22 Liu KX, Li N, Yin YH, et al. An in-house database-driven untargeted identification strategy for deep profiling of chemicalome in Chinese medicinal formula. *J Chromatogr A*. 2022;1666:462862. https://doi.org/10.1016/j.chroma.2022.462862.
23 Sumner LW, Amberg A, Barrett D, et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*. 2007;3(3):211-221. https://doi.org/10.1007/s11306-007-0082-2.
24 Tian ZT, Liu FZ, Li DQ, et al. Strategies for structure elucidation of small molecules based on LC-MS/MS data from complex biological samples. *Comput Struct Biotechnol J*. 2022;20:5085-5097. https://doi.org/10.1016/j.csbj.2022.09.004.
25 Tsugawa H. Advances in computational metabolomics and databases deepen the understanding of metabolisms. *Curr Opin Biotechnol*. 2018;54:10-17. https://doi.org/10.1016/j.copbio.2018.01.008.
26 Kind T, Tsugawa H, Cajka T, et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom Rev*. 2018;37(4):513-532. https://doi.org/10.1002/mas.21535.
27 Chaleckis R, Meister I, Zhang P, et al. Challenges, progress and promises of metabolite annotation for LC-MS-based metabolomics. *Curr Opin Biotechnol*. 2019;55:44-50. https://doi.org/10.1016/j.copbio.2018.07.010.
28 Sawada Y, Nakabayashi R, Yamada Y, et al. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry*. 2012;82:38-45. https://doi.org/10.1016/j.phytochem.2012.07.007.
29 Kopka J, Schauer N, Krueger S, et al. GMD@CSB. DB: the Golm Metabolome Database. *Bioinformatics*. 2005;21(8):1635-1638. https://doi.org/10.1093/bioinformatics/bti236.
30 Wang M, Carver JJ, Phelan VV, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol*. 2016;34(8):828-837. https://doi.org/10.1038/nbt.3597.
31 Horai H, Arita M, Kanaya S, et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom*. 2010;45(7):703-714.

https://doi.org/10.1002/jms.1777.
32 Wishart DS, Guo A, Oler E, et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res*. 2022;50(D1):D622-D631. https://doi.org/10.1093/nar/gkab1062.
33 Sud M, Fahy E, Cotter D, et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res*. 2007;35(SI1):D527-D532. https://doi.org/10.1093/nar/gkl838.
34 O'Donnell VB, Dennis EA, Wakelam MJO, et al. LIPID MAPS: serving the next generation of lipid researchers with tools, resources, data, and training. *Sci Signal*. 2019;12(563):eaaw2964. https://doi.org/10.1126/scisignal.aaw2964.
35 Ni Z, Wolk M, Jukes G, et al. Guiding the choice of informatics software and tools for lipidomics research applications. *Nat Methods*. 2023;20(2):193-204. https://doi.org/10.1038/s41592-022-01710-0.
36 Sheldon MT, Mistrik R, Croley TR. Determination of ion structures in structurally related compounds using precursor ion fingerprinting. *J Am Soc Mass Spectrom*. 2009;20(3):370-376. https://doi.org/10.1016/j.jasms.2008.10.017.
37 Xue J, Guijas C, Benton HP, et al. METLIN MS2 molecular standards database: a broad chemical and biological resource. *Nat Methods*. 2020;17(10):953-954. https://doi.org/10.1038/s41592-020-0942-5.
38 Guijas C, Montenegro-Burke JR, Domingo-Almenara X, et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal Chem*. 2018;90(5):3156-3164. https://doi.org/10.1021/acs.analchem.7b04424.
39 Cho K, Mahieu N, Ivanisevic J, et al. isoMETLIN: a database for isotope-based metabolomics. *Anal Chem*. 2014;86(19):9358-9361. https://doi.org/10.1021/ac5029177.
40 Huang FQ, Dong X, Yin X, et al. A mass spectrometry database for identification of saponins in plants. *J Chromatogr A*. 2020;1625:461296. https://doi.org/10.1016/j.chroma.2020.461296.
41 Popov RS, Ivanchina NV, Silchenko AS, et al. A mass spectrometry database for sea cucumber triterpene glycosides. *Metabolites*. 2023;13(7):783. https://doi.org/10.3390/metabo13070783.
42 Chen YH, Li SY, Wang D, et al. Combinatorics-based chemical characterization and bioactivity comparison of different parts of traditional Chinese medicinal plants through LC-Q-TOF-MS/MS, multivariate statistical analysis and bioassay: *Marsdenia tenacissima* as an example. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2023;1228:123850. https://doi.org/10.1016/j.jchromb.2023.123850.
43 Wang X, Guo AQ, Wang R, et al. AnnoSM: an automated annotation tool for determining the substituent modes on the parent skeleton based on a characteristic MS/MS fragment ion library. *Anal Chem*. 2024;96(9):3817-3828. https://doi.org/10.1021/acs.analchem.3c04946.
44 Krettler CA, Thallinger GG. A map of mass spectrometry-based *in silico* fragmentation prediction and compound identification in metabolomics. *Brief Bioinform*. 2021;22(6):bbab073. https://doi.org/10.1093/bib/bbab073.
45 Hufsky F, Bocker S. Mining molecular structure databases: identification of small molecules based on fragmentation mass spectrometry data. *Mass Spectrom Rev*. 2017;36(5):624-633. https://doi.org/10.1002/mas.21489.
46 Bocker S. Searching molecular structure databases using tandem MS data: are we there yet. *Curr Opin Chem Biol*. 2017;36:1-6. https://doi.org/10.1016/j.cbpa.2016.12.010.
47 Grimme S. Towards first principles calculation of electron impact mass spectra of molecules. *Angew Chem Int Ed Engl*. 2013;52(24):6306-6312. https://doi.org/10.1002/anie.201300158.
48 Bauer CA, Grimme S. How to compute electron ionization mass spectra from first principles. *J Phys Chem A*. 2016;120(21):3755-3766. https://doi.org/10.1021/acs.jpca.6b02907.
49 Wang S, Kind T, Bremer PL, et al. Quantum chemical prediction of electron ionization mass spectra of trimethylsilylated metabolites. *Anal Chem*. 2022;94(3):1559-1566. https://doi.org/10.1021/acs.analchem.1c02838.
50 Wang S, Kind T, Tantillo DJ, et al. Predicting *in silico* electron ionization mass spectra using quantum chemistry. *J Cheminform*. 2020;12(1):63. https://doi.org/10.1186/s13321-020-00470-3.
51 Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*. 2014;11(1):98-110. https://doi.org/10.1007/s11306-014-0676-4.
52 Hu GL, Qiu MH. Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Nat Prod Rep*. 2023;40(11):1735-1753. https://doi.org/10.1039/D3NP00025G.
53 Wang F, Liigand J, Tian SY, et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem*. 2021;93(34):11692-11700. https://doi.org/10.1021/acs.analchem.1c01465.
54 Wang F, Pasin D, Skinnider MA, et al. Deep learning-enabled MS/MS spectrum prediction facilitates automated identification of novel psychoactive substances. *Anal Chem*. 2023;95(50):18326-18334. https://doi.org/10.1021/acs.analchem.3c02413.
55 Kind T, Liu KH, Lee DY, et al. LipidBlast *in silico* tandem mass spectrometry database for lipid identification. *Nat Methods*. 2013;10(8):755-758. https://doi.org/10.1038/nmeth.2551.
56 Kind T, Okazaki Y, Saito K, et al. LipidBlast templates as flexible tools for creating new *in-silico* tandem mass spectral libraries. *Anal Chem*. 2014;86(22):11024-11027. https://doi.org/10.1021/ac502511a.
57 Wolf S, Schmidt S, Müller-Hannemann M, et al. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*. 2010;11:148. https://doi.org/10.1186/1471-2105-11-148.
58 Ruttkies C, Schymanski EL, Wolf S, et al. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. *J Cheminform*. 2016;8:3. https://doi.org/10.1186/s13321-016-0115-9.
59 Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics*. 2019;20(1):376. https://doi.org/10.1186/s12859-019-2954-7.
60 Tsugawa H, Kind T, Nakabayashi R, et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal Chem*. 2016;88(16):7946-7958. https://doi.org/10.1021/acs.analchem.6b00770.
61 Heinonen M, Shen H, Zamboni N, et al. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics*. 2012;28(18):2333-2341. https://doi.org/10.1093/bioinformatics/bts437.
62 Vaniya A, Fiehn O. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends Analyt Chem*. 2015;69:52-61. https://doi.org/10.1016/j.trac.2015.04.002.
63 Hufsky F, Scheubert K, Böcker S. Computational mass spectrometry for small-molecule fragmentation. *Trac-Trends Anal Chem*. 2014;53:41-48. https://doi.org/10.1016/j.trac.2013.09.008.
64 Rasche F, Svatos A, Maddula RK, et al. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem*. 2011;83(4):1243-1251. https://doi.org/10.1021/ac101825k.
65 Duhrkop K, Shen H, Meusel M, et al. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proc Natl Acad Sci USA*. 2015;112(41):12580-12585. https://doi.org/10.1073/pnas.1509788112.
66 Duhrkop K, Nothias LF, Fleischauer M, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol*. 2021;39(4):462-471. https://doi.org/10.1038/s41587-020-0740-8.
67 Bocker S, Letzel MC, Liptak Z, et al. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics*. 2009;25(2):218-224. https://doi.org/10.1093/bioinformatics/btn603.
68 Bocker S, Duhrkop K. Fragmentation trees reloaded. *J Cheminform*. 2016;8:5. https://doi.org/10.1186/s13321-016-0116-8.
69 Duhrkop K, Fleischauer M, Ludwig M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*. 2019;16(4):299-302. https://doi.org/10.1038/s41592-019-0344-8.
70 Hoffmann MA, Nothias LF, Ludwig M, et al. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol*. 2022;40(3):411-421. https://doi.org/10.1038/s41587-021-01045-9.
71 Stravs MA, Duhrkop K, Bocker S, et al. MSNovelist: *de novo* structure generation from mass spectra. *Nat Methods*. 2022;19(7):865-870. https://doi.org/10.1038/s41592-022-01486-3.
72 Zhang M, Otsuki K, Li W. Molecular networking as a natural products discovery strategy. *Acta Mater Med*. 2023;2(2):126-141.
73 Fox Ramos AE, Evanno L, Poupon E, et al. Natural products targeting strategies involving molecular networking: different manners, one goal. *Nat Prod Rep*. 2019;36(7):960-980. https://doi.org/10.1039/C9NP00006B.
74 da Silva RR, Wang M, Nothias LF, et al. Propagating annotations of molecular networks using *in silico* fragmentation. *PLoS Comput Biol*. 2018;14(4):e1006089. https://doi.org/10.1371/journal.pcbi.1006089.
75 Watrous J, Roach P, Alexandrov T, et al. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA*. 2012;109(26):E1743-E1752. https://doi.org/10.1073/pnas.1203689109.
76 Aron AT, Gentry EC, McPhail KL, et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat Protoc*. 2020;15(6):1954-1991. https://doi.org/10.1038/s41596-020-0317-5.
77 Nothias LF, Petras D, Schmid R, et al. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods*. 2020;17(9):905-908. https://doi.org/10.1038/s41592-020-0933-6.
78 Schmid R, Petras D, Nothias LF, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun*. 2021;12(1):3832. https://doi.org/10.1038/s41467-021-23953-9.
79 Hao HP, Cui N, Wang GJ, et al. Global detection and identification of nontarget components from herbal preparations by liquid chromatography hybrid ion trap time-of-flight mass spectrometry and a strategy. *Anal Chem*. 2008;80(21):8187-8194. https://doi.org/10.1021/ac801356s.
80 Wang XX, Li C, Li ZF, et al. A structure-guided molecular network strategy for global untargeted metabolomics data annotation. *Anal Chem*. 2023;95(31):11603-11612. https://doi.org/10.1021/acs.analchem.3c00849.
81 Morehouse NJ, Clark TN, McMann EJ, et al. Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. *Nat Commun*. 2023;14(1):308. https://doi.org/10.1038/s41467-022-35734-z.
82 Wang XY, Mei J, Zhang F, et al. A ternary correlation multi-symptom network strategy based on *in vivo* chemical profile identification and metabolomics to explore the molecular basis of Ephedra Herb against viral pneumonia. *J Sep Sci*. 2024;47(11):e2400090. https://doi.org/10.1002/jssc.202400090.
83 Li XL, Guo ZF, Wen XD, et al. A molecular networking-assisted automatic database screening strategy for comprehensive annotation of small molecules in complex matrices. *J Chromatogr A*. 2023;1710:464417. https://doi.org/10.1016/j.chroma.2023.464417.
84 Cui ZR, Wang YY, Li JX, et al. Natural and pseudonatural lindenane heterodimers from *Sarcandra glabra* by molecular networking. *Org Lett*. 2022;24(49):9107-9111. https://doi.org/10.1021/acs.orglett.2c03769.
85 Li J, Cui Z, Li Y, et al. Chlospicenes A and B, cyclopropane cracked lindenane sesquiterpenoid dimers with anti-nonalcoholic steatohepatitis activity from *Chloranthus henryi*. *Chin Chem Lett*. 2022;33(9):4257-4260. https://doi.org/10.1016/j.cclet.2022.01.084.
86 Li YY, Cui ZR, Li Y, et al. Integrated molecular networking strategy enhance the accuracy and visualization of components identification: a case study of *Ginkgo biloba* leaf extract. *J Pharm Biomed Anal*. 2022; 209:114523. https://doi.org/10.1016/j.jpba.2021.114523.
87 Chen L, Li HL, Zhou HJ, et al. Feature-based molecular network-assisted

cannabinoid and flavonoid profiling of *Cannabis sativa* leaves and their antioxidant properties. *Antioxidants*. 2024;13(6):749. https://doi.org/10.3390/antiox13060749.
88 Zhang YH, Gao ZQ, Cai YL, et al. A novel strategy integrating gas phase fractionation with staggered mass range and LC-MS/MS molecular network for comprehensive metabolites profiling of Gui Ling Ji in rats. *J Pharm Biomed Anal*. 2023;222:115092. https://doi.org/10.1016/j.jpba.2022.115092.
89 Quinn RA, Melnik AV, Vrbanac A, et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature*. 2020;579(7797):123-129. https://doi.org/10.1038/s41586-020-2047-9.
90 Li YY, Zhao S, Sun YP, et al. Automatic MS/MS data mining strategy for discovering target natural products: a case of lindenane sesquiterpenoids. *Anal Chem*. 2022;94(23):8514-8522. https://doi.org/10.1021/acs.analchem.2c01559.
91 Liu FJ, Jiang Y, Li P, et al. Diagnostic fragmentation-assisted mass spectral networking coupled with *in silico* dereplication for deep annotation of steroidal alkaloids in medicinal Fritillariae Bulbus. *J Mass Spectrom*. 2020;55(9):e4528. https://doi.org/10.1002/jms.4528.
92 Chen L, Lu WY, Wang L, et al. Metabolite discovery through global annotation of untargeted metabolomics data. *Nat Methods*. 2021;18(11):1377-1385. https://doi.org/10.1038/s41592-021-01303-3.
93 Shen XT, Wang RH, Xiong X, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun*. 2019;10(1):1516. https://doi.org/10.1038/s41467-019-09550-x.
94 Zhou ZW, Luo MD, Zhang HS, et al. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun*. 2022;13(1):6656. https://doi.org/10.1038/s41467-022-34537-6.
95 Bazzano CF, de Felicio R, Alves LFG, et al. NP3 MS workflow: an open-source software system to empower natural product-based drug discovery using untargeted metabolomics. *Anal Chem*. 2024;96(19):7460-7469. https://doi.org/10.1021/acs.analchem.3c05829.
96 Nothias LF, Nothias-Esposito M, da Silva R, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. *J Nat Prod*. 2018;81(4):758-767. https://doi.org/10.1021/acs.jnatprod.7b00737.
97 Gong P, Cui N, Wu L, et al. Chemicalome and metabolome matching approach to elucidating biological metabolic networks of complex mixtures. *Anal Chem*. 2012;84(6):2995-3002. https://doi.org/10.1021/ac3002353.
98 Wang L, Ye H, Sun D, et al. Metabolic pathway extension approach for metabolomic biomarker identification. *Anal Chem*. 2017;89(2):1229-1237. https://doi.org/10.1021/acs.analchem.6b03757.
99 Oberacher H, Sasse M, Antignac JP, et al. A European proposal for quality control and quality assurance of tandem mass spectral libraries. *Environ Sci Eur*. 2020;32(1):43. https://doi.org/10.1186/s12302-020-00314-9.
100 Vinaixa M, Schymanski EL, Neumann S, et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *Trac-Trends Anal Chem*. 2016;78:23-35. https://doi.org/10.1016/j.trac.2015.09.005.
101 Blazenovic I, Kind T, Ji J, et al. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*. 2018;8(2):31. https://doi.org/10.3390/metabo8020031.
102 Cai YP, Zhou ZW, Zhu ZJ. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *Trac-Trends Anal Chem*. 2023;158:116903. https://doi.org/10.1016/j.trac.2022.116903.
103 Jarmusch AK, Wang M, Aceves CM, et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat Methods*.2020;17(9):901-904. https://doi.org/10.1038/s41592-020-0916-7.
104 Gauglitz JM, West KA, Bittremieux W, et al. Enhancing untargeted metabolomics using metadata-based source annotation. *Nat Biotechnol*. 2022;40(12):1774-1779. https://doi.org/10.1038/s41587-022-01368-1.
105 Zuffa S, Schmid R, Bauermeister A, et al. microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat Microbiol*. 2024;9(2):336-345. https://doi.org/10.1038/s41564-023-01575-9.
106 Mohanty I, Mannochio-Russo H, Schweer JV, et al. The underappreciated diversity of bile acid modifications. *Cell*. 2024;187(7):1801-1818. https://doi.org/10.1016/j.cell.2024.02.019.
107 Gentry EC, Collins SL, Panitchpakdi M, et al. Reverse metabolomics for the discovery of chemical structures from humans. *Nature*. 2024;626(7998):419-426. https://doi.org/10.1038/s41586-023-06906-8.
108 Yan TT, Nie LL, Hao HP. Reverse metabolomics as a novel strategy to annotate the human metabolome. *Chin J Nat Med*. 2024;22(4):289-290. https://doi.org/10.1016/S1875-5364(24)60589-3.
109 Nie QX, Luo X, Wang K, et al. Gut symbionts alleviate MASH through a secondary bile acid biosynthetic pathway. *Cell*. 2024;187(11):2717-2734. https://doi.org/10.1016/j.cell.2024.03.034.
110 Li Q, Yin YH, Liu ZW, et al. FNICM: a new methodology to identify core metabolites based on significantly perturbed metabolic subnetworks. *Anal Chem*. 2024;96(8):3335-3344. https://doi.org/10.1021/acs.analchem.3c04131.
111 Hou YL, He DD, Ye L, et al. An improved detection and identification strategy for untargeted metabolomics based on UPLC-MS. *J Pharm Biomed Anal*. 2020;191:113531. https://doi.org/10.1016/j.jpba.2020.113531.
112 Lin JC, Yang X, Wang AH, et al. LC-MS/MS profiling of colon oxysterols and cholesterol precursors in mouse model of ulcerative colitis. *J Chromatogr A*. 2024;1722:464865. https://doi.org/10.1016/j.chroma.2024.464865.
113 Yuan BF, Zhu QF, Guo N, et al. Comprehensive profiling of fecal metabolome of mice by integrated chemical isotope labeling-mass spectrometry analysis. *Anal Chem*. 2018;90(5):3512-3520. https://doi.org/10.1021/acs.analchem.7b05355.
114 Zhao S, Li H, Han W, et al. Metabolomic coverage of chemical-group-submetabolome analysis: group classification and four-channel chemical isotope labeling LC-MS. *Anal Chem*. 2019;91(18):12108-12115. https://doi.org/10.1021/acs.analchem.9b03431.
115 Qin SY, Gao MY, Zhang QQ, et al. High-coverage strategy for multi-subcellular metabolome analysis using dansyl-labeling-based LC-MS/MS. *Anal Chem*. 2023;95(26):10034-10043. https://doi.org/10.1021/acs.analchem.3c01343.
116 Zhao AQ, Zheng JY, Chen C, et al. Enzyme-driven LC-HRMS approach for specific recognition of 12α-hydroxy bile acids. *Anal Chem*. 2024;96(21):8613-8621. https://doi.org/10.1021/acs.analchem.4c00676.
117 Wandy J, Davies V, van der Hooft JJJ, et al. *In silico* optimization of mass spectrometry fragmentation strategies in metabolomics. *Metabolites*. 2019;9(10):219. https://doi.org/10.1101/744227.
118 Guo T, Shi YY, Zheng L, et al. Rapid and simultaneous determination of sulfonate ester genotoxic impurities in drug substance by liquid chromatography coupled to tandem mass spectrometry: comparison of different ionization modes. *J Chromatogr A*. 2014;1355:73-79. https://doi.org/10.1016/j.chroma.2014.05.079.
119 Xu JD, Xu MZ, Zhou SS, et al. Effects of chromatographic conditions and mass spectrometric parameters on the ionization and fragmentation of triterpene for saponins of in liquid chromatography-mass spectrometry analysis. *J Chromatogr A*. 2019;1608:460418. https://doi.org/10.1016/j.chroma.2019.460418.
120 Wang XX, Sun XS, Wang FB, et al. Enhancing metabolome annotation by electron impact excitation of ions from organics-molecular networking. *Anal Chem*. 2024;96(4):1444-1453. https://doi.org/10.1021/acs.analchem.3c03443.
121 Zhang YH, Liao JY, Le WQ, et al. Improving the data quality of untargeted metabolomics through a targeted data-dependent acquisition based on an inclusion list of differential and preidentified ions. *Anal Chem*. 2023;95(34):12964-12973. https://doi.org/10.1021/acs.analchem.3c02888.
122 Chen YH, Bi JH, Xie M, et al. Classification-based strategies to simplify complex traditional Chinese medicine (TCM) researches through liquid chromatography-mass spectrometry in the last decade (2011–2020): theory, technical route and difficulty. *J Chromatogr A*. 2021;1651:462307. https://doi.org/10.1016/j.chroma.2021.462307.
123 Fan YL, Liu RZ, Tan Q, et al. A database-guided integrated strategy for comprehensive chemical profiling of traditional Chinese medicine. *J Chromatogr A*. 2022;1674:463145. https://doi.org/10.1016/j.chroma.2022.463145.
124 Dai C, Wang C, Zhang CH, et al. A reference substance free diagnostic fragment ion-based approach for rapid identification of non-target components in Pudilan Xiaoyan Oral Liquid by high resolution mass spectrometry. *J Pharm Biomed Anal*. 2016;124:79-92. https://doi.org/10.1016/j.jpba.2016.02.020.
125 Cheng XL, Wan JY, Li P, et al. Ultrasonic/microwave assisted extraction and diagnostic ion filtering strategy by liquid chromatography-quadrupole time-of-flight mass spectrometry for rapid characterization of flavonoids in. *J Chromatogr A*. 2011;1218(34):5774-5786. https://doi.org/10.1016/j.chroma.2011.06.091.
126 Qi LW, Wang HY, Zhang H, et al. Diagnostic ion filtering to characterize ginseng saponins by rapid liquid chromatography with time-of-flight mass spectrometry. *J Chromatogr A*. 2012; 1230:93-99. https://doi.org/10.1016/j.chroma.2012.01.079.
127 Ding M, Jiang Y, Gao W, et al. Characterization and quantification of chemical constituents in Angong Niuhuang Pill using ultra-high performance liquid chromatography tandem mass spectrometry. *J Pharm Biomed Anal*. 2023;228:115309. https://doi.org/10.1016/j.jpba.2023.115309.
128 Tian JX, Tian Y, Xu L, et al. Characterisation and identification of dihydroindole-type alkaloids from processed Semen Strychni by high-performance liquid chromatography coupled with electrospray ionisation ion trap time-of-flight mass spectrometry. *Phytochem Anal*. 2014;25(1):36-44. https://doi.org/10.1002/pca.2457.
129 Huang AX, Li JM, Yang L, et al. A mass defect filtering combined background subtraction strategy for rapid screening and identification of metabolites in rat plasma after oral administration of Yindan Xinnaotong Soft Capsule. *J Pharm Biomed Anal*. 2023;231:115400. https://doi.org/10.1016/j.jpba.2023.115400.
130 Zeng SL, Duan L, Chen BZ, et al. Chemicalome and metabolome profiling of polymethoxylated flavonoids in Citri Reticulatae Pericarpium based on an integrated strategy combining background subtraction and modified mass defect filter in a Microsoft Excel Platform. *J Chromatogr A*. 2017;1508:106–120. https://doi.org/10.1016/j.chroma.2017.06.015.
131 Xie T, Liang Y, Hao HP, et al. Rapid identification of ophiopogonins and ophiopogonones in extract with a practical technique of mass defect filtering based on high resolution mass spectrometry. *J Chromatogr A*. 2012;1227:234-244. https://doi.org/10.1016/j.chroma.2012.01.017.
132 Zhou W, Shan JJ, Meng MX. A two-step ultra-high-performance liquid chromatography-quadrupole/time of flight mass spectrometry with mass defect filtering method for rapid identification of analogues from known components of different chemical structure types in herb pair extract and in rat's blood. *J Chromatogr A*. 2018;1563:99-123. https://doi.org/10.1016/j.chroma.2018.05.067.
133 Yu T, Chen JM, Liu W, et al. In-depth characterization of cycloartane triterpenoids and discovery of species-specific markers from three *Cimicifuga* species guided by a strategy that integrates in-source fragment elimination, diagnostic ion recognition, and feature-based molecular networking. *J Chromatogr A*. 2024;1728:465015. https://doi.org/10.1016/j.chroma.2024.465015.