Chinese Journal of Natural Medicines

Advances in intelligent mass spectrometry data processing technology for *in vivo* analysis of natural medicines

Simian CHEN, Binxin DAI, Dandan ZHANG, Yuexin YANG, Hairong ZHANG, Junyu ZHANG, Di LU, Caisheng WU

Citation: Simian CHEN, Binxin DAI, Dandan ZHANG, Yuexin YANG, Hairong ZHANG, Junyu ZHANG, Di LU, Caisheng WU, Advances in intelligent mass spectrometry data processing technology for *in vivo* analysis of natural medicines, Chinese Journal of Natural Medicines, 2024, 22(10), 900−913. doi: 10.1016/S1875−5364(24)60687−4.

View online: https://doi.org/10.1016/S1875−5364(24)60687−4

Available online at www.sciencedirect.com
ScienceDirect
Chinese Journal of Natural Medicines 2024, **22**(10): 900-913
doi: 10.1016/S1875-5364(24)60687-4
Chinese Journal of Natural Medicines

•Review•

Advances in intelligent mass spectrometry data processing technology for *in vivo* analysis of natural medicines

CHEN Simian[1Δ], DAI Binxin[1Δ], ZHANG Dandan[1], YANG Yuexin[1], ZHANG Hairong[1], ZHANG Junyu[1], LU Di[1], WU Caisheng[1, 2*]

[1] *Fujian Provincial Key Laboratory of Innovative Drug Target Research and State Key Laboratory of Cellular Stress Biology, School of Pharmaceutical Sciences, Xiamen University, Xiamen 361102, China;*
[2] *Xiamen Key Laboratory for Clinical Efficacy and Evidence-Based Research of Traditional Chinese Medicine, Xiamen University, Xiamen 361102, China*

Available online 20 Oct., 2024

[ABSTRACT] Natural medicines (NMs) are crucial for treating human diseases. Efficiently characterizing their bioactive components *in vivo* has been a key focus and challenge in NM research. High-performance liquid chromatography-high-resolution mass spectrometry (HPLC-HRMS) systems offer high sensitivity, resolution, and precision for conducting *in vivo* analysis of NMs. However, due to the complexity of NMs, conventional data acquisition, mining, and processing techniques often fail to meet the practical needs of *in vivo* NM analysis. Over the past two decades, intelligent spectral data-processing techniques based on various principles and algorithms have been developed and applied for *in vivo* NM analysis. Consequently, improvements have been achieved in the overall analytical performance by relying on these techniques without the need to change the instrument hardware. These improvements include enhanced instrument analysis sensitivity, expanded compound analysis coverage, intelligent identification, and characterization of nontargeted *in vivo* compounds, providing powerful technical means for studying the *in vivo* metabolism of NMs and screening for pharmacologically active components. This review summarizes the research progress on *in vivo* analysis strategies for NMs using intelligent MS data processing techniques reported over the past two decades. It discusses differences in compound structures, variations among biological samples, and the application of artificial intelligence (AI) neural network algorithms. Additionally, the review offers insights into the potential of *in vivo* tracking of NMs, including the screening of bioactive components and the identification of pharmacokinetic markers. The aim is to provide a reference for the integration and development of new technologies and strategies for future *in vivo* analysis of NMs.

[KEY WORDS] High-performance liquid chromatography–High-resolution mass spectrometry; Data-acquisition; Data-processing; Artificial Intelligence; Metabolomics

[CLC Number] R917 **[Document code]** A **[Article ID]** 2095-6975(2024)10-0900-14

Introduction

Compared with chemical drugs, the primary challenge in researching natural medicines (NMs) lies in their complexity. Each NM consists of hundreds of chemical compounds, which significantly increases the difficulty of investigation. NMs are typically intricate systems, and efficiently characterizing their active components and understanding the associated mechanisms has been a focal challenge in research. The efficacy of most NMs relies on their active ingredients, which interact with receptors upon entering the systemic circulation and reaching target sites. Therefore, *in vivo* analysis of NMs can investigate the effective utilization of their components after administration and help narrow down the screening of natural active compounds, providing essential support for understanding the material basis of NMs' efficacy.

The high-performance liquid chromatography-high-resolution mass spectrometry (HPLC-HRMS) system offers significant advantages in terms of sensitivity, resolution, and accuracy. Therefore, it is highly suitable for *in vivo* analysis of

[Received on] 28-Jun.-2024
[Research funding] This work was supported by the National Natural Science Foundation of China (Nos. 82222068, 82141215 and 82173779), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (No. ZYYCXTD-D-202206), the Science and Technology Project of Fujian Province (Nos. 2022J02057, 2021J02058 and 2021I0003), the S&T Program of Hebei Province (No. 23372508D).
[*Corresponding author] E-mail: wucsh@xmu.edu.cn
[Δ]These authors contributed equally to this work.
These authors have no conflict of interest to declare.

NMs. However, due to the complexity of NMs, several critical scientific challenges remain unresolved. First, NM compounds exhibit considerable structural diversity, resulting in substantial variations in chromatographic and mass spectrometric behaviors. Therefore, achieving higher sensitivity, comprehensive compound analysis, and enhanced efficiency in secondary data collection are essential prerequisites for the effective analysis of NM components *in vivo*. Second, the data collected from the HRMS system constitutes a massive dataset, necessitating intelligent and efficient methods to filter the MS data pertinent to NM components *in vivo*. Finally, correlating prototypes and metabolites is necessary for their structural identification. Various novel intelligent MS data technologies are urgently required to resolve these critical scientific challenges.

Over the past two decades, intelligent spectral data processing techniques based on various principles and algorithms have been developed and applied for NM substance analysis in *in vivo* studies. Initially, MS data mining techniques primarily relied on templates based on prototype drugs or their core structures, such as the mass defect filter (MDF) and isotope pattern filter (IPF). However, these algorithms were designed around specific parent nuclei or substructures and faced clear limitations when applied to complex systems such as NMs with multiple components. Subsequently, nontargeted techniques, such as background subtraction (BS) and the metabolomics approach (MA), have been widely applied. These techniques do not require prior knowledge of compound structures and rely on differences between test and control samples to achieve nontargeted acquisition and mining of MS data from NMs *in vivo*. As a powerful tool for data processing and analysis, artificial intelligence (AI) plays an increasingly important role in NM research, contributing to structural identification, noise reduction, and advanced data mining and pattern recognition. Thus, the integration of HPLC-HRMS with intelligent MS data processing techniques provides a powerful tool for the *in vivo* analysis of NMs.

Previous reviews have outlined the application of MS technology for drug component analysis, metabolite identification, and absorption, distribution, metabolism, and excretion (ADME) profiling [1-5]. Building on these studies, this review aims to systematically summarize the development of intelligent MS data processing techniques over the past two decades. Additionally, it examines recent advances in the *in vivo* analysis strategies for NMs, focusing on structural diversity, biological variability, and the integration of AI-driven neural network algorithms. Despite the importance of this multifaceted approach, it has not been comprehensively reviewed to date.

The workflow combining HPLC-HRMS with intelligent MS data-processing techniques for *in vivo* analysis of NMs is shown in Fig. 1. This review aims to provide a valuable reference for the future integration and evolution of novel techno-

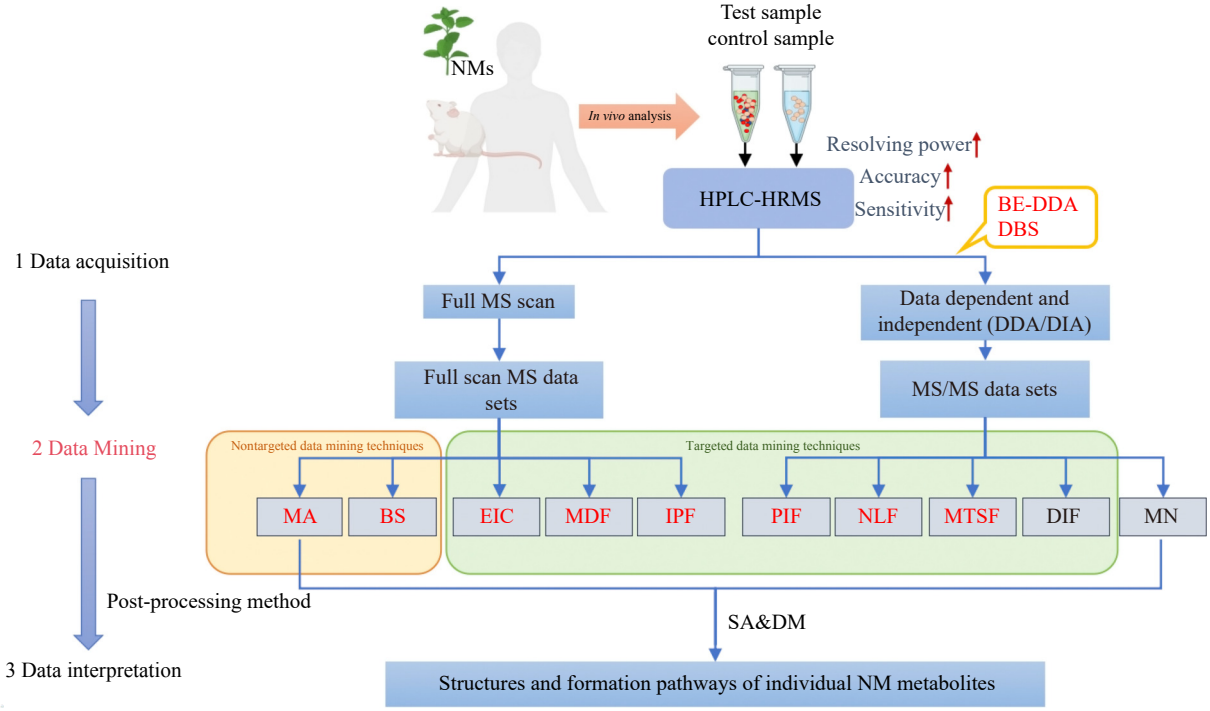

Fig. 1 The workflow of HPLC-HRMS combined with intelligent MS data-processing technology for *in vivo* analysis of NMs. NM, natural medicine; BE-DDA, background exclusion data-dependent acquisition; DBS, dynamic background subtraction; MA, metabolomics approach; BS, background subtraction; EIC, extracted ion chromatography; MDF, mass defect filter; IPF, isotope pattern filter; PIF, product ion filter; NLF, neutral loss filter; MTSF, mass spectral trees similarity filter; DIF, diagnostic ion filter; MN, molecular networking; SA, statistical analysis; DM, database matching.

logies and methodologies in the *in vivo* analysis of NMs.

Development of Intelligent MS Data Processing Techniques Based on NM Structural Characteristics

The active components of NMs include multiple compounds with different structures, such as flavonoids, phenylpropanoids, anthraquinones, steroidal saponins, and alkaloids. These metabolites, despite their structural diversity, frequently share a common parent or core structure. This commonality facilitates the development of targeted data-processing techniques grounded in structural characterization.

Application of single-template chemical–drug metabolite analysis to in vivo analysis of NMs

Targeted postprocessing techniques using prototype drugs as template compounds for *in vivo* analysis of single compounds can be generally categorized into three types. (1) Detection of conventional metabolites formed from known or predictable metabolic pathways, such as hydroxylation and *N*-dealkylation. High-resolution extracted ion chromatography (HR-EIC) can then be used to identify the expected metabolites based on their predicted molecular weights. This method involves acquiring a full scan in a liquid chromatography-tandem mass spectrometry (LC-MS/MS) instrument and applying an ion extraction window to the acquired full-scan MS dataset to improve selectivity and sensitivity, thereby identifying the desired metabolite ion chromatogram. (2) Postprocessing techniques based on primary MS data such as MDF and IPF. (3) Postprocessing techniques based on multistage MS data, including a product ion filter (PIF) and neutral loss filter (NLF). The advantages and disadvantages of these MS smart-processing methods are shown in Table 1. These techniques can be combined to enable the effective discovery of relevant metabolites of chemically synthesized drugs or NM single-active components and have been effectively applied in metabolite identification studies of chemically synthesized drugs.

For the *in vivo* analysis of NM monomers, which are essentially analogous to chemical drugs, a combination of these techniques can also be applied to achieve rapid and facile discovery of the NM monomer metabolites *in vivo* [6-9]. ZHAO *et al*. analyzed the soybean glycoside metabolism in rats using ultrahigh performance liquid chromatography (UPLC)-LTQ-Qrbitrap MS. By using various data mining methods, such as HR-EIC, MDF, NLF, and diagnostic product ion, 59 metabolites, including the prototype compounds, were finally identified and characterized [7]. In addition to revealing the potential pharmacodynamic forms of soybean glycosides, this study contributed to establishing practical strategies for the rapid screening and identification of metabolites of natural compounds.

Furthermore, since components of the same class in NMs often share a common carbon skeleton or substructure, they tend to produce similar fragment ions. Techniques, such as PIF, NLF, and MDF, are commonly used for *in vivo* component screening based on the structural characteristics of compounds. These techniques have achieved favorable results in identifying parent components with similar structures to compounds found in NMs. PIF and NLF, in particular, are postprocessing techniques focusing on multilevel data from HRMS. They rely on tandem MS data to identify *in vivo* metabolites by detecting fragment ions or neutral losses generated during MS fragmentation, which reflect the characteristic cleavage patterns of particular compound classes [10-12]. For example, QIAO *et al*. applied techniques, such as neutral loss, precursor ion scanning, and selected-reaction monitoring, to analyze the *in vivo* components of traditional Chinese medicine (TCM) [10]. In this study, 131 metabolites were detected in rats administered Gegen-Qinlian Decoction, 85 of which were identified for the first time. Structurally similar compounds, such as original components and their metabolites, often exhibit minor mass losses, which MDF utilizes to rapidly screen high-resolution primary MS data and identify structurally related components of NMs *in vivo* [13-15]. For example, active components such as cyclic enol ether terpenoids and glycosides are present in the extract of *Dioscorea*. TAO *et al*. used MDF to characterize the metabolites of catalpol and acteoside in biological samples after the oral administration of extracts to rats. This method was successfully applied to compare the metabolic differences between normal rats and those with chronic kidney disease [15].

Development of multitemplate-targeted screening technology based on NM characteristics

NMs are complex drug systems, and screening techniques that only use a single template have significant limitations. To enhance the efficiency of *in vivo* screening of NMs,

Table 1 Summary of the characteristics of single-template MS processing techniques for NMs *in vivo* analysis

Technology	Advantages	Limitations
HR-EIC	High selectivity and sensitivity	Only applicable to known or predictable metabolites
MDF	Effectively remove interference of complex biological matrices; detection of both predictable and unpredictable metabolites.	Prone to generating false positive results; The performance of MDF technology is greatly influenced by parameter settings.
IPF	Effective removal of matrix ions that do not follow the principle of isotopes; detection of both predictable and unpredictable metabolites.	Loss of low intensity data points; Isobaric interference; requirements of higher resolution and quality accuracy; isotope effect.
PIF/NLF	Effectively screen for metabolites with the same structure; detection of both predictable and unpredictable metabolites.	It is necessary to summarize and determine the characteristic product ion and neutral loss fragment; poor selectivity in DIA acquisition.

several multi-template screening techniques have been developed. In 2008, WANG *et al*. searched for common diagnostic ions from all experimentally generated ions. The components sharing the exact same ions (mass error < 5 mDa) were classified into a family. All families were then connected into a coherent network by the bridging components [16]. This strategy enables a narrowing of the database hits, and ingredients included in the network were successfully identified from the tested herbal preparations. Drug metabolites are the result of metabolic transformations of administered drugs, and their structures have similarities with the parent drugs or other metabolites. These similarities extend to their mass spectral data. In 2013, JIN *et al*. constructed a mass spectral trees similarity filter (MTSF) based on this basic assumption (Fig. 2A). This technique organizes the collected MS data into a dendrogram format *via* intelligent recognition algorithms. The mass spectral dendrograms of potential NM-related components are then compared with those from template compound libraries (e.g., medicinal liquid components) to assess structural similarity. The method calculates a match probability score, with a maximum value of 1000, based on parameters such as the mass deviations between "parent" and "daughter" ions and the match of "daughter ion" types. If no peak matches, no score is assigned. This scoring system allows researchers to screen NM components related to *in vivo* metabolites. Using the same principles, correlations between *in vivo* components can be identified, enabling the construction of a metabolic network for NMs. MTSF is particularly effective because it leverages multilevel MS data to identify both exogenous and endogenous NM components, without being constrained by predictable metabolic pathways. JIN *et al*. successfully applied this technique to identify five flavonoid monomers and 115 metabolites in *Epimedium* extract in rats, uncovering the metabolic profiles of *Epimedium* and proposing the relevant metabolic pathways [17].

MDF is a metabolite mining technique developed specifically for discovering unpredictable metabolites. However, because of the composition of NMs, various of transformations can occur, such as the hydrolysis of multiple glycosidic bonds. These changes result in significant variability in the mass loss of metabolites comparing their original forms. Based on MDF, a multiple mass defect filter (MMDF) was introduced by several instrument companies. This technique allows multiple template compounds to be established simultaneously, selecting appropriate windows of mass ranges and mass defects, connecting each extreme value to form a corresponding rectangular filtering region, and screening *in vivo* components with large motif changes. The technique has been successfully built into several commercially available software packages, including PeakView (Bruker Corporation, USA), MetWorks (Thermo Fisher Scientific, USA), MetaboLynx and MetID (Waters Corporation, USA), that are generalizable and widely used for mining and identifying NM *in vivo* components [18-22]. ZHOU *et al*. used a two-step MDF screen to definitively or preliminarily identify 36 compounds in extract and 16 metabolites from plasma administered in Fructus Gardeniae-Fructus Forsythiae herb pair extract [18]. This enabled the rapid identification of unknown analogs from representative compounds of various structural types, and by using the compounds identified in the first step, other unknown components can be swiftly identified. ZHANG *et al*. applied the MMDF method to detect the *in vivo* metabolites of baicalin in rats and used a total of five MDF templates simultaneously, including a drug filter, a substructure filter, and three conjugate filters, to identify the metabolite of 6'-hydroxy-3,4,5,2',4'-pentamethoxychalcone and baicalin [23]. However, a major limitation of the technique is the large number of false-positive peaks that may be generated when using a wider mass defect window for screening [24]. To address such shortcomings, GAO *et al*. proposed a new post-acquisition processing method, Molecule- and Fragmentation-driven Mass Defect Filters (MF-MDFs), which was developed based on the fragmentation pattern of the parent drug. To study the metabolism of paclitaxel, six filter templates (one for molecular ions and the remaining five for frag-

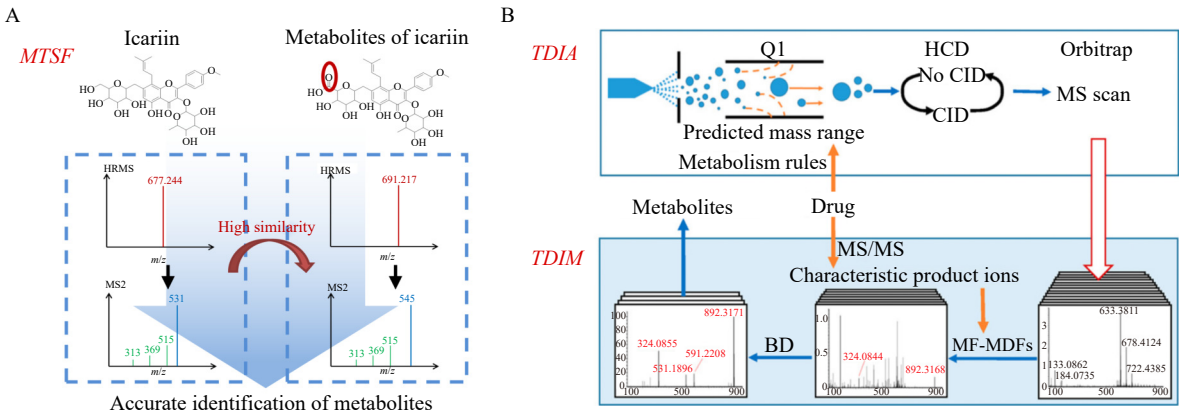

Fig. 2 (A) A Schematic diagram of using HPLC-HRMS and the MTSF technique for the discovery and identification of metabolites [17]. (B) Schematic diagram of TDIAM. BD: blind deconvolution; MF-MDFs: molecule- and fragmentation-driven mass defect filter [25]. (A higher resolution/color version of this figure is available in the electronic copy of the article).

ment ions) were designed and applied to the dataset, which was processed by blinded deconvolution with the resolution, precision, and filtering parameters. A total of 10 paclitaxel metabolites were then identified by MF-MDFs (Fig. 2B) [25].

Intelligent MS Data-processing Technology Developed Based on Differences Between Test Samples and Control Samples

Since NMs contain hundreds to thousands of structurally diverse compounds, any targeted approach is a highly time-consuming and labor-intensive task. Crucially, many key active ingredients in NMs are derived through multistep metabolic transformations, making targeted metabolite discovery techniques prone to failure. To overcome the limitations of targeted techniques, various nontargeted data-mining techniques have been developed, represented by the BS technology that compares the differences between control and administered test samples. These technological strategies have been used for HRMS data acquisition throughout the entire process of nontargeted analysis.

HRMS data-acquisition technology based on biological sample differences

With the advent of tandem HRMS, the scanning speed of HRMS can meet the demands of UPLC combined use, and the automated acquisition of HRMS data is not technically challenging. For the acquisition of MS/MS datasets, data-dependent acquisition (DDA) and data-independent acquisition (DIA) have been commonly used. DIA can fragment all ions within a mass-isolated window to achieve nontargeted data acquisition, which has the advantages of comprehensive data acquisition and data reproducibility. Theoretically, DIA can obtain data on all precursor ions and their product ions; however, the weak correlation between MS^1 and MS/MS complicates the identification and differentiation of endogenous and exogenous compounds [26-27]. In the DDA mode, after acquiring the MS^1 spectra, the precursor ions of the top N ions are selected to be fragmented to provide the MS/MS spectrum with improved relevant connections. This data-acquisition method will capture a high abundance of MS^1, which is then fragmented [28], while most of the pharmacophoric components are not easily detected at low levels *in vivo* and produce unavailable MS/MS spectra.

To enhance the efficiency of the acquisition of NM *in vivo* components, CHEN *et al*. used dynamic DDA data to analyze the major components of NMs and identify typical fragments. Then, DIA MS/MS data of the NM-dosing solution was used to confirm the ability of DIA to deliver typical fragments similar to DDA. Subsequently, a high-resolution PIF was used to extract data from the administered test samples collected *via* DIA. The integration of data-processing techniques enables the efficient, nontargeted identification of NM-related components *in vivo*, identifying 71 anthraquinone prototype components, including 36 new discoveries [29]. ZHU *et al*. utilized the differences between control and administered test samples to propose an intelligent nontargeted acquisition method for secondary data called background exclusion-data-dependent acquisition (BE-DDA) in the pharmacological mechanism study of Xiaoke Pill [30]. The principle of BE-DDA is illustrated in Fig. 3A. After performing four sequential HPLC-HRMS injections, comprehensive and accurate MS data is captured for drug constituents in plasma and urine components. The first injection is a control sample to obtain the exclusion list. The second injection is a test sample, and according to the MS^1 data collected in real-time, ions in the exclusion list are deducted and dynamically collected based on the strength of the ions in the real-time data. Successfully collected ions were dynamically added to the exclusion list. Repeat feeds of the same test sample are performed for a third and fourth time, and the exclusion list is dynamically updated based on this collection. During this process, continuous elimination of the interference of background ions and reduction of duplicate acquisitions enabled the efficient collection of information for low-abundance NM components. The results showed that 97.4% and 93.6% of the MS/MS data of Xiaoke Pill constituents were acquired from rat plasma and urine, respectively. In summary, BE-DDA has great sensitivity and selectivity in triggering MS/MS acquisition of unknown NM constituents in complex samples. It enhances analytical specificity, aiding in the detection and analysis of target compounds while reducing non-specific interference. In 2022, GUO *et al*. established a simple and efficient online stepwise BS-based UPLC-quadrupole time-of-flight tandem MS (UPLC-Q-TOF-MS/MS) dynamic detection method, which combines BS analysis with dynamic MS/MS detection to acquire as many ions of interest as possible until no more significant increase in new ions occurs in the target. Compared with conventional MS/MS detection, the step-by-step dynamic MS/MS method significantly improves the capture of target ions. In addition, this technique minimizes the interference of background ions from the data acquisition source and reduces the burden of post-acquisition data analysis compared to BS methods for offline post-acquisition data processing. Moreover, this technique was used to effectively access NM prototypes and metabolites in plasma samples of rats following the administration of Qiang Huo Sheng Shi decoction [31].

With the development of intelligent MS acquisition technology, an increasing number of research teams are using commercial software (e.g., Metabolite-Pilot™ developed by SCIEX; UNIFI developed by Waters) to integrate data acquisition techniques like dynamic background subtraction (DBS) and MMDF with DDA scanning (Fig. 3B). This approach aims to improve the efficiency of identifying drug metabolites *in vivo* using DBS and MMDF methods while minimizing the detection of irrelevant ions. Research groups led by FENG [22], ZHANG [32], and LAI [21] have applied MMDF- and DBS-dependent online data acquisition techniques for the comprehensive and systematic detection and identification of *in vivo* metabolites of Zhi-zi-chi decoction, *Cirsium japonicum* DC., and Duzhi pill in rat tissues, respectively.

Nontargeted mining techniques based on BS HRMS datasets

The BS technique can be applied in data collection and has been extensively used in nontargeted metabolite profiling *in vivo*. By comparing test samples with control samples, the chromatographic peaks of endogenous components in the test samples were removed, allowing the selection of chromatographic peaks corresponding to exogenous components. This method is characterized by its simplicity, speed, and ability to efficiently streamline HRMS data, making it a valuable tool for simplifying complex datasets.

In 2008, ZHANG *et al*. developed a BS algorithm for better sample control. This algorithm aimed to completely remove matrix-related interference from HPLC-HRMS data and separate ions of interest in metabolites. It examines con-

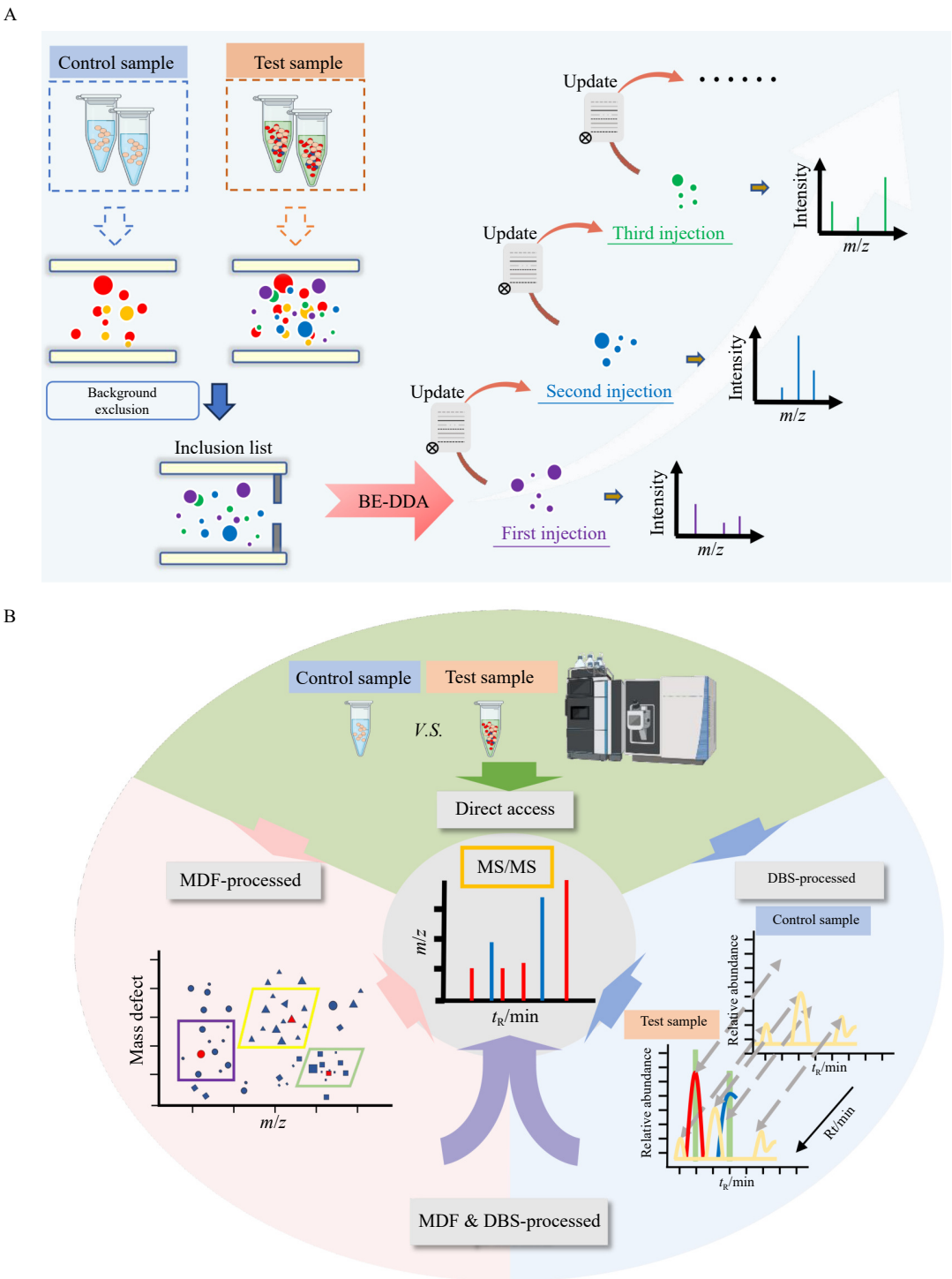

Fig. 3 Schematic diagram of representative nontargeted collection techniques for *in vivo* analysis of NMs. (A) BE-DDA Technology Principle Schematic. (B) MMDF and DBS Technology Principle Schematic.

trol datasets within defined time windows around the analysis dataset to effectively subtract background signals from this dataset [33]. To identify glutathione-reactive metabolites, a controlled scanning time of ± 1 min and a mass error tolerance of ± 5 × 10^{-6} was used. This approach has been validated for the identification of reactive metabolites of diclofenac, clozapine, promethazine, and thalidomide. In the same year, ZHANG *et al*. developed a comprehensive BS algorithm for the identification of metabolites in biological samples (rat plasma, bile, and urine), which was validated by determining the presence of rosiglitazone metabolites in biological matrices [34]. The principle is shown in Fig. 4A. However, simply subtracting the background signal from biological matrices is not highly efficient and may increase electrical noise. In 2009, ZHU *et al*. proposed an additional noise reduction function for BS based on the ZHANG *et al*. algorithm. This function, referred to as the BS and Noise Reduction Algorithm (BgS-NoRA), helps to eliminate noise from residual matrix ions after BS [35]. In this procedure, if the target ion is present within the mass tolerance and retention time window of the control sample, its intensity is multiplied by a predefined scaling factor and subtracted from the intensity of ions in the test sample. In typical LC-MS data, noise tends to be random, while true analyte signals are consistent across multiple scans. Accordingly, NoRA further refines the data after BS by removing inconsistent ion signals from adjacent scans. A detailed workflow of the BgS-NoRA algorithm is illustrated in Fig. 4B. However, BgS-NoRA is a time-intensive process, and to address this issue, Shekar *et al*. developed a novel and effective BS algorithm (A-BgS) in 2016, which significantly reducing processing time compared with previous methods [36]. This improvement is mainly attributed to the use of graphics processing units (GPUs), which execute repetitive calculations at high speed. Compared with using CPU computation alone, the processing speed is improved by thousands of times. An overview of the A-BgS software is presented in Fig. 4C.

Owing to the complexity and diversity of NM components, achieving ideal results with currently commercialized BS techniques remains difficult. The techniques reported for studying NM components *in vivo* are mostly laboratory-developed. In previous studies, based on the fundamental principle of BS, we designed an in-house precise and thorough background subtraction technique (PATBS) targeting the comprehensive exploration of NM-related components in complex biological samples [37]. In this process, the precise mass numbers of the control sample data are first defined within the specified time window (such as background or matrix, i.e., background ion peaks) and the mass accuracy window of ions in the analysis sample data (± 10 × 10^{-6}). A comprehensive analysis of the collected HPLC-HRMS data spectra was conducted using the dynamic scan algorithm. Once background ion peaks were detected in the analysis sample, the algorithm identified the same ions within the same mass difference window in the control sample. The highest intensity of the identified ions in the control sample was determined and multiplied by a specified scaling factor (usually between 1 and 100, based on the sensitivity to endo-

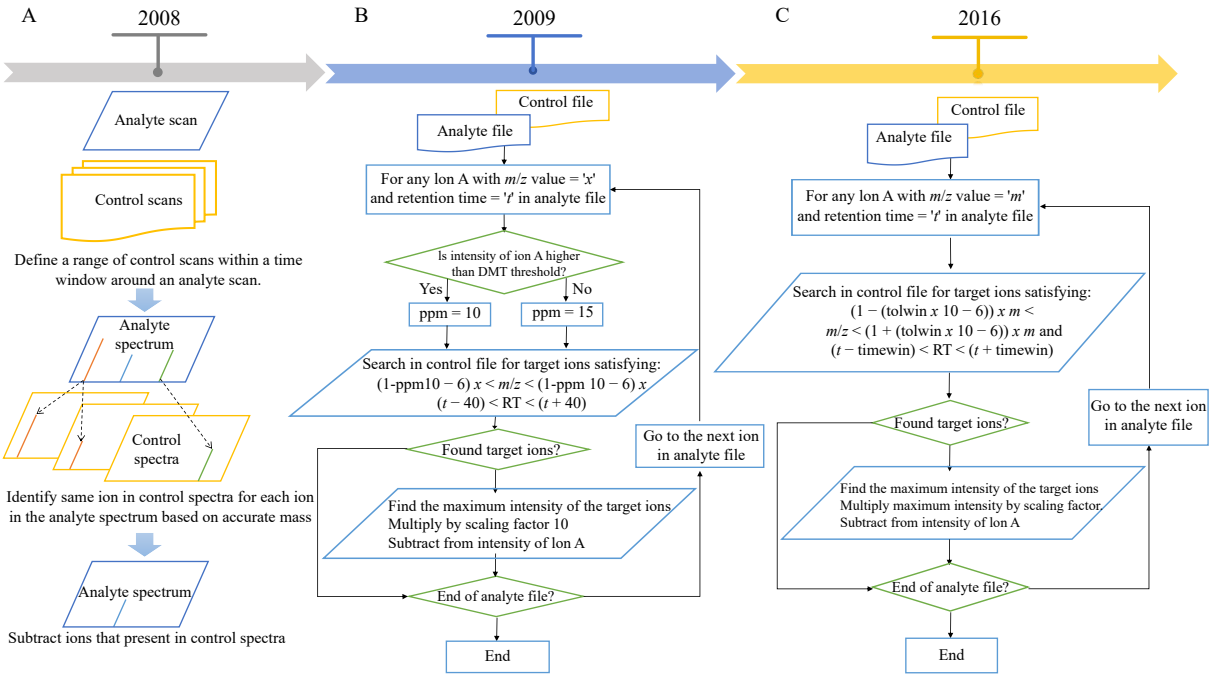

Fig. 4 (A) Schematic illustration of thorough BS of an analyte spectrum. The algorithm is looped throughout the analyte scans in an LC/MS data set. (Analyte and control denote that the scans and spectra are associated either with an analyte sample or a control sample) [34]. (B) Flowcharts for the background-subtraction algorithm A and the NoRA algorithm B [35]. (C) Development of ABgS algorithm [36].

genous fluctuations between samples), and the ion intensity was then subtracted from the scanning spectrum of the analyte. Thus, within a specific time window, the intensity of all ions present simultaneously in the analysis and control samples can be significantly reduced, whereas the intensity of ions solely present in the analysis sample remains unchanged, ultimately achieving comprehensive subtraction of background ion peaks without subtracting signals from exogenous components (NM components). This technique has been successfully applied to the comprehensive exploration of components related to the metabolism of Xiaoke Pill in zebrafish and rats [30, 38]. After PATBS processing of the TIC of biological samples from rats and zebrafish following administration, major endogenous interference peaks are precisely and completely subtracted. Ultimately, 115 and 109 Xiaoke Pill-related components were respectively identified from rat plasma and urine; 39 compounds (13 prototype components and 26 metabolites) and 10 compounds were screened from zebrafish tissues and culture media, demonstrating the efficient identification of NM-related components in complex biological matrices.

Nontargeted mining technique based on HRMS datasets using metabolomics

Metabolomics examines the dynamic patterns of small-molecule metabolites in organisms, focusing on their diversity and quantity under various physiological conditions, such as during disease or therapeutic intervention. By applying the biomarker screening model of metabolomics to identify *in vivo* NM components, the nontargeted and unbiased detection of relevant NM components can theoretically be achieved. This metabolomics strategy is based on statistical differences observed in samples before and after dosing, and when combined with high-throughput analytical techniques and multivariate statistical methods—such as principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and orthogonal partial least squares discriminant analysis (OPLS-DA)—it allows for the identification of significantly different components between control and test samples. PCA, an unsupervised technique, reduces the dimensionality of datasets by transforming the original variables into a smaller number of uncorrelated principal components (PCs). For example, Plumb *et al.* used LC-TOF-MS with PCA to reveal clusters of ions between treated and control groups (Fig. 5A), allowing for the detection of GSK-X drug metabolites without requiring prior knowledge of the test substance's chemical structure [39]. PLS-DA, a supervised method, combines partial least squares regression with discriminant analysis to uncover latent variables that optimize category separation and correlate with response variables. OPLS-DA improves upon PLS-DA by adding an orthogonal component to separate variation unrelated to class differences, thus enhancing the clarity of class-related information. These methods generate primary component-metabolite correlation coefficients, enabling the distinction between pre- and post-administration groups. Using S-plots, significant MS signals can be extracted, identifying components with marked differences between samples before and after drug administration. This allows for the deduction of both prototype and metabolite components of NMs *in vivo*. For instance, GUO *et al.* applied UPLC-Q-TOF-MS to elucidate the prototypical components and metabolites of Shuanghuanglian injection (SHLI) in human serum, using OPLS-DA to distinguish drug-related exogenous components from endogenous ones (Fig. 5B). S-plots identified 35 drug-related components, including 23 prototype compounds and 12 metabolites, revealing that SHLI primarily induces phase II metabolic reactions in humans [40]. Similarly, Minale *et al*. used OPLS-DA and S-plots to screen metabolite changes in plasma, urine, and feces following administration of Brahmi essence from *Bacopa monnieri* (L.) Wettst. They identified 15, 7, and 17 metabolites in plasma, urine, and fecal samples, respectively, *via* online databases and literature [41]. WAN *et al*. conducted a systematic metabolomic analysis of Angelicae Pubescentis Radix (APR) in rats using UPLC-Q-TOF-MS, identifying 213 APR-related compounds, including 41 prototypes and 107 phase I/ 65 phase II metabolites, with 134 being potential novel compounds [42]. JIANG *et al*. analyzed the *in vivo* related components of the western drug nefazodone and the Glycyrrhizae Radix et Rhizoma (GC) *via* three nontargeted analytical methods: metabolomics, PATBS, and MDF. The metabolomics analysis utilized MZmine and MS-DIAL software for peak extraction, followed by volcano plot statistics to identify compounds that showed increased concentrations post-administration. These were then confirmed as potentially relevant components, allowing GC to be used to identify nefazodone-related components in rats [43]. Additionally, PATBS and MDF were employed to find the relevant in vivo components of the two drugs, revealing significant differences in coverage between the metabolomics and PATBS approaches, indicating that they offer complementary insights (Fig. 5C).

Mathematical algorithms (MAs) are also widely adopted in metabolomics analysis to identify and cluster target analytes, facilitating the discovery of numerous new heterogeneous components while considerably reducing labor and time. However, this process cannot distinguish endogenous components with significant differences, potentially leading to false-positive results. Additionally, the MA requires a sufficiently large number of samples and a large amount of work, which is more limited when applied to the search for relevant components in the actual NM components *in vivo*. To improve the quality and efficiency of data analysis and reduce the false-positive rate, ZHANG *et al*. optimized the settings of threshold parameters for peak extraction, parameters for statistical data analysis (mainly FC), and the sample concentration ratio during the application of metabolomics to establish a complete set of analytical strategies for the *in vivo* exposure profiling of nontargeted TCM (Fig. 5D). This strategy

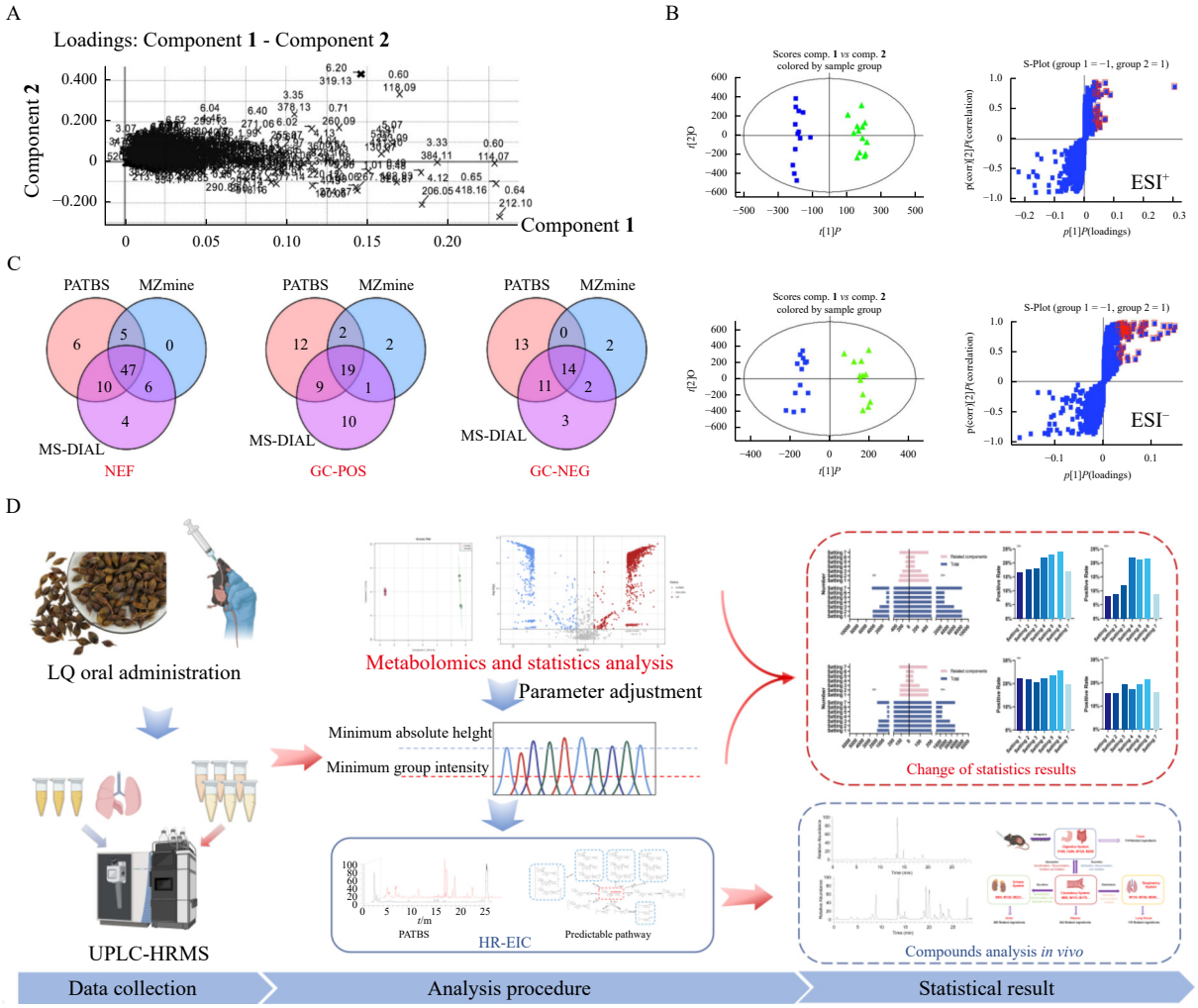

Fig. 5 The application of MA in the search for relevant substances in the *in vivo* analysis of NMs. (A) Resulting PCA scores plot from male control, low and high dose [39]. (B) S-plots of of human serum samples between Control and SHLI groups in positive ion mode and negative ion mode [40]. (C) Comparison of PATBS, MZmine, and MS-DIAL for untargeted profiling of unknown xenobiotics in plasma from NEF- and GC-dosed rats [43]. (D) Flowchart of NMs *in vivo* composition study based on metabolomics [44]. (A higher resolution/color version of this figure is available in the electronic copy of the article).

was applied to the identification of the relevant components of the TCM *Forsythia* in mice [44]. This study demonstrated that more peaks were extracted by the software when higher sample concentrations were used and when threshold parameters and fold change (FC) values were set lower. Under these conditions, metabolomics analysis detected an increased number of drug-related componentsallowing for a more comprehensive profile of the *in vivo* metabolism-related components of the drug. However, this high detection rate came with a drawback: a lower positive identification rate. In contrast, when lower sample concentrations were used, and threshold and FC values were increased, endogenous interference was reduced, resulting in a higher positive identification rate. Ultimately, this approach enabled a detailed elucidation of *Forsythia*'s *in vivo* metabolic processes and constructed a comprehensive metabolic profile of *Forsythia* in mice by using higher sample concentrations and lower threshold and FC values.

In summary, metabolomics technology provides an efficient and thorough method for identifying NM-related components *in vivo*, ensuring that even those present in low concentrations are not overlooked. However, the technique is not without limitations, including a high false-positive rate and challenges with computerized peak extraction from complex datasets.

We summarized the advantages and limitations of multi-template-targeted and nontargeted intelligent MS data processing techniques in Table 2.

Novel MS Data-processing Techniques Based on AI

In recent years, AI technology has emerged as a powerful tool for processing and analyzing NM MS data. AI algorithms can be used to denoise, enhance signals, and correct baseline offsets in raw MS data, significantly improving both data acquisition efficiency and the accuracy of data mining.

Table 2 Summary of the characteristics of multitemplate-targeted and nontargeted MS processing techniques in NMs *in vivo* analysis

Technology	Advantages	Limitations
MTSF	Efficient correlation determination between *in vivo* or *in vitro* compounds and their metabolic transformation; Quickly establish metabolic transformation relationships.	Discrimination and rejection of false-positive data. Metabolite with multi-step transformation will lead to a shortage of the ions observed in the templates, potentially yielding low-to-no similarity score; Building a template compound library is quite time-consuming.
MMDF	The technology is mature, and can be achieved through multiple commercial softwares. Compared to MDF, MMDF can uncover more target structural analogs	A larger quality loss filtering window will lead to a too broad screening range, resulting in false-positive results.
MF-MDFs	Simultaneously picking out molecular ions and characteristic fragment ions related to potential metabolites in a targeted mass defect range. Provides structure-enriched evidence (including molecular ion and fragment ion) to confirm the existence of and elucidate the structure of metabolites.	MF-MDFs is a self-developed technique in the laboratory and has not yet been made available for public use; Need template compounds.
BS	Comprehensive and non-targeted screen of potential drug-related components. The technology is mature, and can be achieved through multiple commercial software; No need for template compounds.	The differences in parameter settings may lead to incomplete subtraction or excessive subtraction.
PATBS	The background elimination is cleaner and does not lose the signal of low-concentration metabolites, allowing for the rapid and comprehensive discovery of the chemical components of traditional Chinese medicine entering the body and their metabolites; No need for template compounds.	PATBS is a self-developed technique in the laboratory and has not yet been made available for public use
Metabolomics	Capable of comprehensive, unbiased, high throughput analysis of samples, without discriminating against low concentrations of components, aiding in the discovery of new drug metabolites.	High false-positive rate; Statistical results will be influenced by various parameter settings, including peak extraction and screening parameters.

Convolutional Neural Networks (CNNs), autoencoders, and other neural network algorithms can learn features from MS signals, automatically removing noise and signal offsets, which enhances the quality and precision of raw MS signals. For example, Seddiki *et al*. demonstrated the use of CNNs for denoising MS data, showing that this method effectively reduces noise levels while improving the resolution and quantitative accuracy of MS signals [45]. Similarly, wavelet transform denoising autoencoders have been used to enhance MS signals by improving signal intensity and clarity. JIN *et al*. applied a deep learning-based denoising autoencoder to predict noise-corrected peaks from the original peak distribution, improving the accuracy of peak detection [30]. Another technique, Deep Learning Baseline Correction (DLBC), uses neural networks to automatically detect and remove baseline offsets in MS data. LIU *et al.* applied DLBC to MS data, showing that it effectively eliminates baseline offsets [46]. AI has also been used in practical examples of *in vivo* analysis. ZENG *et al*. developed a CNN to reduce false positives in peak detection using Continuous Wavelet Transform, which, when combined with computer prediction, improved the chemical coverage for targeted screening of in vivo metabolites *via* LC-HRMS fingerprinting [47]. This method was applied to identify and characterize *in vivo* metabolites after oral administration of Pericarpium Citri Reticulatae and Fructus Aurantii.

Furthermore, AI can integrate multimodal and multidimensional data, allowing the learning of compound features from extensive MS datasets. This capability enables AI to assist in the classification and identification of prototype compounds, metabolites, and active NM components *in vivo*. For instance, GUO *et al*. used a fully connected neural network to learn MDF features from MS signals, facilitating the intelligent classification of metabolites from QiangHuoShengShi decoction. This approach successfully identified 58 prototype compounds and 111 metabolites in rat plasma following oral administration of the decoction [31]. AI algorithms can also integrate multimodal data from MS spectra and pharmacological analyses, enabling the examination of correlations between NM component MS signals and biological activity. Using linear or nonlinear models, AI can identify potential bioactive molecules in NMs. OUYANG *et al*. used multiblock partial least squares, partial least squares, and support vector machine algorithms to explore the relationship between MS signals from carbonized Pollen Typhae and the hemostatic effects both *in vivo* and *in vitro* by using multiblock partial least squares, partial least squares, and support vector machine algorithms. They predicted and validated 9 key components responsible for the hemostatic effect of carbonized Pollen Typhae [48]. Paula *et al*. applied genetic search and decision tree algorithms to correlate MS signals with the anti-inflammatory properties of leaf extracts from 57 species of *Asteraceae*, identifying 11 natural compounds with potential anti-inflammatory effects [49].

All the abovementioned applications of HPLC-HRMS in NM qualitative analysis *in vivo* are summarized in Supplementary Table.

Discussion and Future Perspectives

The pharmacological substance basis of natural medicines (NMs) and their mechanisms of action have consistently been critical yet challenging areas of research, playing an essential role in advancing the modernization of NMs [17]. The natural components that enter the organism and their resulting metabolites are closely linked to the pharmacological effects of NMs. A thorough investigation of the *in vivo* processes of NMs can provide researchers with crucial insights into the exposure and metabolic transformations of NM components within the body, thereby improving the efficiency and reliability of active ingredient screening in NMs. By using the results of *in vivo* component analysis as targets for pharmacological screening, the reliability of selecting prototype drug components is enhanced, and the selection range is narrowed. However, the complexity of the resulting metabolites requires more efficient methods for screening pharmacologically active components in NMs. The integration of *in vivo* component analysis with rapid screening methods is evolving as an important tool in NM research. The comprehensive *in vivo* screening of NMs generally involves three main steps. First, intelligent MS technology is developed to establish a comprehensive system for *in vivo* analysis of NMs, encompassing data acquisition and postprocessing. This system allows for the nontargeted identification and characterization of NM components and the generation of metabolic profiles for major exposed compounds. Second, a rapid screening method for NM component activity is implemented, utilizing approaches such as biochromatography, computer-assisted techniques, and zebrafish models. These methods help establish correlations between in vivo components, active components, and targeted biological pathways, enabling the more efficient identification of potential pharmacological ingredients. Finally, the active components and their mechanisms of action are validated at both the cellular and animal levels (Fig. 6). A detailed explanation of this strategy and its application in screening active components in NMs is provided in the supplementary materials.

Upon *in vivo* administration, identifying pharmacologically active compounds in NMs presents several challenges, including diverse metabolic pathways, low metabolite concentrations, and significant interference from complex biological matrices or endogenous components. These factors complicate the detection and identification of NM components after administration. Currently, HRMS is the most commonly used method for *in vivo* analysis of NMs, relying heavily on advanced intelligent MS data processing techniques. The development of AI is creating new opportunities to enhance the *in vivo* analysis of NMs. Firstly, although NM compound structure identification still largely depends on manual analysis combined with network databases, AI can accelerate and improve the accuracy of component identification through machine learning and data mining techniques. Secondly, AI can aid in managing noise reduction and peak shift calibration in MS data, thus increasing the accuracy and interpretability of NM metabolite datasets. Thirdly, AI can assist in data analysis and pattern recognition, predict metabolic pathways, and identify potential active ingredients. When combined with *in vivo* analysis, AI can significantly improve the success rate of screening for bioactive NM components. Lastly, AI can integrate and analyze multisource, heterogeneous data, facilitating cross-domain knowledge discovery and deepening our understanding of NMs, thereby promoting their development and application.

MS-based nontargeted metabolomics has proven highly successful in providing comprehensive, unbiased, and quantitative characterizations of metabolites in a biological system. However, the vast chemical and compositional diversity of the metabolome poses significant challenges in separating and identifying metabolites. Ion mobility (IM) adds an additional dimension to LC-MS-based nontargeted metabolomics, enhancing coverage, sensitivity, and resolving power, particularly for metabolite isomers. However, the high dimensionality of IM-resolved metabolomics data presents processing challenges, limiting its widespread application. To address this issue, LUO *et al.* developed Met4DX, a computational tool designed for peak detection, quantification, and metabolite identification in 4D untargeted metabolomics. This tool advances metabolite discovery by deciphering complex 4D metabolomics data [50]. Given the complexity of NMs, which often contain multiple compounds that may co-elute or share similar fragments, nontargeted metabolomics with IM offers a powerful additional dimension for resolving and identifying NM-related compounds in biological systems.

Currently, *in vivo* NM research has concentrated on NMs that originate from plants. NMs derived from animals have a wide range of clinical applications with good activities and unique curative effects. While plant-derived NMs often contain small secondary metabolites that are easier to detect, animal-derived drugs are composed mainly of polysaccharides, amino acids, peptides, and proteins, many of which contribute to their therapeutic effects. However, these components are challenging to analyze *in vivo*. Polysaccharides, amino acids, peptides, and proteins are the main components of animal drugs, and many of them are the source of the unique efficacy of animal drugs, but many difficulties exist in the analysis and study of these components *in vivo*. For example, proteins and other biological macromolecules have large molecular weights and complex structures that are generally not metabolized by CYP450 enzymes. Compared with that of small-molecular compounds, the metabolic process of biological macromolecules is more complex. Current methods for the *in vivo* analysis of animal-derived NMs are limited to simple pharmacokinetic (PK) studies of known ingredients using chromatography-MS. Therefore, novel targeted analysis strategies are urgently needed to improve the *in vivo* analysis of animal-derived NMs and address the unique chal-

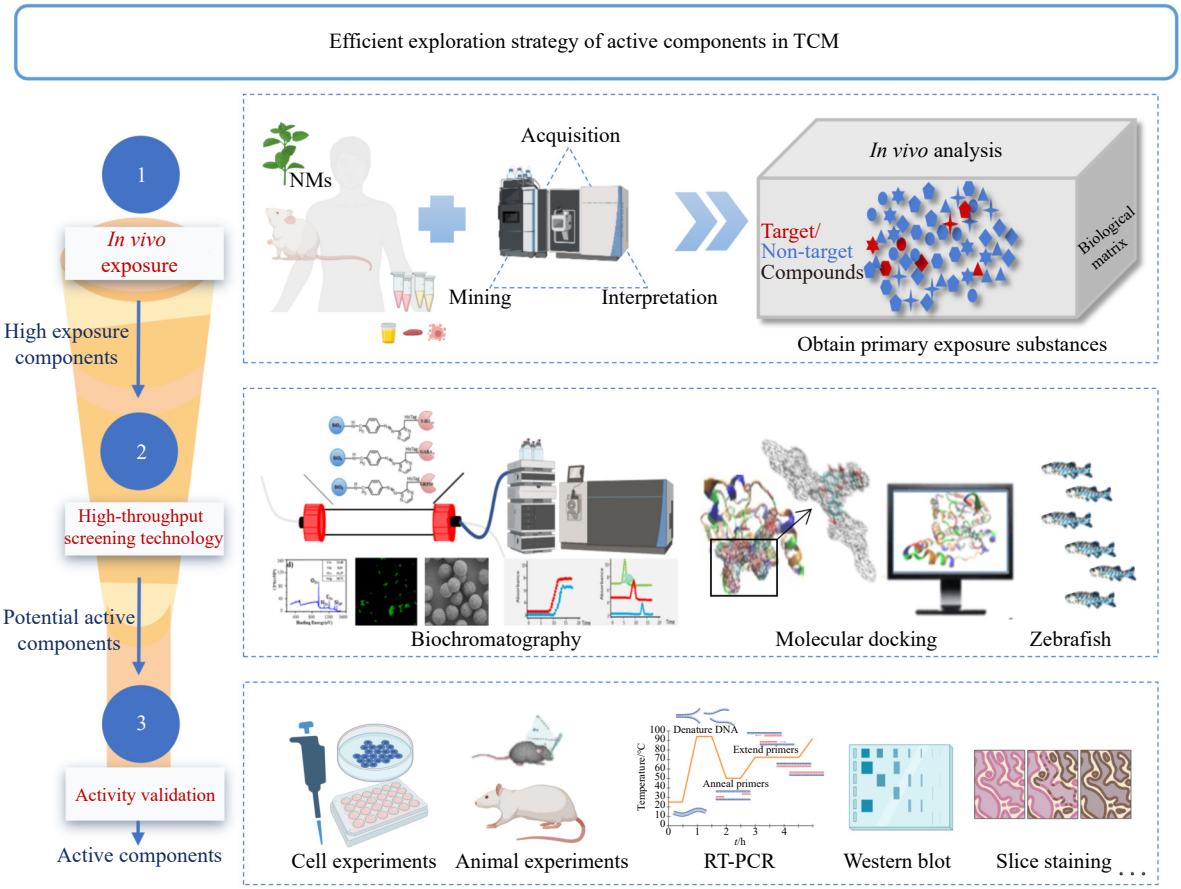

Fig. 6 A schematic diagram of a strategy for efficiently discovering active components in NMs by combining natural drug *in vivo* processes with rapid activity screening technology.

lenges posed by their complex compositions.

Conclusions

NMs have broad prospects and considerable clinical implications. However, their multiple components and targets, coupled with low affinity and selectivity, pose challenges for research and drug development. *In vivo* analysis of NMs using HPLC-HRMS technology can help identify active components in the blood, providing valuable insights into the pharmacological, toxicological, and mechanistic aspects of TCM. This approach offers strong support for the development of new drugs derived from TCM and enhances medication safety. The MS datasets generated by HPLC-HRMS are inherently complex, making the *in vivo* analysis of NMs highly dependent on the advancement of intelligent MS data processing techniques. This review focuses on the intelligent data processing methods used in NM research. For targeted approaches, techniques, such as EIC, MDF, IPF, PIF, NLF, and MTSF, are highlighted for their role in identifying NM components *in vivo*. For nontargeted technical approaches, the review emphasizes two key methods: BS and metabolomics. Additionally, this review introduces MS data processing techniques based on AI and discusses the current state and future directions of data postprocessing based on HPLC-HRMS. In summary, numerous effective data processing techniques for *in vivo* analysis of NMs have been developed to help fulfill the requirements of the *in vivo* analysis of NMs. However, due to the complexity of NMs and their diverse metabolic pathways *in vivo*, the development of more intelligent MS data processing techniques is essential for improving the detection and identification of NM-related components in biological matrices. This remains a crucial area of focus in NM research.

References

[1] Zhu MS, Zhang HY, Humphreys WG. Drug metabolite profiling and identification by high-resolution mass spectrometry [J]. *J Biol Chem* , 2011, **286**(29): 25419-25425.

[2] Yu Y, Yao C, Guo DA. Insight into chemical basis of traditional Chinese medicine based on the state-of-the-art techniques of liquid chromatography-mass spectrometry [J]. *Acta Pharm Sin B*, 2021, **11**(6): 1469-1492.

[3] Bai X, Zhu CY, Chen JY, *et al*. Recent progress on mass spectrum based approaches for absorption, distribution, metabolism, and excretion characterization of traditional Chinese medicine [J]. *Curr Drug Metab*, 2022, **23**(2): 99-112.

[4] Wen B, Zhu MS. Applications of mass spectrometry in drug metabolism: 50 years of progress [J]. *Drug Metab Rev*, 2015, **47**(1): 71-87.

[5] Wu HF, Guo J, Chen SL, *et al*. Recent developments in qualitative and quantitative analysis of phytochemical constituents and their metabolites using liquid chromatography-mass spectrometry [J]. *J Pharm Biomed Anal*, 2013, **72**: 267-291.

[6] Wang HP, Chen CY, Liu Y, *et al*. Metabolic profile of Yi-Xin-

Shu capsule in rat by ultra-performance liquid chromatography coupled with quadrupole time-of-flight tandem mass spectrometry analysis. [J]. *RSC Adv*, 2015, **5**(98): 80583-80590.
[7] Zhao WJ, Shang ZP, Li QQ, *et al*. Rapid screening and identification of daidzein metabolites in rats based on UHPLC-LTQ-Orbitrap mass spectrometry coupled with data-mining technologies [J]. *Molecules*, 2018, **23**(1): 151.
[8] Zhu HD, Wu XD, Huo JY, *et al*. A five-dimensional data collection strategy for multicomponent discovery and characterization in traditional Chinese medicine: Gastrodia Rhizoma as a case study [J]. *J Chromatogr A*, 2021, **1653**: 462405.
[9] Zuo R, Wang HJ, Si N, *et al*. LC-FT-ICR-MS analysis of the prototypes and metabolites in rat plasma after administration of Huang-Lian-Jie-Du decoction [J]. *Acta Pharm Sin B*, 2014, **49**(02): 237-243.
[10] Qiao X, Wang Q, Wang S, *et al*. Compound to extract to formulation: a knowledge-transmitting approach for metabolites identification of Gegen-Qinlian decoction, a traditional Chinese medicine formula [J]. *Sci Rep*, 2016, **6**(1): 39534.
[11] Yao M, Ma L, Duchoslav E, *et al*. Rapid screening and characterization of drug metabolites using multiple ion monitoring dependent product ion scan and postacquisition data mining on a hybrid triple quadrupole-linear ion trap mass spectrometer [J]. *Rapid Commun Mass Sp*, 2009, **23**(11): 1683-1693.
[12] Huang ML, Cheng ZZ, Wang L, *et al*. A targeted strategy to identify untargeted metabolites from *in vitro* to *in vivo*: rapid and sensitive metabolites profiling of licorice in rats using ultra-high performance liquid chromatography coupled with triple quadrupole-linear ion trap mass spectrometry [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2018, **1092**: 40-50.
[13] Ni S, Qian D, Duan JA, *et al*. UPLC-QTOF/MS-based screening and identification of the constituents and their metabolites in rat plasma and urine after oral administration of *Glechoma longituba* extract [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2010, **878**(28): 2741-2750.
[14] Shi T, Yao ZH, Qin ZF, *et al*. Identification of absorbed constituents and metabolites in rat plasma after oral administration of Shen-Song-Yang-Xin using ultra-performance liquid chromatography combined with quadrupole time-of-flight mass spectrometry [J]. *Biomed Chromatogr*, 2015, **29**(9): 1440-1452.
[15] Tao JH, Zhao M, Wang DG, *et al*. UPLC-Q-TOF/MS-based screening and identification of two major bioactive components and their metabolites in normal and CKD rat plasma, urine and feces after oral administration of *Rehmannia glutinosa* Libosch extract [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2015, **1001**: 98-106.
[16] 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 [J]. *Anal Chem*, 2008, **80**(21): 8187-8194.
[17] Jin Y, Wu CS, Zhang JL, *et al*. A new strategy for the discovery of epimedium metabolites using high-performance liquid chromatography with high resolution mass spectrometry [J]. *Anal Chim Acta*, 2013, **768**: 111-117.
[18] 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 Fructus Gardeniae-Fructus Forsythiae herb pair extract and in rat's blood [J]. *J Chromatogr A*, 2018, **1563**: 99-123.
[19] Tian TT, Jin YR, Ma YH, *et al*. Identification of metabolites of oridonin in rats with a single run on UPLC-Triple-TOF-MS/MS system based on multiple mass defect filter data acquisition and multiple data processing techniques [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2015, **1006**: 80-92.
[20] Liu MY, Zhao SH, Wang ZQ, *et al*. Identification of metabolites of deoxyschizandrin in rats by UPLC–Q-TOF-MS/MS based on multiple mass defect filter data acquisition and multiple data processing techniques [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2014, **949-950**: 115-126.
[21] Lai HQ, Ouyang Y, Tian GH, *et al*. Rapid characterization and identification of the chemical constituents and the metabolites of Du-zhi pill using UHPLC coupled with quadrupole time-of-flight mass spectrometry [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2022, **1209**: 123433.
[22] Feng W, Dong QJ, Liu MY, *et al*. Screening and identification of multiple constituents and their metabolites of Zhi-zi-chi decoction in rat urine and bile by ultra-high-performance liquid chromatography quadrupole time-of-flight mass spectrometry [J]. *Biomed Chromatogr*, 2017, **31**(10): 3978.
[23] Zhang JY, Cai W, Zhou Y, *et al*. Profiling and identification of the metabolites of baicalin and study on their tissue distribution in rats by ultra-high-performance liquid chromatography with linear ion trap-Orbitrap mass spectrometer [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2015, **985**: 91-102.
[24] Xie C, Zhong DF, Yu K, *et al*. Recent advances in metabolite identification and quantitative bioanalysis by LC-Q-TOF MS [J]. *Bioanalysis*, 2012, **4**(8): 937-959.
[25] Gao Y, Zhang RP, Bai JF, *et al*. Targeted data-independent acquisition and mining strategy for trace drug metabolite identification using liquid chromatography coupled with tandem mass spectrometry [J]. *Anal Chem*, 2015, **87**(15): 7535-7539.
[26] Wang RH, Yin YD, Zhu ZJ. Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology [J]. *Anal Bioanal Chem*, 2019, **411**(19): 4349-4357.
[27] Zha HH, Cai YP, Yin YD, *et al*. SWATHtoMRM: development of high-coverage targeted metabolomics method using SWATH technology for biomarker discovery [J]. *Anal Chem*, 2018, **90**(6): 4062-4070.
[28] Guo J, Huan T. Evaluation of significant features discovered from different data acquisition modes in mass spectrometry-based untargeted metabolomics [J]. *Anal Chim Acta*, 2020, **1137**: 37-46.
[29] Chen JY, Ding R, Jin Y, *et al*. HRPIF data mining based on data-dependent/independent acquisition for Rhei Radix et Rhizoma metabolite screening in rats [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2022, **1190**: 123095.
[30] Zhu CY, Cai TT, JIN Y, *et al*. Artificial intelligence and network pharmacology based investigation of pharmacological mechanism and substance basis of Xiaokewan in treating diabetes [J]. *Pharmacol Res*, 2020, **159**: 104935.
[31] Guo JD, Shang Y, Yang XH, *et al*. An online stepwise background subtraction-based ultra-high pressure liquid chromatography quadrupole time of flight tandem mass spectrometry dynamic detection integrated with metabolic molecular network strategy for intelligent characterization of the absorbed chemical-fingerprint of QiangHuoShengShi decoction *in vivo* [J]. *J Chromatogr A*, 2022, **1675**: 463172.
[32] Zhang X, Liao M, Cheng XY, *et al*. Ultrahigh - performance liquid chromatography coupled with triple quadrupole and time-of-flight mass spectrometry for the screening and identification of the main flavonoids and their metabolites in rats after oral administration of *Cirsium japonicum* DC. extract [J]. *Rapid Commun Mass Sp*, 2018, **32**(16): 1451-1461.
[33] Zhang HY, Yang YO. An algorithm for thorough background subtraction from high-resolution LC/MS data: application for detection of glutathione-trapped reactive metabolites [J]. *J Mass Spectrom*, 2008, **43**(9): 1181-1190.

[34] Zhang HY, Ma L, He K, *et al*. An algorithm for thorough background subtraction from high-resolution LC/MS data: application to the detection of troglitazone metabolites in rat plasma, bile, and urine [J]. *J Mass Spectrom*, 2008, **43**(9): 1191-200.
[35] Zhu PJ, Ding W, Tong W, *et al*. A retention-time-shift-tolerant background subtraction and noise reduction algorithm (BgS-NoRA) for extraction of drug metabolites in liquid chromatography/mass spectrometry data from biological matrices [J]. *Rapid Commun Mass Sp*, 2009, **23**(11): 1563-1572.
[36] Shekar V, Shah A, Shadid M, *et al*. An accelerated background subtraction algorithm for processing high-resolution MS data and its application to metabolite identification [J]. *Bioanalysis*, 2016, **8**(16): 1693-1707.
[37] Wu CS, Zhang HY, Wang CH, *et al*. An integrated approach for studying exposure, metabolism, and disposition of multiple component herbal medicines using high-resolution mass spectrometry and multiple data processing tools [J]. *Drug Metab Dispos*, 2016, **44**(6): 800-808.
[38] Zhu CY, Jiang XJ, Tian JJ, *et al*. Integrated approach toward absorption, distribution, metabolism, and excretion of Xiaoke pills in zebrafish based on UPLC-HRMS and DESI-MS techniques [J]. *J Chromatogr B Analyt Technol Biomed Life Sci*, 2022, **1200**: 123276.
[39] Plumb RS, Stumpf CL, Granger JH, *et al*. Use of liquid chromatography/time-of-flight mass spectrometry and multivariate statistical analysis shows promise for the detection of drug metabolites in biological fluids [J]. *Rapid Commun Mass Sp*, 2003, **17**(23): 2632-2638.
[40] Guo MX, Zhao BS, Liu HY, *et al*. A metabolomic strategy to screen the prototype components and metabolites of Shuang-Huang-Lian Injection in human serum by ultra performance liquid chromatography coupled with quadrupole time-of-flight mass spectrometry [J]. *J Anal Methods Chem*, 2014, **2014**: 241505.
[41] Minale G, Seasong T, Temkitthawon P, *et al*. Characterization of metabolites in plasma, urine and feces of healthy participants after taking brahmi essence for twelve weeks using LC-ESI-QTOF-MS metabolomic approach [J]. *Molecules*, 2021, **26**(10): 2944.
[42] Wan MQ, Liu XY, Gao H, *et al*. Systematic analysis of the metabolites of Angelicae Pubescentis Radix by UPLC-Q-TOF-MS combined with metabonomics approaches after oral administration to rats [J]. *J Pharm Biomed Anal*, 2020, **188**: 113445.
[43] Jiang XJ, Chen SM, Zhu MS, *et al*. Global xenobiotic profiling of rat plasma using untargeted metabolomics and background subtraction-based approaches: method evaluation and comparison [J]. *Curr Drug Metab*, 2023, **24**(3): 200-210.
[44] Zhang HR, Jiang XJ, Zhang DD, *et al*. An integrated approach for studying the metabolic profiling of herbal medicine in mice using high-resolution mass spectrometry and metabolomics data processing tools [J]. *J Chromatogr A*, 2024, **1713**: 464505.
[45] Seddiki K, Saudemont P, Precioso F, *et al*. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification [J]. *Nat Commun*, 2020, **11**(1): 5595.
[46] Luo Y, Liu Y, Wen Q, *et al*. Comprehensive chemical and metabolic profiling of anti-hyperglycemic active fraction from Clerodendranthi Spicati Herba [J]. *J Sep Sci*, 2021, **44**(9): 1805-1814.
[47] Zeng J, Li YP, Wang CL, *et al*. Combination of *in silico* prediction and convolutional neural network framework for targeted screening of metabolites from LC-HRMS fingerprints: a case study of "Pericarpium Citri Reticulatae-Fructus Aurantii" [J]. *Talanta*, 2024, **269**: 125514.
[48] Ouyang XJ, Li JQ, Zhong YQ, *et al*. Identifying the active ingredients of carbonized Typhae Pollen by spectrum-effect relationship combined with MBPLS, PLS, and SVM algorithms [J]. *J Pharm Biomed Anal*, 2023, **235**: 115619.
[49] Chagas-Paula DA, Oliveira TB, Zhang T, *et al*. Prediction of anti-inflammatory plants and discovery of their biomarkers by machine learning algorithms and metabolomic studies [J]. *Planta Med* , 2015, **81**(6): 450-458.
[50] Luo MD, Yin YD, Zhou ZW, *et al*. A mass spectrum-oriented computational method for ion mobility-resolved untargeted metabolomics [J]. *Nat Commun*, 2023, **14**(1): 1813.

Cite this article as: CHEN Simian, DAI Binxin, ZHANG Dandan, *et al*. Advances in intelligent mass spectrometry data processing technology for *in vivo* analysis of natural medicines [J]. *Chin J Nat Med*, 2024, **22**(10): 900-913.

Prof. WU Caisheng was awarded the title of Professorship under the National Excellent Young Scientists Fund in 2022, at School of Pharmaceutic Sciences, Xiamen University. He mainly focuses on bioanalysis of drugs in complex *in vivo* systems, the demystification of traditional Chinese medicine with *in vivo* metabolic investigation, as well as the identification of clinical diagnostic or therapeutic biomarkers. Prof. WU has published over 80 peer-reviewed research articles in journals including *Adv Sci*, *Acta Pharm Sin B*, and *Anal Chem*, *et al.*, and serving as a Youth Editorial Board Member for journals including *Chin J Nat Med*, *Acta Pharm Sin B*, *J Pharm Anal*, etc.