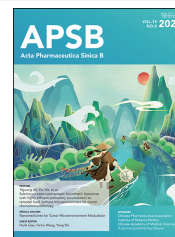




Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

Unveiling the metabolic fate of drugs through metabolic reaction-based molecular networking



Haodong Zhu^{a,b,†}, Xupeng Tong^{c,†}, Qi Wang^{a,†}, Aijing Li^a,
Zubao Wu^a, Qiqi Wang^a, Pei Lin^a, Xinsheng Yao^a, Liufang Hu^{a,d,*},
Liangliang He^{a,*}, Zhihong Yao^{a,*}

^aState Key Laboratory of Bioactive Molecules and Druggability Assessment; International Cooperative Laboratory of Traditional Chinese Medicine Modernization and Innovative Drug Development of Ministry of Education (MOE) of China; Guangdong Basic Research Center of Excellence for Natural Bioactive Molecules and Discovery of Innovative Drugs; Guangdong Province Key Laboratory of Pharmacodynamic Constituents of TCM and New Drugs Research, College of Pharmacy, Jinan University, Guangzhou 510632, China

^bGuangxi Key Laboratory of Special Biomedicine; School of Medicine, Guangxi University, Nanning 530004, China

^cHangzhou Chenfeng Qingxing Technology Co., Ltd., Hangzhou 310000, China

^dCollege of Chemical Engineering, Xiangtan University, Xiangtan 411105, China

Received 24 November 2024; received in revised form 10 March 2025; accepted 1 April 2025

KEY WORDS

Drug metabolism;
MRMN;
LC–MS;
MS² feature degradation
improvement;
Redundant ions
identification;
Endogenous interference
elimination;

Abstract Effective annotation of *in vivo* drug metabolites using liquid chromatography-mass spectrometry (LC–MS) remains a formidable challenge. Herein, a metabolic reaction-based molecular networking (MRMN) strategy is introduced, which enables the “one-pot” discovery of prototype drugs and their metabolites. MRMN constructs networks by matching metabolic reactions and evaluating MS² spectral similarity, incorporating innovations and improvements in feature degradation of MS² spectra, exclusion of endogenous interference, and recognition of redundant nodes. A minimum 75% correlation between structural similarity and MS² similarity of neighboring metabolites was ensured, mitigating false negatives due to spectral feature degradation. At least 79% of nodes, 49% of edges, and 97% of subnetworks were reduced by an exclusion strategy of endogenous ions compared to the Global Natural Products Social Molecular Networking (GNPS) platform. Furthermore, an approach of redundant ions identification was refined,

*Corresponding authors.

E-mail addresses: yaozhong_jnu@163.com, yaozhong.jnu@gmail.com (Zhihong Yao), heliangliang5878@163.com (Liangliang He), huliufang0404@163.com (Liufang Hu).

[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

<https://doi.org/10.1016/j.apsb.2025.03.050>

2211-3835 © 2025 The Authors. Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

YaoLab online platform;
Prototypes and
metabolites “one-pot”
annotation

achieving a 10%–40% recognition rate across different samples. The effectiveness of MRMN was validated through a single compound, plant extract, and mixtures of multiple plant extracts. Notably, MRMN is freely accessible online at <https://yaolab.network>, broadening its applications.

© 2025 The Authors. Published by Elsevier B.V. on behalf of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Drug metabolism is a crucial biochemical process through which drugs are chemically altered by enzymatic reactions^{1,2}. For approximately 75% of all drugs, metabolism is a primary clearance pathway^{3,4} and plays a pivotal role in the activation, inactivation, toxification, and detoxification of small molecules, which significantly influences their efficacy and safety². Especially, the FDA requires further research on drug metabolites if they account for more than 10% of total drug-related exposure in humans at a steady state⁵. Thus, central to this endeavor is the characterization of *in vivo* prototypes and their metabolites⁶.

Nowadays, liquid chromatography coupled with tandem mass spectrometry (LC–MS) is widely regarded as a premier technique for studying drug metabolism due to its high sensitivity capabilities for structural elucidation and metabolite quantification². Traditional strategies for the identification of drug metabolites, such as library matching⁷, mass defect filtering (MDF)⁸, neutral loss filtering⁹, isotope filtering¹⁰, and background subtraction¹¹, have been efficiently utilized in drug metabolism studies for decades. However, these methods are significantly limited by the fact that over 90% of known metabolites in databases like the Human Metabolome Database (HMDB) and METLIN lack standard MS² spectra, complicating accurate identification and annotation, especially for those without readily available chemical standards. Moreover, large-scale metabolite annotation in LC–MS is hindered by high false positive and negative rates due to the complexity of biological matrices and metabolic processes, requiring extensive and time-consuming manual intervention¹².

Molecular networking (MN) is an efficient method for annotating compounds in LC–MS data. This innovative technique enables the grouping of similar MS/MS spectra into clusters, where each node represents a distinct MS/MS spectrum. The structural similarity between *in vivo* drug prototypes and their metabolites provides the feasibility for applying MN techniques in drug metabolites discovery and annotation^{13–17}. Nonetheless, MNs grapple with persistent challenges in the analysis of *in vivo* metabolite annotation, including: 1) the need for more accurate discovery of metabolite-related ions for robust analysis; 2) the importance of edge resolution between known and unknown components for precise compound annotation within MNs; 3) the decline in MS² spectral similarity between metabolites and prototypes as prototypes undergo metabolic transformations, leading to a disproportionate increase in false negatives^{13,16,18,19}; 4) The inherent complexities of LC–MS data, encompassing various types of adduct ions, in-source fragmentation, and co-eluting components, exacerbate challenges in exogenous substance identification, culminating in elevated false positive rates^{20–22}. To tackle these challenges, we introduce metabolic reaction-based molecular networking (MRMN) for comprehensive metabolic profiling of drugs. MRMN leverages metabolic reactions between prototypes and their related metabolites to construct MNs, thereby

enabling streamlined “one-pot” discovery of absorbable components and related metabolites. We demonstrate the robust capabilities of MRMN through the accurate discovery of metabolite-related ions, extensive metabolite exploration, and redundant ion identification across a single compound, plant extract, and a mixture of multiple plant extracts. Additionally, we have established a website (<https://yaolab.network>) to provide free online access to MRMN analysis.

2. Materials and methods

2.1. Chemicals and reagents

Isosilybin (purity $\geq 98\%$) was purchased from Chengdu Chroma-Biotechnology Co., Ltd. (Chengdu, China) and the structures were characterized by ¹H NMR, ¹³C NMR, and MS spectrum analysis before the sale. Commercially available *Astragali Radix* (AR), *Saposhnikovia Radix* (SR), and *Atractylodis Macrocephalae Rhizoma* (AMR) were purchased from Tong Ren Tang Pharmacy Co., Ltd. (Guangzhou, China). The voucher samples were stored at the Institute of Traditional Chinese Medicine and Natural Products of Jinan University. The carboxymethyl cellulose (CMC, purity $\geq 98\%$) was purchased from Shanghai Macklin Biochemical Technology Co., Ltd. (Shanghai, China). Ultra-pure water was purchased by A.S. Watson Group (Hong Kong, China). LC–MS grade acetonitrile, methanol, and formic acid were purchased from Fisher Scientific Co., Ltd. (Fair Lawn, USA). Other chemicals and materials were all analytical grade.

2.2. Sample preparation

The metabolite characterization study was applied through oral administration of a single compound, plant extract, and mixtures of multiple plant extracts, respectively. For one batch of preparation of multiple plant extract mixtures, a combination of AR (60 g), SR (20 g), and AMR (20 g) was extracted twice with 600 mL of deionized water under heating reflux for 1.5 h each time. The extracts were collected and freeze-dried to yield 40 g of traditional Chinese medicine of Yupingfengsan (YPFS), which was dissolved in deionized water prior to administration. Additionally, for plant extract preparation, 100 g of fresh ginger was juiced, and filtered through a 75 μ m mesh, and the filtrate was freeze-dried to obtain 3.8 g of ginger extract powder, also to be dissolved in deionized water before administration. Isosilybin was prepared as a suspension using a CMC aqueous solution prior to administration.

2.3. Experimental animals and administration protocol

Six-week-old male BALB/c mice of SPF level were obtained from Beijing Vital River Laboratory Animal Technology Co., Ltd.

(Beijing, China). The mice were allowed to acclimate for one week before the experiment. All mice were housed at an ambient temperature 23 ± 2 °C, humidity $60 \pm 5\%$, and a 12 h light/dark cycle. Then, they were randomly divided into four groups: single compound (isosilybin) group, plant extract group (fresh ginger), mixtures of multiple plant extracts (YPFS) group, and blank group. Thereafter, these animals were fasted with free access to water in stainless steel metabolic cages separately overnight. The treatment groups received daily doses of isosilybin (50 mg/kg/day), fresh ginger extract (500 mg/kg/day), and YPFS extract (3 g/kg/day) for a duration of 3 Days, administered in two divided doses per day with 0.3 mL per administration. Water was orally administrated to the blank group ($n = 4$) in the same way. All experimental procedures were executed according to the protocols approved by the Guide for the Care and Use of Laboratory Animals of Jinan University.

2.4. Collection and preprocessing methods for biological samples

After the final administration, experimental mice were anesthetized using gas anesthesia. Following the final administration, blood samples at 0.5, 1, 2, and 4 h were obtained from the orbital from all groups, and immediately transferred into heparinized tubes, respectively. Then, plasma was obtained by centrifugation at 12,000 rpm for 10.0 min, and plasma from each time point was taken and mixed to produce the pooled plasma of each group, respectively. Similarly, three mice from both the treat groups and the blank group were individually placed in metabolic cages to collect urine and feces from 0 to 8 h. Urine samples were centrifuged at 14,000 rpm for 10 min to remove impurities (Sorvall Legend Micro 17R Centrifuge, Thermo Scientific). All samples were subsequently stored at -80 °C before use.

For the pooled plasma sample of each group, three aliquots of 1 mL were drawn and processed. Each aliquot was mixed with 3 mL of a methanol-acetonitrile solution (v/v, 1:1), vortexed for 3 min, and centrifuged at 14,000 rpm for 10 min at 4 °C (Sorvall Legend Micro 17R Centrifuge, Thermo Scientific). The supernatant was transferred to clean EP tubes, dried under nitrogen gas at room temperature, and the residue was reconstituted with 200 μ L of methanol. After another centrifugation (14,000 rpm, 20 min, 4 °C, Sorvall Legend Micro 17R Centrifuge, Thermo Scientific), the supernatant was collected for LC–MS analysis.

Urine samples were thawed on ice and subjected to solid-phase extraction (SPE). The columns were first activated with pure methanol, equilibrated with 50% methanol, and then pure water. Subsequently, 3 mL of the urine sample was loaded onto the column, and the eluent was processed accordingly. The eluate obtained from the pure methanol wash was collected into clean EP tubes, dried under nitrogen gas, and the residue was reconstituted with 500 μ L of methanol. After centrifugation (14,000 rpm, 20 min, 4 °C, Sorvall Legend Micro 17R Centrifuge, Thermo Scientific), the supernatant was prepared for LC–MS analysis.

Fecal samples were air-dried, ground, soaked in methanol (10-fold weight-to-volume ratio) for 24 h, and sonicated for 1 h. The extracted liquid was then centrifuged at 14,000 rpm for 10 min, and 3 mL of the supernatant was transferred to clean EP tubes. The samples were dried under nitrogen gas, and the residue was reconstituted with 1 mL of methanol. After centrifugation (14,000 rpm, 20 min, 4 °C, Sorvall Legend Micro 17R Centrifuge, Thermo Scientific), the supernatant was passed through an SPE column to collect the pure methanol solution, which was dried

under nitrogen gas and reconstituted with methanol for LC–MS analysis.

2.5. Data acquisition of UHPLC/Q-exactive-orbitrap–MS

Waters ACQUITY™ HSS T3 column (2.1 mm $\times$ 50 mm, 1.8 μ m) equipped with Waters VanGuard™ pre-column (1.8 μ m, 2.1 mm $\times$ 5 mm) was installed a U3000 UHPLC system (Thermo Scientific Co., Ltd., MA, USA), column temperature 45 °C, single needle injection volume 2 μ L, mobile phase consisting of 0.1% formic acid-water (A) and 0.1% formic acid-acetonitrile (B), flow rate set to 0.3 mL/min, elution gradient: 0–2 min, 2%–5% (B); 2–3 min, 5%–11% (B); 3–12 min, 11%–40% (B); 12–15 min, 40%–60% (B); 15–17 min, 60%–100% (B); 17–18 min, 100% (B); 18–18.5 min, 100%–2% (B); 18.5–20 min, 2% (B).

An ESI ion source was employed by a Q-exactive-orbitrap mass spectrometer (Q-exactive-orbitrap–MS) (Thermo Scientific Co., Ltd., MA, USA), positive ion mode, spray voltage (+): 3.5 kV, capillary temperature 320 °C, sheath gas pressure 35 arb, auxiliary gas pressure 35 arb, solvent evaporation temperature 400 °C, ion scanning mode: Full MS-DD-MS², MS¹ scanning range m/z 50–1050, MS² scanning range m/z 70–1200, DDA data acquisition mode for intensity-dependent TOP5, collecting MS² information of precursor ions responding to TOP five in the scanning window. Xcalibur software (version 4.1, Thermo Scientific Co., Ltd., MA, USA) was used for instrument management and data acquisition.

2.6. Data preprocessing

All LC–MS data were imported into MS-DIAL software (version 4.9, CompMS, Tokyo, Japan) for raw data deconvolution, peak detection, peak alignment, and export of MS¹ peak list files and MS² MGF files for alignment results. Key parameters were set as follows: retention time deviation of 0.1 min, m/z deviation of 0.015 Da, peak detection threshold of 500,000, and retention of MS² information relative to the corresponding threshold of 3%. After alignment, the MS¹ peak list file was exported as a txt file with the “Raw data matrix (Height)” option, and the MGF file was exported as an MGF format file with the “GNPS export” option. The blank filter option should not be selected to avoid incorrect ion ID matching between the MS¹ and MS² files, which could lead to errors in subsequent algorithms.

2.7. Structural similarity evaluation

The similarity algorithm based on molecular fingerprints is one of the most important algorithms for evaluating the structural similarity of two compounds. The Morgan & Dice algorithm has been shown to be one of the algorithms with the best correlation between structural similarity calculation results and the similarity of corresponding MS² spectra of compounds. Therefore, the Morgan & Dice algorithm was chosen to evaluate the structural similarity between drug prototypes and their metabolites. 1) Convert the compound's mol format data to text-based SMILES format; 2) Utilize RDKit Python scripts (available at <https://www.rdkit.org/docs/index.html>) to convert compound SMILES structures to Morgan molecular fingerprint patterns²³; 3) Use the Dice molecular fingerprint similarity algorithm function in RDKit Python scripts to calculate the molecular fingerprint similarity between compounds to represent their structural similarity.

2.8. Construction of metabolic reaction database

The MRMN relies on edge generation based on metabolic reaction matching. Constructing a convincing metabolic reaction database is significant for the comprehensive discovery of metabolic reaction pairs. First, we obtained 8479 authentic metabolic reaction pairs from the KEGG database²⁴, calculated the molecular weight changes and molecular formula differences between substrates and products in all metabolic reaction pairs, and organized these metabolic reactions into a foundational metabolic reaction database²⁵. Subsequently, through literature searches and manual screening, we compiled a database of common metabolic reactions, including those frequently found in KEGG and related to natural products. The final database includes 115 types of metabolic reactions, along with their corresponding IDs, reaction descriptions, molecular weight changes, and molecular formula differences, as detailed in Supporting Information Table S1. To enhance the database's coverage, we developed a user-customized metabolic reaction module on an online platform, allowing users to input metabolic reaction descriptions and their corresponding molecular weight changes. Ultimately, we constructed a database combining 115 common metabolic reactions with user-customized metabolic reactions, ensuring the reliability of the metabolic reaction database.

We also provided access to an expanded database of 1536 metabolic reactions that cover the vast majority of known metabolic reaction types from KEGG. Additionally, we have enriched the expanded database with information on the potential metabolic pathways and associated enzymes for each reaction (updated in MRMN at YaoLab with a downloadable link). This comprehensive database enhances the interpretability of edges in MRMN's molecular networks. This data is accessible when clicking on edges in the molecular network, offering valuable references for edge annotations.

2.9. MS² spectral similarity evaluation

The most mainstream algorithm for evaluating the similarity between MS² spectra is the cosine similarity algorithm, which vectorizes the compound MS² spectra and evaluates their similarity based on the cosine similarity of MS² vectors. Although there are many other algorithms, such as the dot product algorithm, Spec2Vec algorithm based on machine learning, and MS² DeepScore algorithm based on deep learning^{26,27}, it has been found that these algorithms tend to suffer from either insufficient detection capability or excessive redundancy when calculating MS² spectra of exogenous-related ions in complex matrices. Therefore, this chapter employs three main cosine similarity algorithms from the matchms package for calculating MS² spectral similarity²⁸. The specific algorithms involved are as follows: 1) Cosine greedy algorithm (CosGre): this algorithm vectorizes the *m/z* and intensity values of MS² spectra and calculates their cosine value as the similarity value; 2) Neutral loss cosine algorithm (NLCos): considering that compounds of the same structural type may exhibit certain differences in MS² spectra, but their MS² fragmentation follows similar patterns, the neutral loss patterns are relatively close. Therefore, this algorithm utilizes the $\Delta m/z$ values between fragment ions in the MS² spectra and the average response of the two ions involved in the calculation to construct a data matrix. It then calculates the similarity of NL spectra between two MS² spectra, which is beneficial for focusing on smaller structural changes in similar compounds; 3) Modified cosine

algorithm (ModCos): this algorithm integrates the CosGre and NLCos algorithms. It considers both the direct similarity of MS² spectra and the decrease in MS² spectral similarity caused by overall MS² offset.

2.10. Construction of MRMN analysis platform

This study encapsulated the relevant innovative algorithms and original scripts of MRMN into a comprehensive package to enhance the usability and broader application of MRMN. An open-source web analysis platform (<http://www.yaolab.network>) was developed on GitHub, offering an interactive analysis workflow and providing detailed analysis tutorials.

2.11. Feature-based molecular networking analysis

First, the LC–MS raw data were preprocessed using MS-DIAL software (CompMS) to obtain MS¹ peak list files and MS² MGF files (took about 30 min)²⁹. The processed files were then imported into GNPS, the MS² similarity threshold was set to 0.7, and the rest of the parameters were set to default to construct the feature-based molecular networking (FBMN) (took about 3 h)³⁰. Finally, the constructed MN file was imported into Cytoscape software (V3.8.0) for subnetwork screening and manual annotation of isosilybin metabolites (took 15 min). Ion identity molecular networking (IIMN) analysis was performed based on the above FBMN process according to its official tutorial to compare the redundant ion recognition capabilities of IIMN and MRMN^{22,30}.

2.12. MetID analysis

The LC–MS raw data were directly imported into Compound Discoverer 3.3 software, and the “Expected and Unknown Met ID Workflow with MMDF” for isosilybin metabolites were analyzed. Taking isosilybin as the prototype compound, the relevant exogenous ions were filtered based on MMDF, and the built-in metabolite prediction and FISH Scoring were used to judge the metabolites. All parameters were set to default values.

3. Results

3.1. Edge generation based on high information entropy of metabolic reactions

The cornerstone of MN construction lies in generating edges by evaluating the relationships between nodes. Traditional MNs predominantly rely on MS² spectral similarity for edge generation, a measure fundamentally linked to chemical structural similarity¹⁷. However, this approach falters, as only about 50% of adjacent metabolites exhibit MS² spectral similarity greater than 0.5¹⁴, leading to inevitable misannotations and the need for laborious, supplementary interpretation of many edges, particularly in drug metabolite analysis. The structural correlation between substrates and products in metabolic reactions is primarily driven by enzymatic processes. Drugs and endogenous substances share the same metabolic enzyme systems *in vivo*. Consequently, within the complex chemical background of biological matrices, a relatively constrained spectrum of metabolic reactions more accurately reflects drug metabolism compared to the diverse and unknown substrates and products. Here, we propose an extension of the

edge generation mechanism beyond MS² spectral similarity to encompass metabolic reaction matching.

To enhance the annotation of drug metabolism reaction pairs, it is essential to recognize the potential for metabolic reactions to contain richer informational entropy. Our analysis of known metabolic reaction pairs (8479 pairs and 5651 metabolites) from the KEGG database revealed that the distribution of changes in molecular formula and mass differences caused by metabolic reactions was neither uniform nor random (Supporting Information Fig. S1)²⁴. Certain mass differences occurred far more frequently than others (Fig. 1A), with only 26 types appearing 30 times or more. We then ranked the top 16 mass differences most associated with natural products from these 26 types and identified their corresponding molecular formula changes (Fig. 1B). Notably, the $\Delta m/z = 2.01$ Da change, representing the molecular formula change of H₂, and the $\Delta m/z = 15.99$ Da change, representing the molecular formula change of O, were the two most frequent types, appearing 936 and 769 times, respectively. These 16 types of mass differences accounted for only 0.74% of all mass difference types, yet represented 46.18% of the metabolic reactions in the database (3916 reaction pairs). This indicated that the

mass differences corresponding to metabolic reactions have higher information density or entropy.

Our findings underscored the potential of integrating metabolic reaction information into MN construction. By leveraging both MS² spectral similarity and metabolic reaction matching (Fig. 1C), this new edge generation method achieved more accurate and comprehensive annotation of drug metabolism processes. This dual-approach methodology not only reduced false positives but also provided a clearer and more precise representation of metabolic pathways, facilitating a deeper understanding of drug metabolism in complex biological matrices.

3.2. Workflow of metabolic reactions based molecular networking

As shown in Fig. 2, the construction of MRMN involves the following steps: 1) Deconvolution of raw data and targeted filtering of metabolites-related ions: LC–MS data of biological samples and drugs were imported into data processing software (mainly MS-DIAL and MZmine)³¹. The data underwent

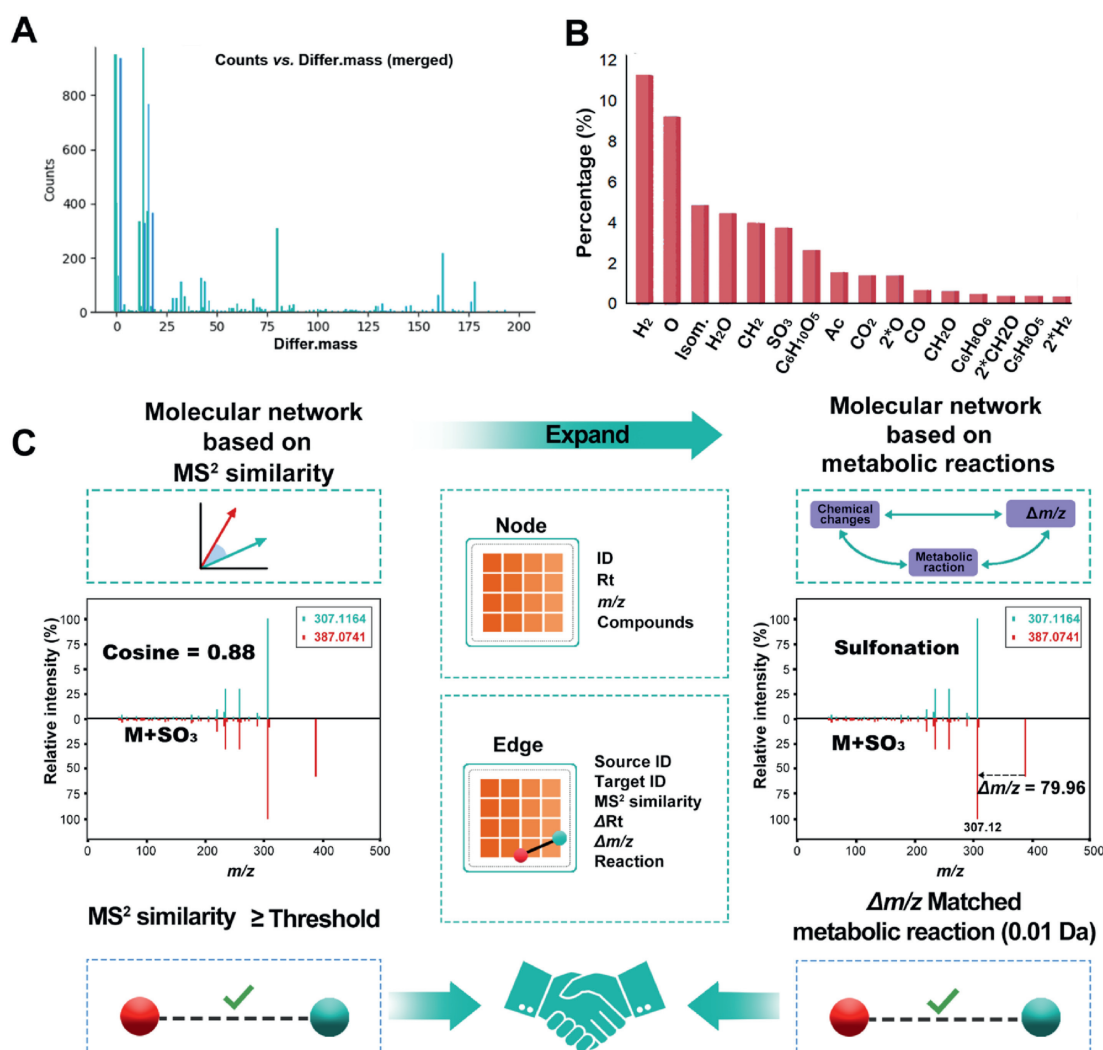


Figure 1 High information entropy of metabolic reactions and edge generation mechanism of MRMN. (A) Frequency distribution of mass changes caused by metabolic reactions (0–200 Da). (B) The top 16 mass changes most correlated with natural products and their corresponding molecular formula differences, representing 46.18% of metabolic reactions. (C) Edge generation mechanism integrating MS² spectral similarity and metabolic reaction matching.

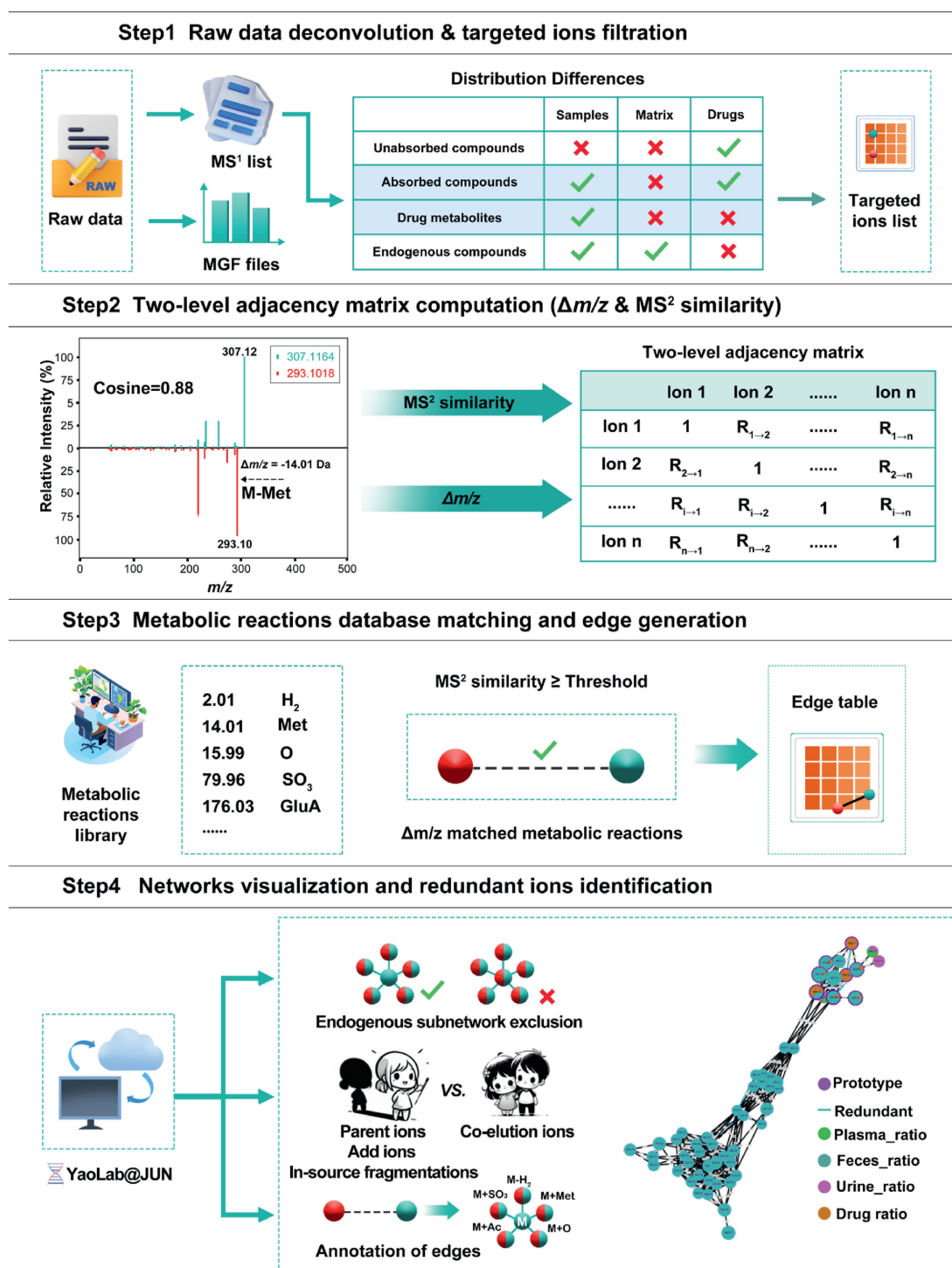


Figure 2 Logic diagram of metabolic reactions based molecular networking.

deconvolution, peak identification, and peak alignment, and were exported as MS^1 list files and MGF files. The response distribution correlation of prototypes and their metabolites-related ions in different biological samples and drugs was applied to filter out exogenous-related ions, preliminarily removing background interference from biological matrices. 2) Establishment of MS^2 spectral similarity & $\Delta m/z$ two-layer adjacency matrix of metabolite-related ions: develop a similarity calculation method tailored for drug metabolism research to compute the MS^2 spectral

similarity adjacency matrix and the $\Delta m/z$ adjacency matrix for all related ions. This method assigned dual-layer attributes of MS^2 spectral similarity and $\Delta m/z$, facilitating subsequent network construction. 3) Metabolic reaction database matching and edge generation: construct a metabolic reaction library containing 115 common *in vivo* metabolic reactions (Table S1). Determine whether the relationship between ion pairs in the dual-layer adjacency matrix satisfied the MS^2 spectrum similarity threshold and $\Delta m/z$ matched the metabolic reaction library within a given error

for generating the edge. 4) Networks visualization and annotation: the MRMN network was visualized based on the relationships in the generated edge files. Endogenous subnetworks were removed to reduce background interference and improve annotation accuracy. Potential redundant ions, in-source fragments, adduct ions, and co-eluting ions were identified using ion distribution characteristics and their MS² spectral similarity. Metabolites were then linked to their corresponding metabolic reaction pairs, enhancing the guidance for metabolite annotation.

Finally, we developed an integrated analysis platform (<https://yaolab.network>), where users can perform interactive MRMN analysis simply by uploading MS¹ list files and MGF files. The platform effectively identifies prototypes, removes endogenous subnetworks, automatically detects redundant data, and accurately annotates metabolites, offering a powerful and precise tool for the analysis of exogenous substances in complex samples.

3.3. Develop an algorithm to improve the feature degradation of MS² spectra caused by metabolic reactions

In MNs, the MS² spectral similarity between two nodes is crucial for determining their connectivity. Traditionally, it is understood that drug metabolism, being metabolic reactions based on the drug's core structure, should ideally not alter or minimally change the judgment of MS² spectral similarity between the prototypes and their metabolites. To validate this subjective assessment, we compared the structural similarity between adjacent metabolites in 3916 metabolic reaction pairs using the Morgan & Dice molecular fingerprint algorithm. The results showed that over 75% of the metabolic reactions have structural similarities between neighboring metabolites greater than 0.5 (Supporting Information Fig. S2). However, in metabolic reactions characterized by minor chemical structural changes (such as M-H₂, M + OH, and M-CH₂), a paradox arises: despite the high structural similarity between the substrate and the product, the MS² spectral similarity often exhibits a precipitous decline (Fig. 3A), falling significantly below the threshold typically used for edge generation in MNs. This rapid reduction in MS² spectral similarity between prototypes and their metabolites, induced by metabolic reactions, is identified as MS² feature degradation.

This degradation of features can have severe consequences, as even structurally similar prototypes and their metabolites may not be recognized as connected nodes in the network due to the abnormal decrease in MS² spectral similarity, leading to false negatives in the analysis. Using cimifugin as an example, we calculated the structural similarity and MS² spectral similarity between cimifugin and its six metabolites using RDKit and MS-match, respectively. The traditional cosine algorithm for MS² spectral similarity showed a strong correlation between M + SO₃ and cimifugin, with a structural similarity of 0.659 and MS² spectral similarity of 0.88. However, a significant discrepancy was observed between M-CH₂ and cimifugin, where the structural similarity was as high as 0.867, but the MS² spectral similarity was only 0.06, indicating typical feature degradation (Fig. 3A). Of the six metabolites, only M + SO₃ and M + C₆H₈O₆ had MS² spectral similarity with the prototype significantly above the default threshold of 0.7. The remaining metabolites showed feature degradation (Fig. 3B and Supporting Information Fig. S3), with a high occurrence rate of 66.6%.

Addressing feature degradation is essential for applying MNs in drug metabolite discovery. To calculate the MS² spectrum similarity between cimifugin and its six metabolites, three

algorithms were used: Cosine Greedy (CosGre), Neutral Loss Cosine (NLCos), and Modified Cosine (ModCos)²⁷. Results showed that the ModCos algorithm was the most effective in mitigating feature degradation, with all compounds located in the upper right corner, indicating a strong correlation between MS² spectral similarity and structural similarity. To validate the improvement of feature degradation by the ModCos algorithm on a broader scale, we used CFM-ID4.0 to construct SMILES structure databases and virtual MS² databases for 3949 metabolites across 3349 metabolic reactions^{32,33}, involving 12 common metabolic reactions (Supporting Information Fig. S4). Under the traditional cosine MS² spectral similarity algorithm, feature degradation rates for individual reactions ranged from 36.5% to 77.1%, with an overall rate of 51.5% (Fig. 3C). Violin plots of MS² spectral similarity scores (Fig. 3D) showed that NLCos performed the worst, CosGre had nearly 50% of scores below 0.5, and ModCos performed the best with at least 75% of scores above 0.5. Correlation analysis between MS² spectral similarity and structural similarity showed that ModCos results were mainly in the upper right corner, highlighting its potential to mitigate feature degradation compared to CosGre and NLCos (Fig. 3E–G).

To evaluate whether ModCos might increase the similarity between originally unrelated samples (false positives), we created two datasets: one consisting of 500 randomly selected real metabolic reaction pairs from the KEGG database and another with 500 randomly generated, non-existent reaction pairs. We compared the results using three MS² similarity algorithms and assessed the correlation between MS² similarity and structural similarity. From the real metabolic reaction pairs, ModCos ensured that the MS² similarity between substrates and products was greater than 0.5, and there was a strong correlation with structural similarity (Supporting Information Fig. S5). Overall, ModCos performed better than the other two algorithms. In contrast, for the generated non-existent reaction pairs, we observed that ModCos did not significantly increase the MS² similarity between unrelated compounds, with the majority of false products and substrates showing MS² similarity below 0.5 (Supporting Information Fig. S6). This result was supported by low structural similarity between these compounds. Although a few of the false reaction pairs had MS² similarity greater than 0.5 (Fig. S6), none of them matched with the KEGG metabolic reaction database. The dual mechanism of MS² similarity scoring and metabolic reaction matching in MRMN further reduces the likelihood of false positives.

3.4. Exclusion of endogenous interference ions substantially reduced the molecular network complexity of MRMN

Identifying relevant ion pairs of prototypes and their metabolites within complex biological matrix backgrounds is crucial for efficient compound annotation. This study proposes an automatic exclusion method for endogenous interference, integrating sample distribution differences and subnetwork filtering techniques (Fig. 4A). Firstly, sample distribution difference filtering is employed to obtain preliminary relevant ion information for prototypes and potential drug metabolites. After a drug enters the body, it undergoes absorption and metabolic reactions, resulting in the presence of absorbable prototypes, potential drug metabolites, and endogenous metabolites in the dosed biological samples. Specifically, prototypes are only present in the drug sample and the dosed biological sample, potential drug metabolites are only present in the dosed biological sample, while endogenous

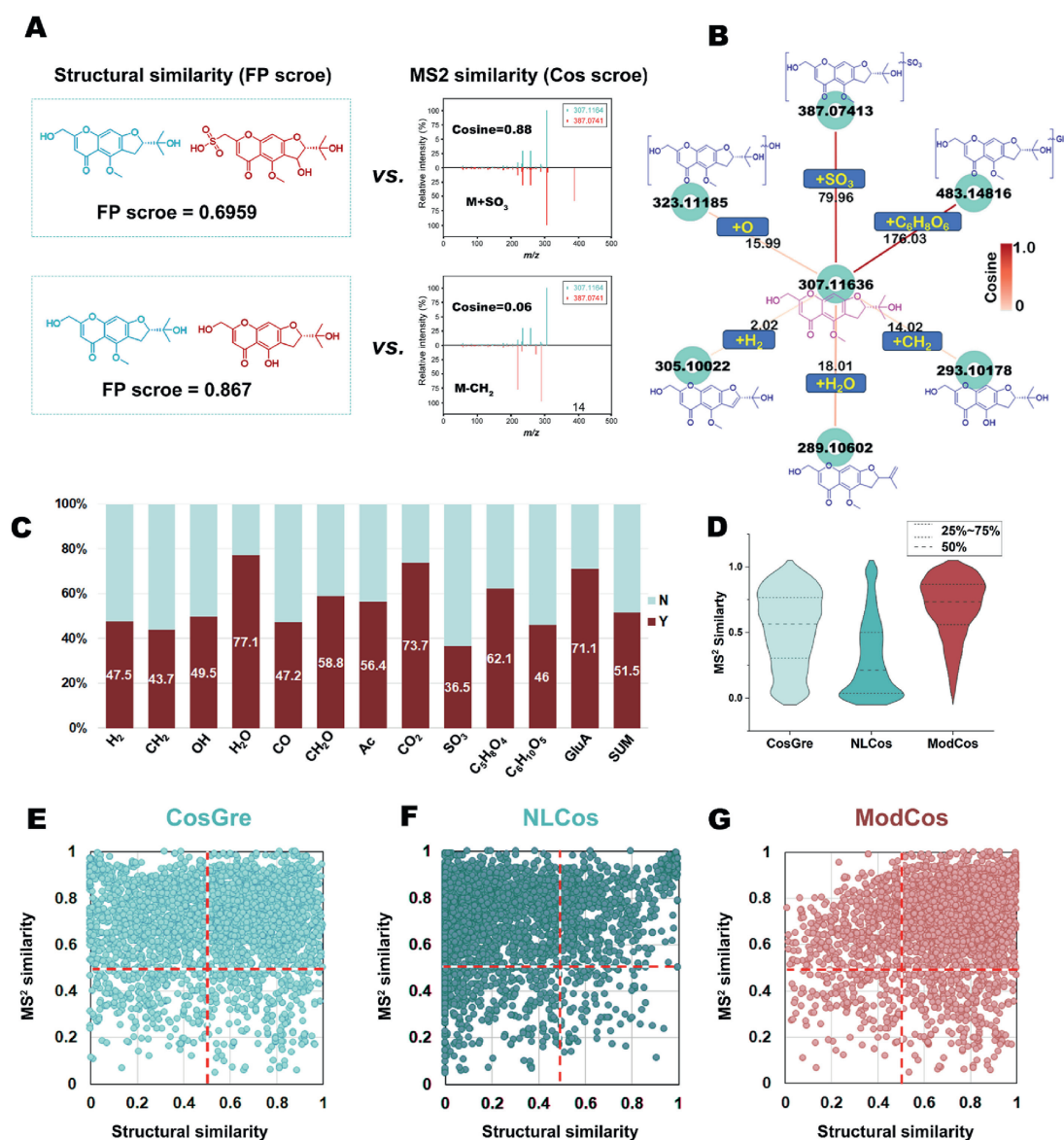


Figure 3 Feature degradation of MS² spectra induced by metabolic reactions and its improvement by the ModCos algorithm. (A) Structural similarity and MS² spectrum similarity of sulfonated cimifugin (M + SO₃) and demethylated cimifugin (M-CH₂) compared to cimifugin. (B) MS² feature degradation phenomenon between cimifugin and its common six metabolites under the traditional cosine similarity algorithm for MS² spectra. Only M + SO₃ and M + C₆H₈O₆ exhibited MS² spectral similarity with the prototype, exceeding the 0.7 threshold. The remaining metabolites showed significant feature degradation, with a 66.6% occurrence rate. (C) Comparison of feature degradation occurrence probabilities for 3359 pairs corresponding to 12 common mass changes under the traditional cosine similarity algorithm for MS² spectra (Feature degradation occurrence: Y, when both structural similarity and MS² spectral similarity are ≥0.5). (D) Violin plot distribution of MS² spectral similarity for 3359 metabolic reaction pairs under CosGre, NLCos, and ModCos MS² spectral similarity algorithms. (E) Scatter plot distribution of the correlation between structural similarity of 3359 metabolic reaction pairs and MS² spectral similarity under CosGre, (F) NLCos, and (G) ModCos algorithms, respectively.

metabolites may be present in both the blank sample and the dosed biological sample. Therefore, by applying a minimum response threshold (500,000) and a response ratio criterion between the sample and the blank matrix (Sample/Matrix ≥30), we obtained ions that are present only in the dosed biological sample (potential drug metabolites and significantly altered endogenous substances) and ions that are present in both the drug sample and the dosed biological sample (prototypes). Secondly, endogenous interference is further filtered based on subnetwork characteristics. In the MRMN constructed based on the ions obtained, prototypes and

their metabolites, which exhibit high MS² spectral similarity, tend to cluster in the same subnetwork. Therefore, subnetworks containing prototype-related nodes are likely linked to exogenous substances, whereas those without are considered endogenous interference and excluded.

A cross-sectional comparison of LC-MS data from *in vivo* metabolic samples, including a single compound (isosilybin, S1), a plant extract (fresh ginger, S2), and a mixture of plant extracts (YuPingFengSan, YPFS, S3), was conducted to validate the universality of the proposed method across different sample

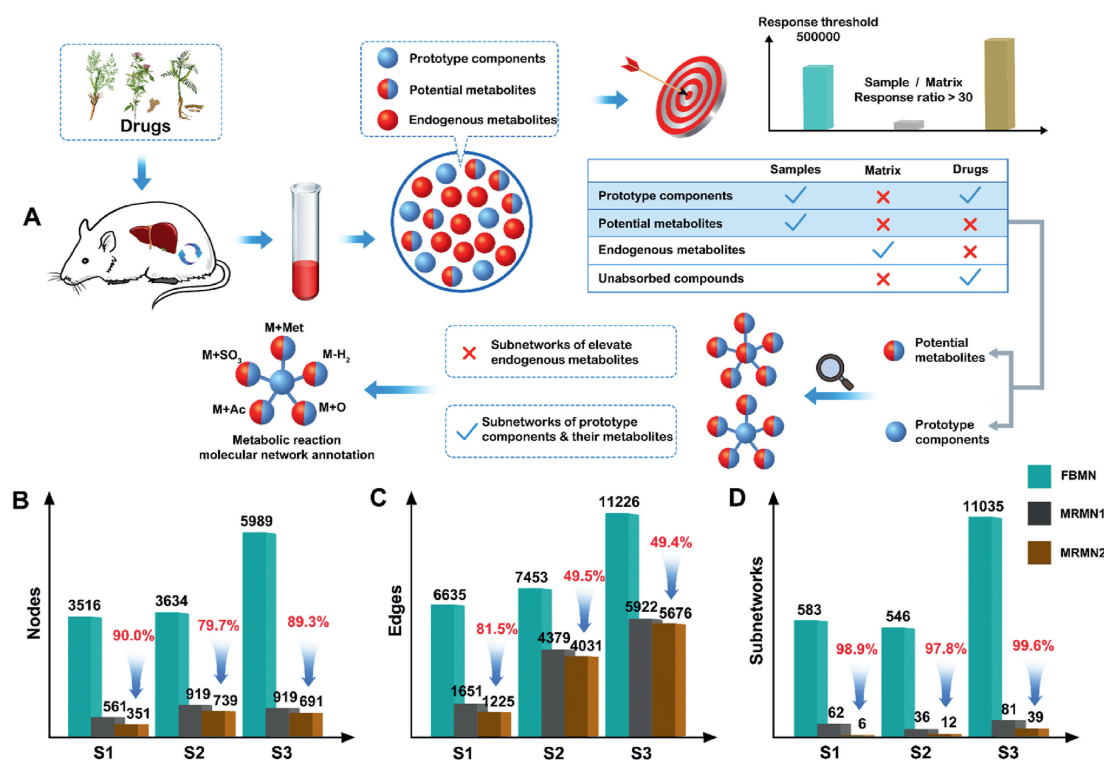


Figure 4 Filtering strategy for exogenous ions and endogenous interference exclusion in MRMN. (A) Ion filtering strategy based on sample distribution differences and exogenous-related subnetworks. (B) Comparison of nodes, edges, and subnetwork quantities in MRMN and FBMN for a single compound (S1: isosilybin), (C) a plant extract (S2: fresh ginger), (D) and a mixture of plant extracts (S3: YPFS), along with the proportion of endogenous interference exclusion.

types, chemical complexity, and structures. The overall visualization effects of feature-based molecular networking (FBMN) by global natural products social molecular networking (GNPS) platform and MRMN for the above LC–MS data were compared (Supporting Information Fig. S7). Results showed that the combination of sample distribution differences and subnetwork filtering significantly reduced the number of nodes, edges, and subnetworks requiring manual annotation. Specifically, node reduction rates were 90%, 79.37%, and 89.3% for S1, S2, and S3, respectively, mitigating endogenous interference. Correspondingly, edges were reduced by 81.5%, 45.9%, and 49.4%, respectively (Fig. 4B–D). Notably, the number of subnetworks requiring manual inspection was reduced to 6, 12, and 39, with a relative reduction ratio of 97.8%–99.6%, even as sample complexity increased. These results highlight the method's effectiveness in mitigating endogenous interference, focusing on subnetworks constructed by exogenous-related ions, and enabling more efficient discovery and annotation of components and their metabolites in complex samples within a biological matrix background.

3.5. Identification of redundant ions in MRMN

3.5.1. Identification of redundant ions based on evaluation of ions distribution similarity

Within the same retention time window, distinguishing redundant data arising from multiple adduct ions, endogenous fragment ions, and co-eluting component-related ions poses a significant challenge to the accurate annotation of molecular networking results. To address this issue, we developed a strategy based on the

similarity of ion sample distribution to identify and differentiate redundant data. Redundant ions and co-eluting component ions exhibit distinct sample distribution characteristics (Fig. 5A). Redundant ions, originating from the same compound molecule as the precursor ion, appear and disappear simultaneously across samples, leading to consistent response distribution trends. In contrast, co-eluting component ions, although peaking at the same retention time, originate from different compounds. Differences in their absorption and metabolism rates result in inconsistent response distribution trends across multiple samples. Heatmaps were employed to illustrate the distribution of redundant ion responses across different sample matrices (Supporting Information Fig. S8). These heatmaps reveal clustering patterns that distinguish redundant ions from co-eluting components. For example, in the same R_t (min) window of S3-172, three distinct clusters were observed. The precursor ion m/z 477.10 ($[M+H]^+$) and its redundant ions (m/z 494.12/ $[M + NH_4]^+$, m/z 301.07/ $[M-GluA + H]^+$, and m/z 285.07/ $[M-OH-GluA + H]^+$) exhibited similar distribution patterns across samples. Conversely, the unknown ion m/z 489.13 formed a separate cluster, suggesting it was a potential co-eluting component. This differentiation not only provides insights into ion behavior across sample groups but also helps correct erroneous connections in MN results. By evaluating the sample distribution similarity between adjacent ion nodes in MRMN, this strategy enables the identification of redundant data from multiple types of adduct ions, in-source fragment ions, and co-eluting components.

Inspired by the MS^2 spectral similarity algorithm, we propose a method for calculating sample distribution similarity between ions (Fig. 5B): Each ion's response values across a sequence of

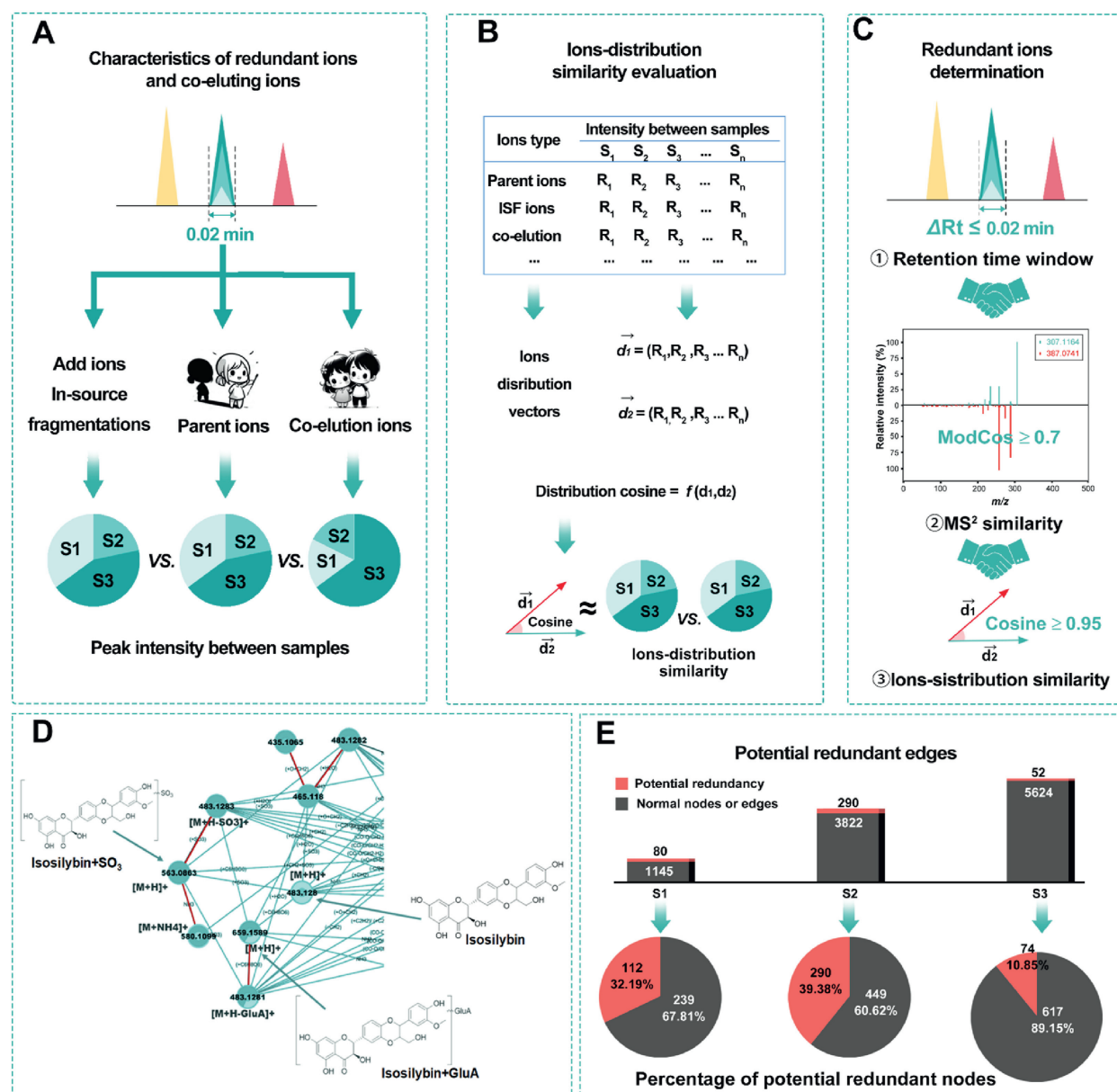


Figure 5 Redundant ions identification based on ions-distribution similarity algorithm and its recognition results in MRMN. (A) Redundant ions and compound precursor ions originate from the same molecule and behave like an object and its shadow, always appearing and disappearing together. In contrast, co-eluted component ions come from different compounds and resemble two people walking side by side, with varying absorption and metabolism rates, leading to inconsistent response distribution trends across samples. (B) A method was proposed to calculate the multi-sample distribution similarity of ions by transforming their response values across a sequence of N samples ($N \geq 2$) into an N -dimensional vector, defining their response distribution. The cosine similarity algorithm was then used to measure the similarity of these distributions by calculating the cosine value between the vectors. A cosine value close to 1 indicates high similarity, while a value close to 0 indicates low similarity. (C) The judgment criteria for identifying redundant ions based on ion sample distribution similarity algorithm: 1) two ions are in the same retention time window ($\Delta Rt \leq 0.02 \text{ min}$); 2) have a high MS² spectrum similarity (ModCos ≥ 0.7) and 3) have a multi-sample response distribution similarity (Cosine ≥ 0.95). (D) Visualization of redundant ions using isosilybin as an example. (E) Recognition effectiveness of redundant identification algorithm in MRMN.

samples (where the number of samples is N , and $N \geq 2$) are transformed into an N -dimensional vector $\vec{d}_i$, as shown in Eq. (1):

$$\vec{d}_i = (R_1, R_2, R_3 \dots R_N) \quad (1)$$

where the response distribution of each ion in the multi-sample sequence is defined by this multi-dimensional vector. Furthermore, the Cosine similarity algorithm is used to calculate the Cosine value between the multi-dimensional vectors of ions, representing their response distribution similarity across multiple

samples, as shown in Eq. (2): The higher the similarity, the closer the Cosine value is to 1; conversely, the closer the Cosine value is to 0, the lower the similarity.

$$\text{Distribution Cosine} = f\left(\vec{d}_1, \vec{d}_2\right) \quad (2)$$

$R_1, R_2, R_3 \dots R_N$ represent the aligned response sequences of ions across a sequence of N samples.

To ensure the accuracy of redundant ion identification, we also integrated MS^2 similarity results as one of the evaluation criteria. After manual inspection and organization of the redundant results in existing data, we established that the identification of redundant ions based on the ion sample distribution similarity algorithm must meet the following three conditions simultaneously (Fig. 5C): 1) The two ions must be within the same retention time window ($\Delta Rt \leq 0.02$ min); 2) The two ions must have a high MS^2 spectral similarity ($\text{ModCosd} \geq 0.7$); and 3) The two ions must exhibit good multi-sample response distribution similarity (Cosine ≥ 0.95).

In the visualization results of MRMN, the edges connecting ions identified as potential redundant nodes are highlighted in red to differentiate them from non-redundant data. For example, in the MRMN results of isosilybin (Fig. 5D), the visualization delineates redundant ions clearly. The sulfonation and glucuronidation products of isosilybin, along with relevant endogenous fragment ions and different types of adduct ions, were precisely annotated by matching with the metabolic reaction database, effectively avoiding false positive annotations caused by redundant ions. As shown in Fig. 5E, the redundant ion identification method detected 80, 209, and 52 potential redundant edges in the MRMN results of samples S1, S2, and S3, respectively. After further deduplication, 112, 290, and 74 potential redundant nodes were determined, with redundancy ratios of 32.19%, 39.38%, and 10.85%, respectively. This method can effectively identify redundant edges and nodes in MRMN, reducing potential false positives in annotations. Moreover, although the redundancy ratio varies with compound structure, the unexpectedly high proportion of redundant nodes further underscores the significance of this redundant identification method.

3.5.2. Comparison of the MRMN redundant ion identification approach with existing tools

One of the well-known methods for handling redundant nodes is Ion IIMN³⁰, which is an efficient tool for redundant ion recognition³⁴. It has been shown to be particularly effective in identifying adduct ions, isotopic peaks, and various types of adduct ions. In terms of redundant ion recognition algorithms, IIMN identifies redundant ions based on features such as the retention time proximity, similar chromatographic peak shapes, and MS^2 similarity of multiple types of adduct ions, isotopic peaks, and in-source fragment ions. In contrast, the MRMN is based on features like retention time proximity, MS^2 similarity, and the distribution similarity of ions across samples.

To further compare the redundancy recognition capabilities of MRMN and IIMN, we analyzed example datasets from two sample types: a single compound (isosilybin, S1) and a complex mixture (YPFS extract, S3). As shown in Supporting Information Fig. S9, the molecular networks generated by IIMN for S1 and S3 contained 7110 and 14,686 nodes, with 16,365 and 26,050 edges, respectively. In contrast, the corresponding networks generated by MRMN contained only 351 and 341 nodes, accompanied by 1225 and 2927 edges, respectively. This significant reduction in network

complexity is a direct benefit of MRMN's interference exclusion algorithm, which reduces the overall complexity of the molecular network and simplifies post-processing. Then, we compared the mapping of redundant ions associated with three precursor ions identified in the study: S1-48 (Isosilybin, R_t 13.485 min, m/z 483.1282, $[M+H]^+$), S3-123 (Calycosin + GluA, R_t 9.32 min, m/z 461.1080, $[M+H]^+$), and S3-172 (Pratensein + GluA, R_t 10.49 min, m/z 477.1025, $[M+H]^+$) in both IIMN and MRMN networks. For clearer comparison, we utilized Cytoscape software to independently process and visualize the sub-networks of these precursor ions in IIMN. The results indicate that in IIMN, precursor ions of S3-172 (Supporting Information Fig. S10A), S3-123 (Fig. S10B), and S1-48 (Fig. S10C) were directly connected to a limited number of redundant ions (indicated by red and blue dashed lines). However, a larger number of redundant ions were indirectly mapped through complex sub-networks. While IIMN does indeed identify more redundant ions overall, it is important to note that this process requires substantial manual intervention to interpret the complex sub-networks effectively. In contrast, MRMN provides a more intuitive visualization, directly mapping some key redundant ions (Fig. S10D and S10E). This approach simplifies the identification process, making MRMN more user-friendly and efficient for metabolite identification tasks.

In summary, both MRMN and IIMN exhibit comparable accuracy in redundant ion recognition. While IIMN identifies a higher number of redundant ions, MRMN excels in providing a more straightforward and efficient analysis. This advantage makes MRMN particularly suitable for applications such as drug metabolite identification, where ease of interpretation and processing efficiency are essential.

3.6. "One-pot" discovery and annotation of prototypes and their metabolites through online MRMN analysis

3.6.1. YaoLab platform for open-source MRMN analysis

To enhance the usability and broader application of MRMN, we have encapsulated the innovative algorithms and original scripts of MRMN into an open-source web analysis platform: YaoLab (<http://www.yaolab.network>). All of the open-source scripts and software for constructing MRMN were listed in Supporting Information Table S2^(23–25,28,32,33). YaoLab provides an intuitive and interactive analysis process with additional modules for MS^2 spectra visualization, prototype drug structure visualization, and feature ion filtering. These features not only simplify the analysis process but also allow users to critically evaluate and refine their results. The MRMN analysis workflow based on YaoLab is outlined in Fig. 6A. Users preprocess raw data using MS-DIAL or MZ-mine software to generate corresponding MS^1 ion lists and MS^2 data in MGF format. These files are then uploaded to the platform for data grouping, which includes biological samples, blank biological samples, and optional drug samples (necessary for filtering endogenous subnetworks). Users can also set up custom metabolic reaction libraries according to their research needs. The platform then executes algorithms to perform various tasks, such as exploring exogenous-related ions, constructing metabolic reaction databases, generating MS^2 spectral similarity and $\Delta m/z$ double-layer adjacency matrices, matching metabolic reaction databases, generating molecular network-related files, identifying redundant nodes, subnetwork filtering, and mapping node color. The platform outputs a visualized MRMN and provides download interfaces for node and edge files, supporting further offline annotation. MRMN's primary focus is on

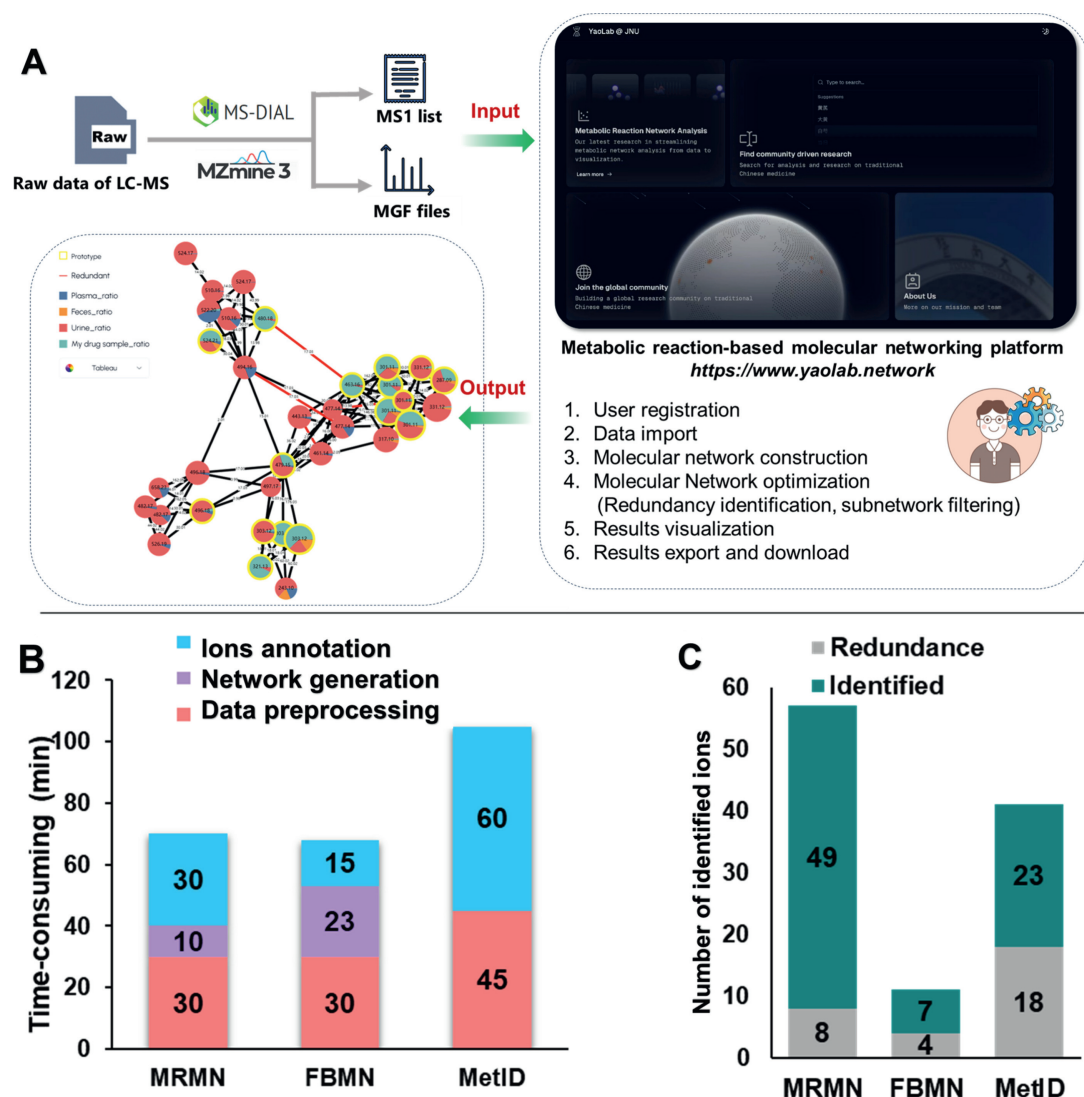


Figure 6 “One-Pot” discovery and annotation of prototypes and their metabolites through online MRMN. (A) Schematic of the MRMN workflow based on the network platform (<https://www.yaolab.network>). (B) Comparison of the time consumption of the whole process of analyzing metabolites in isosilybin-related biological samples using MRMN, FBMN, and MetID; (C) comparison of the results of the number of identified metabolites in isosilybin-related biological samples using MRMN, FBMN, and MetID.

identifying the potential connections between *in vivo* prototypes and their metabolites quickly and visually and further annotation through metabolic reactions and mass fragment characteristics.

As shown in Supporting Information Fig. S11, the MRMN annotation process is illustrated using the example of methyl-nissolin and its glycoside compounds in YPFS. This subnetwork involves 35 ion nodes and 103 edges, with 4 prototypes (red structures) and 15 related metabolites (blue structures) identified through edge attribute labeling. Additionally, 8 potential redundant ions, 2 adduct ions, and source fragmentation ions were identified. Finally, the annotation rate of nodes in this subnetwork reached 100%. Furthermore, we utilized MRMN to profile xenobiotic substances in biological samples after the administration of isosilybin, fresh ginger, and YPFS. As a result, we identified and annotated 49 metabolites of isosilybin (Supporting Information Table S3), 45 prototypes of fresh ginger and their 100 metabolites (Supporting Information Table S4), and 95 prototypes of YPFS and their 192 metabolites (Supporting Information Table S5

and Supporting Information Fig. S12). Thanks to the exclusion of matrix interference, the comprehensive annotation rates (*i.e.*, the number of annotated ions/total ions in the MRMN network) for S1, S2, and S3 samples were 15.6% (55/351), 25.8% (191/739), and 41.9% (290/691), respectively. These results clearly demonstrate the excellent annotation performance of MRMN for exogenous ions of interest in complex biological samples. However, it must be acknowledged that some false-positive network clusters, which were likely generated by endogenous compounds exhibiting significant changes following drug administration, still appeared in the MRMN results. Notably, these ions usually formed isolated sub-networks, which were distinct from the expected classical prototypes of plant extracts or single compounds. For example, as shown in Supporting Information Fig. S13, the MRMN of isosilybin revealed certain phospholipid compounds, such as PC 36:1 (identified by the diagnostic phospholipid ion at *m/z* 184.07), which were visualized but did not cluster with the isosilybin sub-network. The integration of additional professional

judgment, based on retention time, MDF, and in-source database matching, may be an effective strategy to resolve these challenges. Additionally, MRMN provides detailed annotations for all detected ions, including their sub-network classifications, retention times, and MS² spectra, enabling users to independently apply these filters for more accurate outcomes.

3.6.2. Comprehensive comparison between MRMN and existing tools

To compare the efficiency and accuracy of MRMN, FBMN, and MetID (Supporting Information Table S6), we analyzed isosilybin-related data (S2). As shown in Fig. 6B, MRMN and FBMN completed the entire workflow at similar times (MRMN ~70 min and FBMN ~68 min), both faster than MetID, which took 105 min. However, while MRMN identified 49 metabolites of isosilybin, FBMN only 7 metabolites, and MetID only 23 metabolites highlighting MRMN's superior performance in metabolite identification. The data preprocessing times for all three tools were similar, but the key differences were in molecular network construction time (MRMN 10 min vs. FBMN 23 min) and ion annotation time (MRMN 30 min vs. FBMN 10 min vs. MetID 60 min). Unlike FBMN, which handles all ions for broad metabolomics studies, MRMN focuses only on drug-related ions and applies effective background interference elimination. This targeted approach likely accounts for the reduced computational demand, making MRMN's molecular network construction faster. Furthermore, MRMN's advanced algorithms, such as optimized MS² similarity evaluation, metabolic reaction matching, and redundant ion recognition, led to a much higher quantity and accuracy of drug metabolite identification compared to both FBMN and MetID (Fig. 6C). This also helps explain the longer ion annotation time in MetID due to its high false-positive rate and the shorter time in FBMN, which suffers from low sensitivity and insufficient data. Additionally, MetID predicts metabolites based on predefined prototype compounds, ranking them using FISh Scoring, which limits it to single-compound analysis. In contrast, MRMN can analyze multiple matrices (e.g., blood, tissue, feces, urine) without requiring predefined prototypes, making it especially suited for complex samples like traditional Chinese medicine. The superior performance of MRMN is attributed to its targeted design, which enhances both efficiency and accuracy in drug metabolism analysis.

To validate its broader applicability, we reanalyzed raw MS^E LC–MS data (DIA mode) of 6-gingerol samples from a previous study³⁵. The original study identified 40 metabolites using MetaboLynx XS and MDF methods from negative data. In comparison, MRMN annotated 197 metabolites from the same dataset (Supporting Information Table S7), encompassing most of the manually identified metabolites and uncovering numerous novel candidates. As shown in Supporting Information Fig. S14, the MRMN results not only encompassed most of the manually identified metabolites from the original study (30/40) but also revealed numerous potential novel metabolites (168/198). This comparison underscores the enhanced capabilities and broad applicability of the MRMN approach. This demonstrates MRMN's enhanced capacity for comprehensive metabolite annotation and its potential to advance drug metabolism research.

3.6.3. Data analysis and interpretation of MRMN

MRMN provides valuable insights into the connections between a drug's prototype and its metabolites, which is essential for metabolite characterization. Due to the complex metabolic pathways of prototype compounds, multiple candidate metabolites can be generated, and the correlation between these metabolites and

experimental results may not always be clear. Therefore, it is necessary for users to rely on their expertise to make informed decisions. To address the challenge of accurate metabolite identification and improve the analysis process, we have introduced several strategies in the MRMN platform: 1) Metabolite ranking based on network connectivity: MRMN ranks metabolites according to their connectivity within the molecular network. Metabolites located more centrally with higher connectivity are likely to play a more significant role in the metabolic pathway, helping users prioritize those with higher biological relevance; 2) Integration of metabolic pathway information: by incorporating metabolic pathway data, MRMN provides additional context for metabolite ranking. Metabolites associated with known metabolic pathways are prioritized as they are more likely to be involved in critical biological processes; 3) User-defined cut-off criteria: MRMN allows users to set custom cut-off criteria based on their experimental needs. This includes thresholds for MS² similarity scores, retention time alignment, and other parameters, enabling more tailored and precise metabolite selection; 4) Visualization and interactive tools: the platform features interactive visualization tools that assist users in evaluating the molecular networks. Users can adjust cut-off criteria in real-time, enhancing the intuitive nature of the analysis and facilitating better decision-making; 5) Automated filtering algorithms: automated algorithms are employed to filter out noise and redundant information, simplifying the dataset and allowing users to focus on the most relevant metabolites.

These strategies provide a robust framework for metabolite ranking and cut-off criteria establishment, making the analysis of large metabolite datasets more manageable and interpretable. By combining automated algorithms, user-defined parameters, and interactive tools, MRMN ensures a more efficient and effective data analysis process. Looking ahead, accurate identification and sequencing of metabolites remains a key goal for the future development of MRMN. While MRMN cannot fully resolve this challenge at its current stage, we have introduced a SMILES structure import and visualization module for prototype compounds at Yaolab. This feature allows users to interactively visualize the compound's structure and potential reaction sites, facilitating better prediction of metabolites.

4. Discussion and conclusions

In this study, we developed and validated the MRMN approach to enhance the analysis of drug metabolites from LC–MS data. By focusing on the molecular formula differences and structural similarities between prototypes and metabolites in metabolic reactions, MRMN innovatively improved the mechanism of edge generation from a higher-dimensional perspective. This advancement eliminates the need for prior knowledge, such as the pre-specification of prototypes, thereby enabling a more flexible and comprehensive analysis of complex metabolite profiles.

A major contribution of this study is the validation of the ModCos MS² spectral similarity algorithm. Through large-scale predictions using MS² databases, ModCos effectively addressed the inevitable feature degradation between prototypes and metabolites, achieving high correlations between structural and MS² spectral similarity in over 75% of common metabolic reactions. Notably, the algorithm did not introduce false-positive links between unrelated ions, further demonstrating its reliability and precision in metabolite analysis. Additionally, the combination of ion distribution differences across samples with subnetwork feature filtering proved highly effective in removing endogenous

interference. This process reduced interference nodes, edges, and subnetworks by approximately 80%–90%, 45%–81%, and 97%–99.5%, respectively, significantly streamlining the analysis process and reducing manual labor. The study also introduced a novel algorithm for evaluating ion response distribution similarity across multiple samples using the cosine similarity of response vectors. This algorithm successfully identified and reduced redundant nodes caused by adducts and in-source fragmentation ions, decreasing redundancy by 10%–40%. This reduction in redundancy improved both the accuracy and annotation precision of the MRMN workflow. Comparative analyses with existing tools further highlighted MRMN's advantages in computational efficiency, accuracy, and sensitivity. MRMN's targeted molecular network construction focuses on exogenous ions, effectively removing background interference and reducing molecular network complexity. Compared to comprehensive molecular networks, such as those generated by GNPS^{29,30}, MRMN's targeted approach not only decreases computational power requirements but also enhances sensitivity for identifying drug-related ions. Furthermore, MRMN supports simultaneous analysis of multiple matrices (*e.g.*, blood, tissue, feces, urine) within a single workflow. This capability is particularly advantageous for analyzing complex samples, such as plant extracts and traditional Chinese medicine, without the need for prior predictions of absorbable prototypes—a limitation of traditional approaches like MDF or MetID^{8,13}.

In summary, MRMN represents a significant advancement in the field of drug metabolite analysis, offering a robust and comprehensive solution for the “one-pot” discovery and annotation of prototypes and their endogenous substances. Its innovative algorithms, including ModCos for spectral similarity, redundant node recognition, and interference reduction, position MRMN as a superior alternative to existing tools in terms of computational efficiency, sensitivity, and precision. While designed primarily for small-molecule *in vivo* metabolism analysis from LC–MS data, the underlying principles of MRMN are broadly applicable to other scientific domains. These include the chemical and metabolic analysis of plant extract, food^{36,37}, and pesticides^{38,39}, demonstrating MRMN's versatility as a powerful tool for exploring chemical and metabolic profiles across diverse disciplines. Given the diverse metabolic pathways of prototype compounds⁴⁰, multiple candidate metabolites may be generated^{41,42}, and the correlation between candidate metabolites and experimental results is not definitive. This requires users to rely on their expertise and experience to make informed judgments. Therefore, challenges remain in the areas of accurate structural prediction of metabolites^{43–45}, determination of relative configurations⁴⁶, and annotation of unknown metabolites¹⁸. Future studies should address these limitations to further expand the applicability and reliability of MRMN in metabolomics and beyond.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (U23A20500, 82374011, 82474050, 82404818), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (ZYYCXTU-D-202203, China), the Guangdong Basic and Applied Basic Research Foundation (2025A1515011795, 2023A1515011144, 2024A1515012714, 2024A1515011699, China), the Guangzhou Basic and Applied Basic Research

Foundation (2024A04J3398, China), the China Postdoctoral Science Foundation (2023M741395), the Postdoctoral Fellowship Program of CPSF (GZB20240274, China), the Natural Science Foundation of Guangxi (2025GXNSFBA069293, China), the Scientific Research Start-up Funding Project of Guangxi University (ZX01080033724006, China), and supported by the Fundamental Research Funds for the Central Universities (China).

Author contributions

Haodong Zhu designed the conceptualization, methodology, and software, carried out the investigation and data curation, and wrote the original draft. Xupeng Tong designed the conceptualization, methodology, and software. Qi Wang developed the methodology and conducted the investigation and data curation. Aijing Li, Zubao Wu, and Qiqi Wang performed the investigation and data curation. Pei Lin and Xinsheng Yao contributed to the writing review & editing and provided supervision and project administration. Liufang Hu provided supervision, methodology, writing review & editing, and project administration. Liangliang He and Zhihong Yao contributed to the conceptualization, supervision, methodology, writing review & editing, and project administration. All of the authors have read and approved the final manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

Code availability

MRMN is primarily developed using Python and is available for free non-commercial use at <https://www.yaolab.network>. The main page of <https://www.yaolab.network> provides an interactive tutorial for using MRMN and download links for demo data, ensuring that novice users can quickly get started.

Data availability

A dataset serves as crucial supporting documentation for establishing and validating MRMN (<https://doi.org/10.6084/m9.figshare.28294868.v1>), comprising the following components: 1) Predicted MS² database of 3949 compounds in 3349 selected pairs involving 12 common metabolic reactions. 2) MS¹ list data and MS² MGF files derived from biological samples administered Isosilybin, Fresh Ginger, and YPFS, obtained by MS-DIAL for online MRMN analysis on the platform. 3) Node and edge files of MRMN for LC–MS data from biological samples administered Isosilybin, Fresh Ginger, and YPFS, generated at These files are intended for offline MRMN processing using Cytoscape 3.8.0. The GNPS-FBMN MN results for biological samples administered Isosilybin, Fresh Ginger, and YPFS using the same dataset can be accessed at the following links:

<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5f9b4407dfdc4fa69783793b4416b10f> (S1-FBMN).
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=464691a11862423ea7fd720bbebdb4be> (S2-FBMN).
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=21c40eaaef4c4baaafa1915f2109b3f> (S3-FBMN).
<https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=34bff70f353c4897abc22ca9f800e49f> (S1-IIMN).

[https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9ef4972b11114cc3ad40f19388c4e0e3](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9ef4972b11114cc3ad40f19388c4e0e3....) (S2-IIMN).
[https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=36c428c887444ec1a588cda93e91a72d](https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=36c428c887444ec1a588cda93e91a72d....) (S3-IIMN).

Appendix A. Supporting information

Supporting information to this article can be found online at <https://doi.org/10.1016/j.apsb.2025.03.050>.

References

- Park BK, Boobis A, Clarke S, Goldring CEP, Jones D, Kenna JG, et al. Managing the challenge of chemically reactive metabolites in drug development. *Nat Rev Drug Discov* 2011;**10**:292–306.
- Kirchmair J, Göller AH, Lang D, Kunze J, Testa B, Wilson ID, et al. Predicting drug metabolism: experiment and/or computation?. *Nat Rev Drug Discov* 2015;**14**:387–404.
- Wienkers LC, Heath TG. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. *Nat Rev Drug Discov* 2005;**4**:825–33.
- Di L. The role of drug metabolizing enzymes in clearance. *Expet Opin Drug Metabol Toxicol* 2014;**10**:379–93.
- Schadt S, Bister B, Chowdhury SK, Funk C, Hop CECA, Humphreys WG, et al. A decade in the MIST: learnings from investigations of drug metabolites in drug development under the “metabolites in safety testing” regulatory guidance. *Drug Metab Dispos* 2018;**46**:865–78.
- Chen C, Wohlfarth A, Xu H, Su D, Wang X, Jiang HL, et al. Untargeted screening of unknown xenobiotics and potential toxins in plasma of poisoned patients using high-resolution mass spectrometry: generation of xenobiotic fingerprint using background subtraction. *Anal Chim Acta* 2016;**944**:37–43.
- Castro-Perez JM. Current and future trends in the application of HPLC–MS to metabolite-identification studies. *Drug Discov Today* 2007;**12**:249–56.
- Wohlfarth A, Scheidweiler KB, Chen XH, Liu HF, Huestis MA. Qualitative confirmation of 9 synthetic cannabinoids and 20 metabolites in human urine using LC–MS/MS and library search. *Anal Chem* 2013;**85**:3730–8.
- Cuyckens F, Hurkmans R, Castro-Perez JM, Leclercq L, Mortishire-Smith RJ. Extracting metabolite ions out of a matrix background by combined mass defect, neutral loss and isotope filtration. *Rapid Commun Mass Spectrom* 2009;**23**:327–32.
- McLeod MD, Waller CC, Esquivel A, Balcells G, Ventura R, Segura J, et al. Constant ion loss method for the untargeted detection of bis-sulfate metabolites. *Anal Chem (Washington, DC, U S)* 2017;**89**:1602–9.
- Klitgaard A, Nielsen JB, Frandsen RJN, Andersen MR, Nielsen KF. Combining stable isotope labeling and molecular networking for biosynthetic pathway characterization. *Anal Chem* 2015;**87**:6520–6.
- 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 Mass Spectrom* 2008;**43**:1181–90.
- Delcourt V, Barnabé A, Loup B, Garcia P, André F, Chabot B, et al. MetIDfyR: an open-source R package to decipher small-molecule drug metabolism through high-resolution mass spectrometry. *Anal Chem* 2020;**92**:13155–62.
- Aguilar-Mogas A, Sales-Pardo M, Navarro M, Guimera R, Yanes O. iMet: a network-based computational tool to assist in the annotation of metabolites from tandem mass spectra. *Anal Chem* 2017;**89**:3474–82.
- Shen XT, Wang RH, Xiong X, Yin YD, Cai YP, Ma ZJ, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun* 2019;**10**:1516.
- Guo JD, Shang Y, Yang XH, Li J, He J, Gao XM, 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 Chromatogr A* 2022;**1675**:463172.
- Yu JS, Nothias LF, Wang MX, Kim DH, Dorrestein PC, Kang KB, et al. Tandem mass spectrometry molecular networking as a powerful and efficient tool for drug metabolism studies. *Anal Chem* 2022;**94**:1456–64.
- Zhou ZW, Luo MD, Zhang HS, Yin YD, Cai YP, Zhu ZJ. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun* 2022;**13**:6656.
- Xing SP, Hu Y, Yin ZX, Liu M, Tang XY, Fang ML, et al. Retrieving and utilizing hypothetical neutral losses from tandem mass spectra for spectral similarity analysis and unknown metabolite annotation. *Anal Chem* 2020;**92**:14476–83.
- Schollée JE, Schymanski EL, Stravs MA, Gulde R, Thomaidis NS, Hollender J. Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *J Am Soc Mass Spectrom* 2017;**28**:2692–704.
- Zhu HD, He LL, Wu WY, Duan HF, Chen JL, Xiao Q, et al. A compounds annotation strategy using targeted molecular networking for offline two-dimensional liquid chromatography-mass spectrometry analysis: Yipingfeng as a case study. *J Chromatogr A* 2023;**1702**:464045.
- Schmid R, Petras D, Nothias LF, Wang MX, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* 2021;**12**:3832.
- Bento AP, Hersey A, Félix E, Landrum G, Gaulton A, Atkinson F, et al. An open source chemical structure curation pipeline using RDKit. *J Cheminf* 2020;**12**:1–16.
- Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
- Goloborodko AA, Levitsky LI, Ivanov MV, Gorshkov MV. Pyteomics—a Python framework for exploratory data analysis and rapid software prototyping in proteomics. *J Am Soc Mass Spectrom* 2013;**24**:301–4.
- Huber F, Ridder L, Verhoeven S, Spaaks JH, Diblen F, Rogers S, et al. Spec2Vec: improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol* 2021;**17**:e1008724.
- Huber F, van der Burg S, van der Hooft JJJ, Ridder L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Cheminf* 2021;**13**:84.
- Bittremieux W, Schmid R, Huber F, van der Hooft JJJ, Wang MX, Dorrestein PC. Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. *J Am Soc Mass Spectrom* 2022;**33**:1733–44.
- Tsugawa H, Ikeda K, Takahashi M, Satoh A, Mori Y, Uchino H, et al. A lipidome atlas in MS-DIAL 4. *Nat Biotechnol* 2020;**38**:1159–63.
- Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* 2020;**17**:905–8.
- Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrland TS, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* 2023;**41**:447–9.
- Boettiger C. An introduction to Docker for reproducible research. *ACM SIGOPS - Oper Syst Rev* 2015;**49**:71–9.
- Wang F, Liigand J, Tian S, Arndt D, Greiner R, Wishart DS. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;**93**:11692–700.
- Zhang YH, Liao JY, Le WQ, Zhang WD, Wu GS. In-depth analysis of molecular network based on liquid chromatography coupled with tandem mass spectrometry in natural products: importance of redundant nodes discovery. *Anal Chem* 2024;**96**:15888–97.
- He LL, Qin ZF, Li MS, Chen ZL, Zeng C, Yao ZH, et al. Metabolic profiles of ginger, a functional food, and its representative pungent compounds in rats by ultraperformance liquid chromatography

- coupled with quadrupole time-of-flight tandem mass spectrometry. *J Agric Food Chem* 2018;**66**:9010–33.
36. Aherne SA, O'Brien NM. Dietary flavonols: chemistry, food content, and metabolism. *Nutrition* 2002;**18**:75–81.
37. Chen YH, Michalak M, Agellon LB. Importance of nutrients and nutrient metabolism on human health. *Yale J Biol Med* 2018;**91**:95–103.
38. Perez de Albuquerque NC, Carrão DB, Habenschus MD, Moraes de Oliveira AR. Metabolism studies of chiral pesticides: a critical review. *J Pharm Biomed Anal* 2018;**147**:89–109.
39. Zhang JJ, Yang H. Metabolism and detoxification of pesticides in plants. *Sci Total Environ* 2021;**790**:148034.
40. Kumar GN, Surapaneni S. Role of drug metabolism in drug discovery and development. *Med Res Rev* 2001;**21**:397–411.
41. Zhang ZP, Tang W. Drug metabolism in drug discovery and development. *Acta Pharm Sin B* 2018;**8**:721–32.
42. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol* 2019;**20**:353–67.
43. Kazmi SR, Jun R, Yu MS, Jung CJ, Na D. *In silico* approaches and tools for the prediction of drug metabolism and fate: a review. *Comput Biol Med* 2019;**106**:54–64.
44. Hughes TB, Dang NL, Kumar A, Flynn NR, Swamidass SJ. Metabolic forest: predicting the diverse structures of drug metabolites. *J Chem Inf Model* 2020;**60**:4702–16.
45. Shen ZW, Lv C, Zeng S. Significance and challenges of stereo-selectivity assessing methods in drug metabolism. *J Pharm Anal* 2016;**6**:1–10.
46. Garcia-Perez I, Posma JM, Serrano-Contreras JI, Boulange CL, Chan Q, Frost G, et al. Identifying unknown metabolites using NMR-based metabolic profiling techniques. *Nat Protoc* 2020;**15**:1–30.