scientific reports

Check for updates

OPEN

Dereplication of secondary metabolites from *Sophora flavescens* using an LC–MS/MS-based molecular networking strategy

Hua Wang[1]✉ & Hui Ding[2]

A dereplication strategy was developed for the screening of secondary metabolites from *Sophora flavescens*. The strategy consisted of 4 procedures. First, the extract of the *Sophora flavescens* root was subjected to LC–MS/MS analysis with both data-independent acquisition (DIA) mode and data-dependent acquisition (DDA) mode. Then the DIA results were used to construct a molecular networking (MN) according to the GNPS workflow and consequently obtain annotations. In parallel, the DDA results were projected to both MN analysis and direct databases matching to obtain annotations. Finally, the isomers were discriminated and annotated by their extracted ion chromatogram. Through the combination of these approaches, a total of 51 compounds were annotated and dereplicated in the *Sophora flavescens* samples. The annotation results showed DIA and DDA approach are complementary to each other. MN on GNPS can overcome the challenges of trace compound identification compared to direct DB matching. This strategy provides a powerful tool for the dereplication study in plant chemistry.

Sophora flavescens is a perennial subshrub in the genus *Sophora*, family *Fabaceae*. It is distributed throughout China and also in India, Japan, Korea, and some European countries. The secondary metabolites of *Sophora flavescens* are mainly alkaloids, flavonoids, triterpenoids and other active constituents. Typical alkaloids such as matrine, oxymatrine, sophoridine[1], etc., have significant antioxidation, anti-tumor[2], anti-inflammatory[3] and other biological activities[4]. Flavonoids such as kurarinone, xanthohumol, etc., have antioxidant, anti-inflammatory[5] and anti-allergic[6] effects.

The secondary metabolites of *Sophora flavescens* are the basis of evidence-based pharmacy. It is important to characterize the component. Although *Sophora flavescens* has been extensively studied, new compounds are still being discovered[7–9]. In order to find new compounds in a well-studied plant, dereplication is necessary[10]. Dereplication means to rapid identify those previously characterized molecules. To date, there is a lack of a systematic strategy to achieve dereplication together with characterization and relative quantification of the constitution of *Sophora flavescens*.

Liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) is a powerful tool for compound analysis and characterization[11]. Ultra-high performance liquid chromatography coupled to high-resolution mass spectrometry (UPLC-HRMS) provides better separation and more accurate mass-to-charge (*m/z*) ratio of the analyte ion[12]. In mass spectrometry, multiple-reaction monitoring (MRM) acquisition were developed for targeted metabolomics, while either data-independent acquisition (DIA) or data-dependent acquisition (DDA) were used for untargeted metabolomics[13]. Nevertheless, the annotation of metabolites in untargeted datasets represents a significant challenge for researchers[14].

A straightforward dereplication approach is to compare the mass spectrum of the candidate compound with those of known natural products. However, as the amount of library data increases[15] and each LC–MS run generates a significant number of spectra[16], it becomes impossible to match spectra manually. The simple library search often resulted in mismatch because of the data variation from the experimental conditions[17,18].

[1]Instrumental Analysis Center, Yancheng Teachers University, Yancheng 224007, China. [2]School of Pharmacy, Yancheng Teachers University, Yancheng 224007, China. ✉email: wangh@yctu.edu.cn

The development of algorithms to systematically screen and dereplicate these structurally diverse compounds is necessary[19].

Molecular networking (MN) is a method for organizing, visualizing and annotating untargeted MS/MS data[20,21]. In MN, the mass spectra of molecules are connected based on the similarity of fragmentation patterns[15,22]. Molecular networking within the Global Natural Products Social (GNPS) platform[23] has been successfully applied to discovering and identifying natural products[24]. Recently, feature-based molecular networking (FBMN) in the GNPS community has also been used as an analysis tool for chemical annotation[25,26]. Other modified MN approaches have also annotated global metabolites from known to unknowns in untargeted metabolomics[27].

For the purpose of dereplication, we tend to provide more spectral information to MN in order to obtain more component identification. Although all MS/MS fragmentation information is included in the DIA data, the information is complex. Even after the deconvolution, the data from DIA are still complicated and difficult to interpret. On the other hand, the MS/MS spectra in DDA data are simpler and can be used directly for DB matching.

Here, we proposed a dereplication strategy for the metabolomic study of *Sophora flavescens*. The pipeline of strategy was illustrated in Fig. 1.

The analytical pipeline consisted of 4 procedures. First, the extract of the *Sophora flavescens* sample was subjected to LC–MS/MS analysis with both DIA and DDA modes. Then, the raw DIA data were processed and aligned into compatible version to construct the MN. Next, the raw DDA data were searched directly in public databases for the complementary annotation to the MN approach. Finally, the putative annotations were combined and the isomers were annotated by their extracted ion chromatogram (EIC). This dereplication strategy could be a useful tool for the discovery of new metabolites in the plant metabolomics studies.

## Experimental section
Materials and reagents
Standards of matrine, sophoridine, kurarinone, anagyrine, sophoramine, neosophoramine, and calycosin-7-O-beta-D-glucoside were purchased from Chengdu Zhibiao Biotech Corporation Limited (Chengdu, China). Standards of trifolirhizin, oxysophocarpine, xanthohumol and isoxanthohumol were purchased from Wuhan Tianzhi Biotech Corporation Limited (Shanghai, China).

The sample of the *Sophora flavescens* root was purchased from a grower in Wenshan County, Yunnan Province, China. The grower grew and sold the sample with the Wenshan Government license (851749 K). The sample was identified by the Dr. Mei Deng of China Pharmaceutical University. The voucher specimen (No. YCTU-IAC-2024-Mar-26-01) is deposited in the Instrumental Analysis Center of Yancheng Teachers University.

Chromatographic grade methanol and acetonitrile (ACN) were both purchased from Tedia Company, Inc. (Fairfield, USA). Formic acid (purity > 98.0%) was purchased from Tokyo Chemical Industry Co., Ltd. (Tokyo, Japan). Ammonium acetate (purity > 98.0%) was purchased from Sinopharm Chemical Reagent Co., Ltd. (Shanghai, China). Water was purified using a Milli-Q system (Molsheim, France).

Sample preparation
The *Sophora flavescens* root was dried and ground to pass through a 0.1 mm sieve. 50 mg powder was extracted with 10 mL solvent mixture of methanol/water/formic acid (49:49:2; v/v/v) by sonication for 60 min. An aliquot
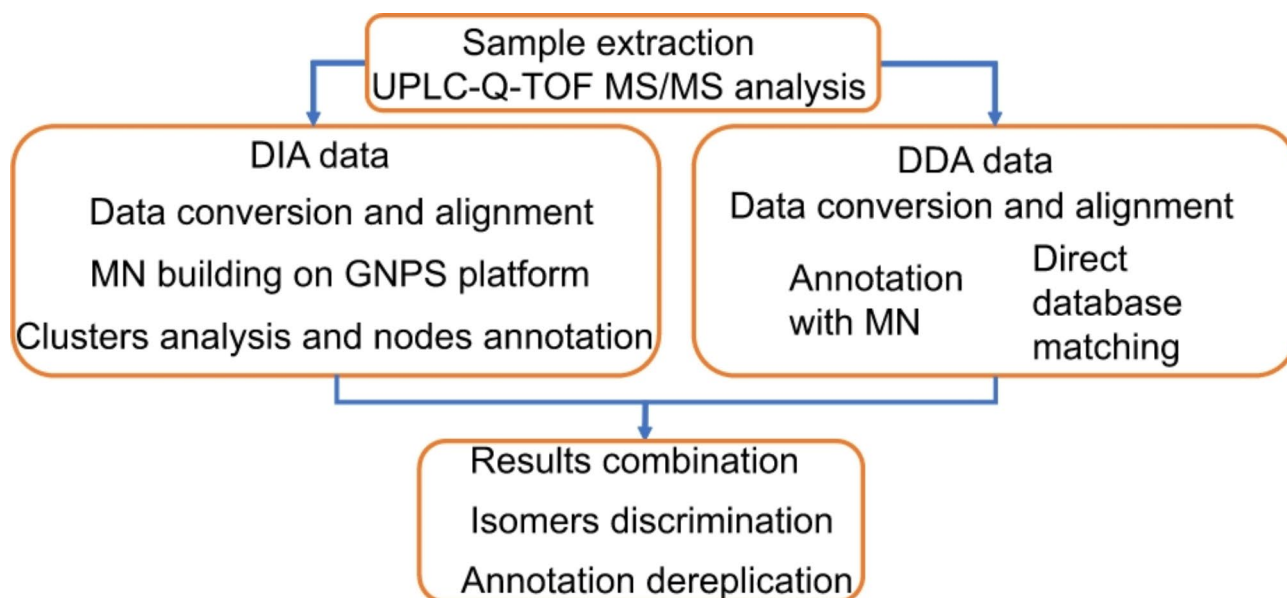
Fig. 1. Schematic workflow of the proposed strategy for dereplication of secondary metabolites in *Sophora flavescens* root.

of extractant was served as blank sample and subjected with the same treatments. After centrifugation, the supernatant was reserved. This extraction was carried out in triplicate. The supernatants were combined and then dried by nitrogen blowing. The dried extract was dissolved in a H_2O/ACN (95:5; v/v) solution. The concentration of *Sophora flavescens* powder in the reconstituted sample solution is 10 mg/mL. The sample solution was filtered through a 0.22 μm polytetrafluoroethylene membrane before testing. The solutions of the standard sample were all prepared at a concentration of 100 ng/mL.

Instrumental experiment procedures

All acquisitions were conducted on an UPLC-Q-TOF system. The system consisted of an Agilent 1290 Infinity LC set (Santa Clara, CA, USA) and an ABSciex triple TOF 5600$^+$ mass spectrometer. A 2.1×150 mm, 1.8 μm ECLIPSE PLUS-C$_{18}$ column (Agilent) was used.

The mobile phase **A** was an ammonium acetate/water solution (8.0 mmol/L). Mobile phase **B** was acetonitrile. The total flow rate was 0.300 mL/min. The column temperature was 40 °C. The LC gradient program was as follows: 3–5% B for 0–3 min, 5–5% for 3–5 min, 5–15% for 5–8 min, 15–60% for 8–12 min, 60–98% for 12–20 min, 98–98% for 20–21 min. The injection volume was 2.0 μL. All samples were analyzed in triplicate.

The mass spectrometer worked in positive mode. The ionization voltage was set at +5.5 kV with 55 psi for the nebulizing gas, 55 psi for the auxiliary gas, and 35 psi for the curtain gas. The source temperature was set at 550 °C. For MS scanning, the TOF scan range covered *m/z* 100–2000.

For DDA analysis, the mass spectrometer was operated in Information Dependent Acquisition (IDA) mode[28]. The mass range of survey scans was set to 100–2000 Da. The top 4 ions were selected for collision induced dissociation. The collision energy was set to 50 eV, the collision energy spread to 10 eV, the ion release delay to 67 ms, and the ion release width to 25 ms.

In order to conduct a DIA analysis, the sequential window acquisition of all theoretical fragment-ion spectra (SWATH)[29] was employed. The precursor ions, spanning the range of 100–1000 Da, were sequentially and cyclically isolated using a 50 Da mass window, and then subjected to fragmentation with a 50 eV collision energy.

Raw data conversion

Two MNs were built for DIA and DDA data separately. The DIA and DDA raw spectra (MassIVE: MSV000096717) were both converted to mzML format data with the software MSConvert (ProteoWizard 3.02).

The MS^2 features in DIA data were then extracted with MS-DIAL (v5.3) to construct the pseudo-MS/MS spectra for MN. The parameter of acquisition type was set to SWATH for imported DIA data. MS^1 tolerance was set at 0.01 Da while MS^2 tolerance at 0.025 Da. For peak detection, the minimum peak height was 50 amplitude and mass slice width was 0.1 Da. The alignment parameters of the triplicate results were set with a 0.1 min RT tolerance and a 0.015 Da MS^1 tolerance. The aligned results were exported to a peak table file and a MS/MS spectral file. The two files were then uploaded to the GNPS platform.

The converted DDA data were processed with MZmine (v4.3.0)[30,31] to perform MS detection, chromatogram building[32], chromatogram resolving, isotope grouping, blank subtraction, and feature alignment. The parameters for isotope grouping were 0.01 m/*z* tolerance and 0.1 min RT tolerance. The alignment of the triplicate results was performed using the parameters of 0.015 m/*z* tolerance, 0.1 min RT tolerance and same charge state. The aligned data were exported to a feature quantification table file and a MS/MS spectral file. The table and spectral file were also uploaded to the GNPS platform.

Molecular networking construction

Both DIA and DDA data based MNs were created using the METABOLOMICS-SNETS-V2 GNPS workflow on GNPS[21]. The data were filtered by removing all MS/MS fragment ions within +/− 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/− 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.5 Da and a MS/MS fragment ion tolerance of 0.5 Da. A network was then created. The cosine score of edges was set to 0.6 or more. The number of matched peaks was set to 4 or more. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS' spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.6 and at least 4 matched peaks.

The molecular networking was further visualized and analyzed using Cytoscape (Ver. 3.10)[33].

Direct spectral matching for DDA data

The converted DDA data were subjected to MS-DIAL (ver. 5.3) to perform direct spectral matching. The MS-DIAL parameters were set as follows: MS^1 accurate mass tolerance at 0.01 Da; MS/MS accurate mass tolerance at 0.1 Da; minimum peak height at 50 amplitude; mass slice width at 0.1 Da. The alignment parameters of the triplicate results were set with a 0.1 min RT tolerance and a 0.015 Da MS^1 tolerance. All identification databases were in MSP format including MSMS-Public_experimentspectra-pos-VS19, MSMS-Pos-GNPS, MoNA-export-Vaniya-Fiehn_Natural_Products_Library, MoNA-export-MassBank, MoNA-export-LC–MS–MS_Positive_Mode, and MoNA-export-GNPS.

Results and discussion

Annotation through DIA data-based MN

Metabolites were assigned with three levels of confidence. Level 1 is the identification. The identification is based on RT and MS/MS spectra that are identical to those of the standard chemical. Level 2 is the annotation. The

annotation corresponds to the metabolites with confident MS/MS spectra. Level 3 is the tentative annotation corresponding to the metabolites with an accurate parent ion mass (< 5 ppm). These three levels comply with classification proposed by Schrimpe-Rutledge and colleagues[34,35]. The compounds name adopted for annotation and tentative annotation was recommended by library hits or literature reports.

The MN (Fig. S1) for DIA data of *Sophora flavescens* samples was built in GNPS (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=eab88b9bca7048cb880c3dc647780ad6).

In cluster 1 (Fig. 2) of the built MN, 7 components (red nodes) were tentatively annotated based on the GNPS library hits. Among them, 6 annotations were achieved at different evidence levels. They are all flavonoids. The information of them is listed in Table 1.

Xanthohumol (**704**) and trifolirhizin (**1041**) were identified through standard material. The MS/MS spectra matching of metabolites and standard were presented in Supplementary List S1 (List S1). The retention time of the metabolites and standard was compared in Supplementary List 2 (List S2).

The presence of kushenol A (**850**) and kushenol I (**1010**) was identified by their precursor ion mass, the fragmentation patterns, and the library spectra match. The mirror plots of library matches are shown in List S1. The node **650** (*m/z* 341.1388) was annotated as 8-Prenylnaringenin. It should be noted that the name of the node **650** was recommended by the library search. Despite the identical MS/MS spectra (List S1- No.1), it can only be regarded as a class-type identification since its molecular structure cannot be characterized with only MS measure. The node **1240** was also a putative annotation at evidence level 2.

In cluster 1, the node **752** was assigned to (2R,3R)-3,7,4'-Trihydroxy-5-methoxy-8-prenylflavanone (*m/z* 371.15) by GNPS library hits. However, there are two components (RT 12.49 min and RT 12.73 min) in the sample that support this assignment. The EIC and MS/MS spectra of them were shown in Fig. S2. Due to the lack of the standard material and reference report, the node **752** was not annotated in this study.

The above identified or annotated components were then employed as seed nodes (at least putatively identified) and allowed the propagation annotation. Totally, 5 components (green nodes in Fig. 2) were annotated with the propagation protocol manually. The information of them is listed in Table 2. Take the kuwanon C (**866**) for example, it was annotated starting from xanthohumol. Its annotation was then corroborated with the accurate precursor ion (experimental *m/z* 423.1805, 0.7 ppm) and MS/MS spectrum.

Fig. 2. The node annotations and propagations in cluster 1 of built MN for DDA data.
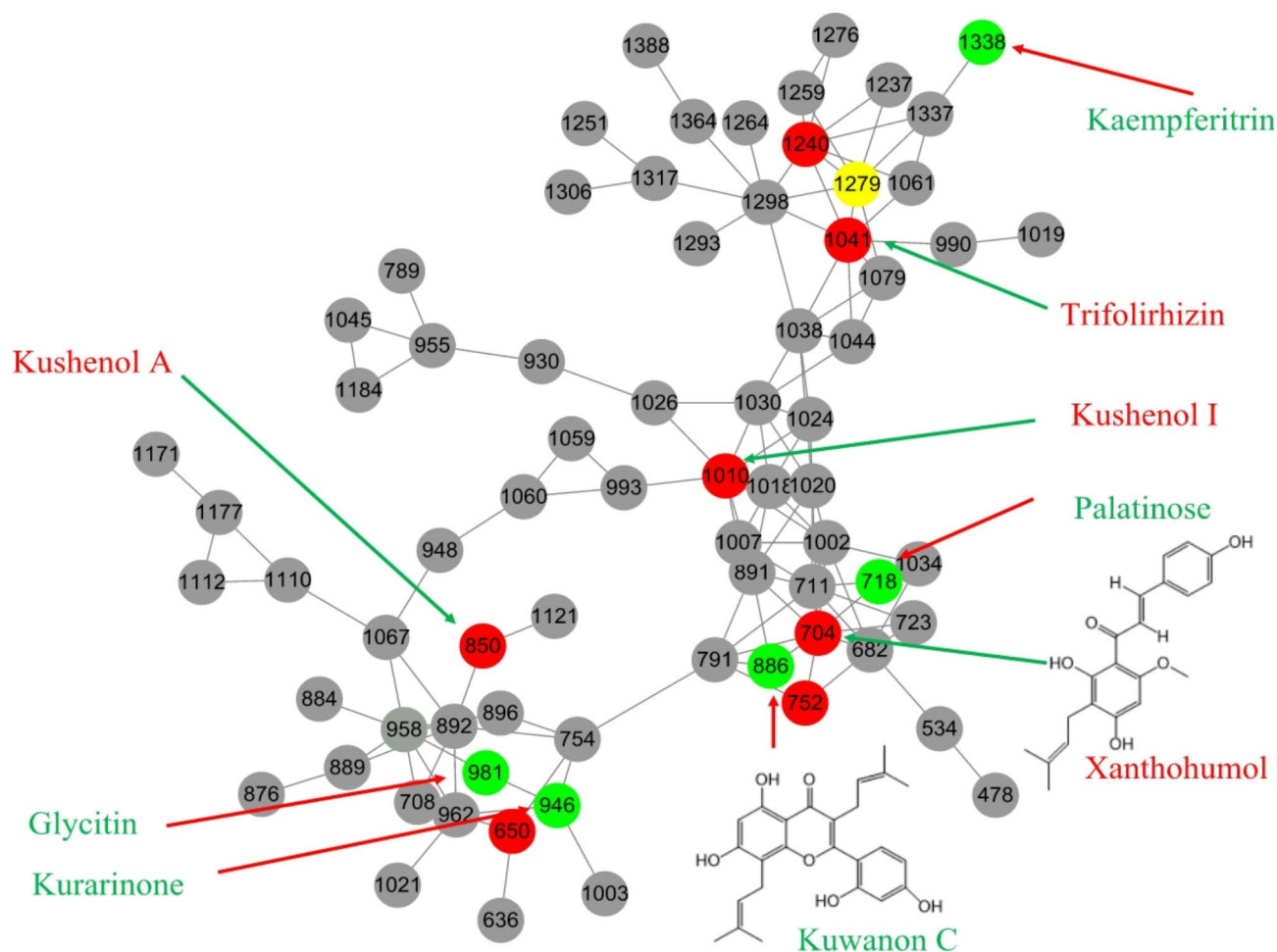
No	USI* no	Compound_Name	RT (min)	Adduct	parent (m/z)	Δm/z (ppm)	Predicted formula	Evidence level
1	650	8-Prenylnaringenin	13.78	[M + H] +	341.1388	1.32	$C_{20}H_{20}O_5$	2
2	704	Xanthohumol	12.97	[M + H] +	355.1545	1.41	$C_{21}H_{22}O_5$	1
3	850	Kushenol A	15.29	[M + H] +	409.2019	2.32	$C_{25}H_{28}O_5$	2
4	1010	Kushenol I	13.32	[M + H] +	455.2071	1.47	$C_{26}H_{30}O_7$	2
5	1041	Trifolirhizin	11.81	[M + NH$_4$] +	464.1560	1.90	$C_{22}H_{22}O_{10}$	1
6	1240	6'-Malonyltrifolirhizin	11.17	[M + H] +	533.1293	0.62	$C_{25}H_{24}O_{13}$	2

Table 1. Annotation in the Cluster 1 of the DIA data MN based on GNPS library hits. *USI (Universal Spectrum Identifier): mzspec:MassIVE:TASK-eab88b9bca7048cb880c3dc647780ad6-spectra/specs_ms.mgf:scan:#

No	USI* no	Compound_Name	RT (min)	Adduct	Parent (m/z)	Δm/z (ppm)	Predicted formula	Evidence level
7	718	Palatinose	1.24	[M + NH4] +	360.1508	2.11	$C_{12}H_{22}O_{11}$	2
8	886	Kuwanon C	12.87	[M + H] +	423.1805	0.70	$C_{25}H_{26}O_6$	2
9	946	Kurarinone	13.57	[M + H] +	439.2121	1.32	$C_{26}H_{30}O_6$	1
10	981	Glycitin	10.57	[M + H] +	447.1298	2.75	$C_{22}H_{22}O_{10}$	3
11	1338	Kaempferitrin	10.53	[M + H] +	579.1714	0.98	$C_{27}H_{30}O_{14}$	3

Table 2. Propagation annotation in the Cluster 1 of the DIA data MN. *USI (Universal Spectrum Identifier): mzspec:MassIVE:TASK-eab88b9bca7048cb880c3dc647780ad6-spectra/specs_ms.mgf:scan:#

No	USI* no	Compound_Name	RT (min)	Adduct	Parent (m/z)	Δm/z (ppm)	Predicted formula	Evidence level
12	104	Sophocarpine	10.81	[M + H] +	247.1804	0.36	$C_{15}H_{22}N_2O$	2
13	110	Sophoridine	10.41	[M + H] +	249.1962	0.24	$C_{15}H_{24}N_2O$	1
14	197	Oxymatrine	8.82	[M + H] +	265.1914	1.32	$C_{15}H_{24}N_2O_2$	2
15	1151	Bryodulcosigenin	17.47	[M + Na] +	497.3601	0.06	$C_{30}H_{50}O_4$	2
16	1249	Nomilin	5.16	[M + Na] +	537.2089	1.12	$C_{28}H_{34}O_9$	2
17	92	Sophoramine	10.73	[M + H] +	245.1650	0.65	$C_{15}H_{20}N_2O$	1
18	181	mamanine	4.70	[M + H] +	263.1757	1.14	$C_{15}H_{22}N_2O_2$	3
19	200	Oxysophoridine	7.46	[M + H] +	265.1913	0.94	$C_{15}H_{24}N_2O_2$	2
20	487	Lycorine	4.80	[M + NH$_4$] +	305.1497	0.39	$C_{16}H_{17}NO_4$	2
21	151	isoliquiritigenin	12.74	[M + H] +	257.0801	2.88	$C_{15}H_{12}O_4$	2
22	254	Formononetin	13.00	[M + H] +	269.0812	1.34	$C_{16}H_{12}O_4$	2
23	375	Maackiain	11.80	[M + H] +	285.0762	1.58	$C_{16}H_{12}O_5$	3
24	1192	Dehydroeburicoic acid monoacetate	10.78	[M + H] +	511.3771	2.13	$C_{33}H_{50}O_4$	3

Table 3. Annotation in the Cluster 2 of the DIA data MN. *USI (Universal Spectrum Identifier): mzspec:MassIVE:TASK-eab88b9bca7048cb880c3dc647780ad6-spectra/specs_ms.mgf:scan:#

In cluster 2 (Fig. S3), the alkaloids of sophocarpine (**104**), sophoridine (**110**), and oxymatrine (**197**) were annotated based on the GNPS library hits. The triterpenoids of bryodulcosigenin (**1151**) and nomilin (**1249**) were annotated as well. Since *Sophora flavescens* contains the alkaloids from the same metabolic pathway, it is easy to propagate the annotation between them. For example, the sophoramine (**92**) was annotated next to sophocarpine as shown in Fig. S3. The propagation is based on that sophoramine is actually the 11,12-dehydrogenated sophocarpine. Along this protocol, other alkaloids including oxysophoridine (**200**), lycorine (**487**), isoliquiritigenin (**151**), formononetin (**254**), maackiain (**375**) and dehydroeburicoic acid monoacetate (**1192**) were annotated. Due to the lack of reference library spectra, mamanine (**181**) was annotated tentatively. The information about these annotations is listed in Table 3.

In other clusters and self-loop nodes, 3 metabolites were annotated based on GNPS library hits. No further propagated annotation was achieved. The information of 3 annotations is listed in Table 4. Calycosin (**372**) was hit by library search as a self-loop node. Sophoflavescenol (**745**) was connected to an unresolved node. Calycosin-7-O-beta-D-glucoside (**980**) was identified with the standard chemical.

There are lots of nodes that were not annotated. For example, cluster 4 comprises 8 nodes but none was even putatively annotated. Although it is not necessary, it always starts from a known node in MN to propagate the annotation and structure relationship[36]. Therefore, cluster 4 and 6 were not analyzed directly.

The GNPS library hit was tentative assignment but not annotation or identification. Without standard chemicals, tentative library hits could not achieve level 1 identification in this study.

No	USI* no	Compound_Name	RT (min)	Adduct	Parent (*m/z*)	Δ*m/z* (ppm)	Predicted formula	Evidence level
25	372	Calycosin	11.97	[M + H] +	285.0758	0.18	$C_{16}H_{12}O_5$	2
26	745	Sophoflavescenol	13.15	[M + H] +	369.1336	0.89	$C_{21}H_{20}O_6$	2
27	980	Calycosin-7-O-beta-D-glucoside	10.72	[M + H] +	447.1291	1.19	$C_{22}H_{22}O_{10}$	1

Table 4. Annotation in the other clusters and self-loop nodes of the DIA data MN. *USI (Universal Spectrum Identifier): mzspec:MassIVE:TASK-eab88b9bca7048cb880c3dc647780ad6-spectra/specs_ms.mgf:scan:#

No	USI* no	Compound_Name	RT (min)	Adduct	Parent (*m/z*)	Δ*m/z* (ppm)	Predicted formula	Evidence level
28	3	Matrine	8.05	[M + H] +	249.1955	2.57	$C_{15}H_{24}N_2O$	1

Table 5. Annotation from the DDA data-based MN that unobserved in DIA data-based MN. *USI (Universal Spectrum Identifier): mzspec:MassIVE:TASK-88ffe047b7ff4baea8fc407d41628752-spectra/specs_ms.mgf:scan:#

DDA data-based MN

In contrast to the DIA data-based MN, a MN based on DDA data was also constructed using the same parameters. The built DDA data-based MN (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=88ffe047b7ff4baea8fc407d41628752) was shown in Fig. S4.

Twelve GNPS library hits were achieved as listed in Table S1. They are 2 alkaloids, 5 flavonoids, 4 isoflavonoids, and 1 aromatic polyketide. Among them, sophocarpine, matrine, calycosin, kurarinone, kushenol I, and trifolirhizin were annotated. As a complementary annotation to DIA data-based MN, matrine (**3**) was further identified with its standard chemical. The identification information of matrine is listed in Table 5.

The compound matrine was not annotated in DIA data-based MN, although it is a main ingredient of *Sophora flavescens*. The possible reason is the feature lost during DIA data process. However, the DIA based MN annotated more compounds which DDA did not achieve. It is beneficial to perform both DDA and DIA data-based MN to dereplicate the metabolites in *Sophora flavescens*.

Direct DB matching

The primary method for identifying natural compounds is to directly match their MS/MS spectra with library spectra. However, DIA spectra are inherently complex, containing uncertain precursor ions. As a result, they cannot be used to match libraries directly. DDA data are well suited for direct spectral matching because each spectrum in a DDA result is an MS/MS spectrum of a single precursor ion.

Complementary to the MN approach, the DDA data of *Sophora flavescens* were subjected to matching the offline database. When the match score was set to 1.4 or more, 18 metabolites were annotated (Table 6) that did not get annotated in both MN approaches. These annotations showed that DDA is the necessary technological approach to achieve dereplication.

Isomer discrimination (LC/MS)

There are lots of isomers in the constituents of *Sophora flavescens*. In order to distinguish the isomers, the chromatographic method in this study was optimized elaborately based on previous work. The LC–MS was then performed in total scanning acquisition mode. The total ion chromatogram (TIC) was shown in Fig. S5. The zoomed TIC between 0 and 7 min was shown in Fig. S6. The EIC of a particular isomer was extracted from the TIC. The Peaks in EIC were analyzed by the feature of RT and accurate *m/z* value. The MS/MS spectrum of the peak was extracted from the DIA data at the corresponding RT. The MS/MS spectrum was then checked with fragmentation patterns, the molecular polarity, and literature reference.

Six matrine isomers (*m/z* 249.1961) were observed in the EIC (Fig. 3). By referencing the RT of the standard chemicals, the peaks at 8.053 min and 10.413 min were identified as matrine and sophoridine, respectively. The remaining isomers were putatively annotated according to their MS/MS patterns in reference databases and in the reported literature[37]. The peaks at 6.643 min, 8.559 min, 9.304 min, and 9.669 min were annotated as (+)-lupanine, α-isolupanine[37], isomatrine[38], and allomatrine[39] respectively. Details are listed in Table S2.

The isomers of sophoramine (*m/z* 245.1648) were annotated through the EIC. After the MS/MS spectrum searching in the database and fragmentation pattern analysis, 3 components (Table S3) at 8.204, 10.727 and 11.268 min were annotated as anagyrine, sophoramine, and neosophoramine, respectively[40,41]. They were all identified with standard chemicals. The MS/MS spectra and retention time matching of all metabolites validated by standards were presented in List S1 and List S2.

Out of all the annotated isomers of matrine and sophoramine, there were 5 metabolites (Table 7) that were not annotated by both the MN and the direct DB matching approaches.

Compared to previous works[38,42,43], more isomers were separated and identified with the strategy proposed in this study. It should be noted that not all the observed isomers are annotated. Those unidentified isomers could be the new candidates for future metabolomics research.

No	Compound_Name	RT (min)	Adduct	Parent (*m/z*)	Δ*m/z* (ppm)	Predicted formula	Evidence level
29	Kuraramine	1.34	[M + H] +	223.144	0.45	$C_{12}H_{18}N_2O_2$	3
30	lupinine	1.96	[M + H] +	170.1539	0.24	$C_{10}H_{19}NO$	2
31	cytisine	3.34	[M + H] +	191.1179	0.05	$C_{11}H_{14}N_2O$	2
32	Adenosine	4.37	[M + H] +	268.1043	1.01	$C_{10}H_{13}N_5O_4$	2
33	Lupanine	6.64	[M + H] +	249.1956	2.17	$C_{15}H_{24}N_2O$	3
34	Oxysophocarpine	9.18	[M + H] +	263.1757	1.14	$C_{15}H_{22}N_2O_2$	1
35	baptifoline	9.60	[M + H] +	261.1599	0.57	$C_{15}H_{20}N_2O_2$	2
36	N-methylcytisine	9.85	[M + H] +	205.1334	0.68	$C_{12}H_{16}N_2O$	2
37	isosophocarpine	10.00	[M + H] +	247.1801	1.58	$C_{15}H_{22}N_2O$	3
38	Isosakuranin	10.19	[M + H] +	449.1432	2.27	$C_{22}H_{24}O_{10}$	3
39	Apiin	10.66	[M + H] +	565.1558	1.1	$C_{26}H_{28}O_{14}$	2
40	Sophoricoside	10.91	[M + H] +	433.1137	1.85	$C_{21}H_{20}O_{10}$	2
41	meso-Zeaxanthin	10.94	[M + H] +	569.4363	1.74	$C_{40}H_{56}O_2$	3
42	Ononin	11.54	[M + H] +	431.1343	1.48	$C_{22}H_{22}O_9$	2
43	Sparteine	11.65	[M + H] +	235.2170	0.51	$C_{15}H_{26}N_2$	2
44	Quebrachamine	13.36	[M + H] +	283.2170	0.42	$C_{19}H_{26}N_2$	2
45	8-Prenylkaempferol	13.87	[M + H] +	355.1181	1.35	$C_{20}H_{18}O_6$	2
46	Isoxanthohumol	15.17	[M + H] +	355.1541	0.28	$C_{21}H_{22}O_5$	1

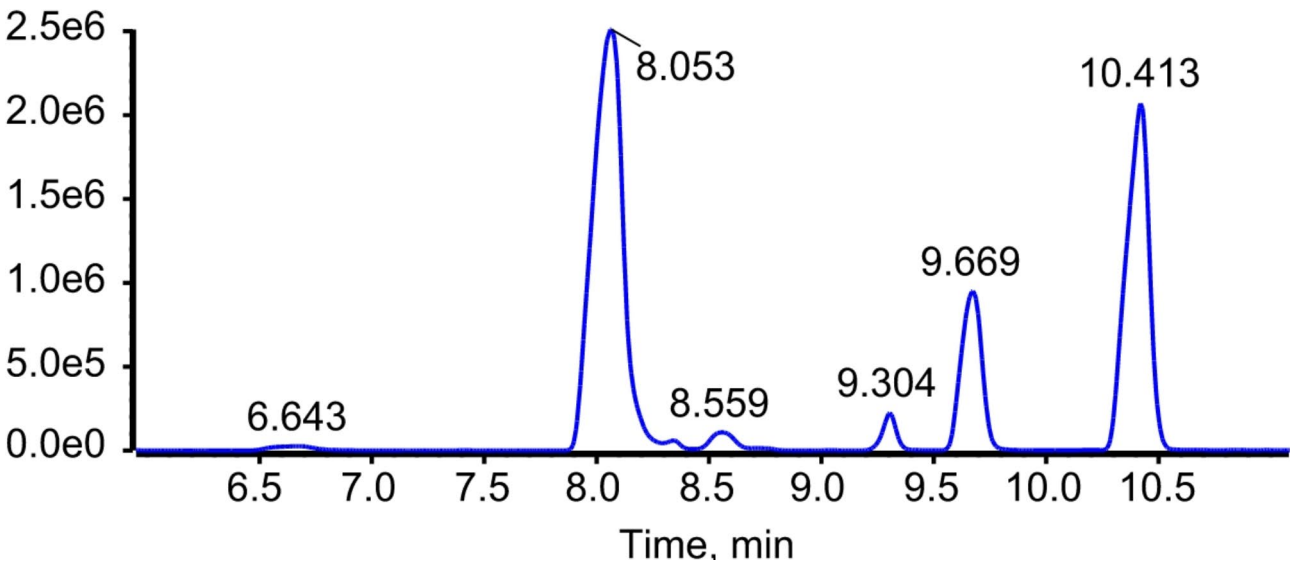**Table 6**. Annotation with offline DB matching for the DDA data.

Fig. 3. Extracted ion chromatogram (EIC) for the discrimination of isomers (*m/z* 249.1961) in *Sophora flavescens*.

No	Compound_Name	RT (min)	Adduct	Parent (*m/z*)	Δ*m/z* (ppm)	Predicted formula	Evidence level
47	α-isolupanine	8.56	[M + H] +	249.1961	0.16	$C_{15}H_{24}N_2O$	2
48	Isomatrine	9.30	[M + H] +	249.1962	0.24	$C_{15}H_{24}N_2O$	2
49	allomatrine	9.67	[M + H] +	249.1959	0.96	$C_{15}H_{24}N_2O$	2
50	anagyrine	8.20	[M + H] +	245.1649	0.24	$C_{15}H_{20}N_2O$	1
51	neosophoramine	11.27	[M + H] +	245.1648	0.16	$C_{15}H_{20}N_2O$	1

Table 7. Complementary annotation with the isomer discrimination.

Quantitation

Although the focus of this study is the dereplication strategy, the quantitation of the identified compounds is always the concern. A simple external standard method was applied to quantify the contents of matrine and oxysophocarpine. The peak area in the EIC was calculated. The calibration curves of them are shown in Fig. S7. As a primary evaluation measure, the quantification method was not rigorously validated for specificity, limit of detection, linear range, and precision. According to the external standard results, the content of matrine in the dried root of *Sophora flavescens* was about 6.05 mg/g. The content of oxysophocarpine was about 5.81 mg/g.

Semi-quantification for all the identified compounds was estimated based on the hypothesis that the peak intensity is proportional to the number of molecules. The peak heights of the annotated compounds in their EIC were recorded in Table S4 for the *Sophora flavescens* sample at a concentration of 10 mg/mL. It shows that the most abundant compound of the tested sample was oxymatrine. Other abundant alkaloids are sophoridine and matrine. The most abundant flavonoids are kurarinone and xanthohumol. It was noteworthy that for the 1.0 mg/g sample, compounds such as calycosin, lycorine, and nomilin were not annotated in both the MN and direct DB matching approaches.

It was found that the MN approach annotated some trace compounds that could not be annotated by direct DB matching. Typical trace compounds such as glycitin, sophoricoside, and lycorine were all annotated through the DIA data-based MN. It proves that MN on GNPS can overcome the challenges of trace compound identification compared to direct DB matching.

Conclusion

The present study proposed a dereplication strategy for the screening of secondary metabolites in *Sophora flavescens*. The UPLC-Q-TOF-MS/MS analysis in both DDA and DIA mode was performed firstly. DIA and DDA data were then used to build a MN for the propagation annotation. The DDA data were directly matched to the offline DB to obtain complementary annotations. Isomers were annotated and distinguished by their retention time difference. Through a combination of GNPS MN annotation, direct DB matching and isomer discrimination, 51 compounds were annotated and dereplicated at 3 evidence levels. The results provided a closer and more comprehensive understanding of the chemical constitution of *Sophora flavescens*. This dereplication strategy can be used as a general approach for the plant metabolomics research.

Data availability

Data are available from the corresponding author upon request.

References

1. Wu, Z., Cheng, S., Wang, S., Li, W. & Liu, J. Biological evaluation of sophoridine by interaction with plasma transport protein and protective effects in cardiomyocytes. *Arab. J. Chem.* **17**, 106014 (2024).
2. Zhou, Y. J., Guo, Y. J., Yang, X. L. & Ou, Z. L. Anti-cervical cancer role of matrine, oxymatrine and sophora flavescens alkaloid gels and its mechanism. *J. Cancer* **9**, 1357–1364 (2018).
3. Zhou, H., Lutterodt, H., Cheng, Z. & Yu, L. (Lucy). Anti-Inflammatory and Antiproliferative Activities of Trifolirhizin, a Flavonoid from *Sophora flavescens* Roots. *J. Agric. Food Chem.* **57**, 4580–4585 (2009).
4. Kwon, M. et al. Inhibitory effects of flavonoids isolated from *Sophora flavescens* on indoleamine 2,3-dioxygenase 1 activity. *J. Enzyme Inhibition Med. Chem.* **34**, 1481–1488 (2019).
5. Nishikawa, S. et al. Anti-inflammatory activity of Kurarinone involves induction of HO-1 via the KEAP1/Nrf2 pathway. *Antioxidants* **9**, 842 (2020).
6. Kim, B.-H. & Lee, S. Sophoricoside from Sophora japonica ameliorates allergic asthma by preventing mast cell activation and CD4+ T cell differentiation in ovalbumin-induced mice. *Biomed. Pharmacother.* **133**, 111029 (2021).
7. Li, J. et al. Two new isoprenoid flavonoids from sophora flavescens with antioxidant and cytotoxic activities. *Molecules* **26**, 7228 (2021).
8. Yang, Y.-F. et al. Flavonoids from the roots of sophora flavescens and their potential anti-inflammatory and antiproliferative activities. *Molecules* **28**, 2048 (2023).
9. Lu, K. et al. Five novel pterocarpan derivatives from *Sophora flavescens*. *Chin. J. Chem.* **39**, 2763–2768 (2021).
10. Quiros-Guerrero, L.-M. et al. Comprehensive mass spectrometric metabolomic profiling of a chemically diverse collection of plants of the Celastraceae family. *Sci Data* **11**, 415 (2024).
11. Zhang, A., Sun, H. & Wang, X. Mass spectrometry-driven drug discovery for development of herbal medicine. *Mass Spectrom. Rev.* **37**, 307–320 (2018).
12. Tsugawa, H., Rai, A., Saito, K. & Nakabayashi, R. Metabolomics and complementary techniques to investigate the plant phytochemical cosmos. *Nat Prod Rep* **38**, 1729–1759 (2021).
13. Guo, J. & Huan, T. Comparison of full-scan, data-dependent, and data-independent acquisition modes in liquid chromatography-mass spectrometry based untargeted metabolomics. *Anal. Chem.* **92**, 8072–8080 (2020).
14. Carriot, N. et al. Integration of LC/MS-based molecular networking and classical phytochemical approach allows in-depth annotation of the metabolome of non-model organisms - The case study of the brown seaweed Taonia atomaria. *Talanta* **225**, 121925 (2021).
15. Mohimani, H. et al. Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications* **9**, 4035 (2018).
16. Scheubert, K. et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun* **8**, 1494 (2017).
17. Cai, Y., Zhou, Z. & Zhu, Z.-J. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *TrAC Trends Anal. Chem.* **158**, 116903 (2023).
18. De Jonge, N. F. et al. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics* **18**, 103 (2022).
19. Gao, Y. et al. Molecular networking as a dereplication strategy for monitoring metabolites of natural product treated cancer cells. *Rapid Comm Mass Spectrom.* **34**, e8549 (2020).

20. Aron, A. T. et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature Protocols* **15**, 1954–1991 (2020).
21. Wang, M. et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* **34**, 828–837 (2016).
22. Shen, X. et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun* **10**, 1516 (2019).
23. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci USA* **109**, E1743-1752 (2012).
24. Gurevich, A. et al. Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **3**, 319–327 (2018).
25. Pakkir Shah, A. K. et al. Statistical analysis of feature-based molecular networking results from non-targeted metabolomics data. *Nat Protoc* https://doi.org/10.1038/s41596-024-01046-3 (2024).
26. Pakkir Shah, A. K. *et al.* Statistical analysis of feature-based molecular networking results from non-targeted metabolomics data. *Nature Protocols* **20**, 92–162 (2025).
27. Zhou, Z. et al. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun* **13**, 6656 (2022).
28. Zhang, H. et al. Arabidopsis proteome and the mass spectral assay library. *Sci. Data* **6**, 278 (2019).
29. Midha, M. K. et al. A comprehensive spectral assay library to quantify the *Escherichia coli* proteome by DIA/SWATH-MS. *Sci. Data* **7**, 389 (2020).
30. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **11**, 395 (2010).
31. Heuckeroth, S. et al. Reproducible mass spectrometry data processing and compound annotation in MZmine 3. *Nat Protoc* **19**, 2597–2641 (2024).
32. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: New algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. *Anal Chem* **89**, 8696–8703 (2017).
33. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* **27**, 431–432 (2011).
34. Godzien, J., Gil De La Fuente, A., Otero, A. & Barbas, C. Metabolite Annotation and Identification. in *Comprehensive Analytical Chemistry* vol. **82**, 415–445 (Elsevier, 2018).
35. Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D. & McLean, J. A. Untargeted metabolomics strategies—challenges and emerging directions. *J. Am. Soc. Mass Spectrom.* **27**, 1897–1905 (2016).
36. Yang, J. Y. et al. Molecular networking as a dereplication strategy. *J Nat Prod* **76**, 1686–1699 (2013).
37. Liu, G., Dong, J., Wang, H., Hashi, Y. & Chen, S. Characterization of alkaloids in Sophora flavescens Ait. by high-performance liquid chromatography–electrospray ionization tandem mass spectrometry. *J. Pharm. Biomed. Anal.* **54**, 1065–1072 (2011).
38. Liu, X.-J. et al. Alkaloids from Sophora flavescens Aition. *Fitoterapia* **81**, 524–527 (2010).
39. Zhang, Y.-B. et al. Matrine-type alkaloids from the roots of sophora flavescens and their antiviral activities against the hepatitis B virus. *J Nat Prod* **81**, 2259–2265 (2018).
40. Wu, Z.-J. et al. Analysis of matrine-type alkaloids using ESI-QTOF. *Int. J. Mass Spectrom.* **341–342**, 28–33 (2013).
41. Li, J.-C. et al. Quinolizidine alkaloids from Sophora alopecuroides with anti-inflammatory and anti-tumor properties. *Bioorganic Chem.* **110**, 104781 (2021).
42. Wang, H. & Ding, H. Screening of Alkaloids in Sophora Flavescens with liquid chromatography tandem data-dependent acquisition mass spectrometry. *J. Pharm. Res. Int.* https://doi.org/10.9734/jpri/2023/v35i317465 (2023).
43. Dong, Y. et al. Determination of Alkaloids and Flavonoids in Sophora flavescens by UHPLC-Q-TOF/MS. *J Anal Methods Chem* **2021**, 9915027 (2021).

Acknowledgements

Author contributions

H. Wang and H. Ding did the same contribution including conceptualization, investigation, methodology, resources, experiment, software, writing—original draft; and writing—review & editing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-94958-3.

Correspondence and requests for materials should be addressed to H.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.