scientific data

OPEN

DATA DESCRIPTOR

Mass Spectral Data of Primary and Secondary Metabolites Changes in Medicinal Plants by Solvent Polarity

Jinyoung Park, Jiho Lee, Chaeyeon Cho, Dong Kyo Lee, Woohyun Song, Kyeongwon Choi, Seong Yeon Choi & Heejung Yang ✉

Solvent polarity is a critical factor in metabolite extraction from plants, influencing the recovery of primary and secondary metabolites. Plants contain not only the primary metabolites but also a variety of secondary metabolites with low polarity, which favor a higher proportion of organic solvents. In this repository, 248 representative Korean medicinal plants were examined, and a total of 744 samples were prepared using water and ethanol as extraction solvents with three different polarities (100% water, 50% ethanol, and 100% ethanol). After the feature extraction of the scans corresponding to single compounds, 63,944 scans in the positive mode and 42,481 in the negative mode were subjected to in silico chemical class annotation tools. The chemical taxonomic types were summarized in the datasets, which will provide the significance of the proper solvent selection in plant metabolome research.

Background & Summary

Plant extracts have long been utilized in traditional medicinal practices across both Eastern and Western countries for centuries. The research and development of pharmaceuticals derived from medicinal plant extract continue to be an area of active and rigorous investigation. Several medicinal plant extract-derived drugs, such as Veregen and Fulyzaq, have been approved by the U.S. Food and Drug Administration (FDA)[1,2]. In 2023, FDA approved Filsuvez, a drug primarily formulated with birch triterpenes extracted from birch trees, for the treatment of epidermolysis bullosa, a rare and debilitating skin condition[3]. Accordingly, plant extracts have been actively used in the development of new pharmaceutical products, serving as valuable sources of diverse bioactive compounds that can be leveraged to create novel therapeutic agents.

Water has been the most common solvent for producing medicinal extracts from plants. In the past, traditional methods such as heating or long-term soaking in water were commonly used to create extracts from medicinal plants for various therapeutic purposes. Water-based extraction methods allowed for the extraction and concentration of water-soluble compounds from the plant material, providing the basis for many traditional herbal remedies and medicines. Water remains the safest extraction solvent used worldwide in the traditional medicine and new natural product drug development. However, its ability to extract compounds of varying polarity is limited. Therefore, relatively harmless organic solvents are currently used alone or in combination with water for extraction purposes[4]. In the modern era, the availability of pure organic solvents has enabled the extraction of a broader range of metabolites. Due to their lower polarity compared to water, organic solvents can simultaneously extract a wide range of organic compounds. The International Conference on Harmonization (ICH) has published a list of organic solvents that have been determined to be suitable for use in pharmaceutical and dietary supplement formulations, based on an assessment of their relative hazard levels to human health[5]. The choice of solvent is critical in extract production but is often made conventionally. Highly polar compounds are most effectively extracted using water alone or water mixed with a low ratio of organic solvent, whereas compounds with lower polarity are better extracted using organic solvents in higher proportions.

Untargeted metabolomics based on mass spectrometry techniques is widely recognized as the most appropriate approach for comprehensive analysis of metabolites, enabling the efficient evaluation of quantitative and qualitative alterations in individual metabolites within sample sets[6,7]. Qualitative analysis of metabolites serves as the foundation of untargeted metabolomics studies, but it also represents the most challenging aspect of this approach. Although matching metabolites to extensive reference libraries has become more feasible, studies

Department of Pharmacy, Kangwon National University, Gangwondaehak-gil 1, Chuncheon, Gangwon, 24341, South Korea. ✉e-mail: heejyang@kangwon.ac.kr

have reported that less than 5% of the total data can be reliably identified through this approach[8]. Recent in silico approaches employing deep learning techniques have demonstrated high reliability in annotating the structures or classes of compounds in natural products. An alternative method, such as the molecular networking approach in the GNPS platform, helps overcome the limitations of individual library matching[9]. This approach facilitates the visualization of structural relationships among compounds with similar MS/MS fragmentation patterns, enabling the propagation of known annotations to structurally similar unknown derivatives and enhancing the reliability of in silico annotation.

While various studies have reported changes in plant metabolite profiles depending on the choice of extraction solvent, these analyses have typically focused on individual species[10–14]. In the study, we acquired 744 samples from 248 medicinal plants using three different solvent extraction systems (100% water, 50% ethanol and 100% ethanol) and measured their metabolites under the positive and negative modes. We employed in silico annotation and molecular networking approaches to comprehensively analyze the changes in metabolite profiles in response to variations in solvent polarity related to their chemical taxonomy. These datasets will facilitate the investigation the biological implications of solvent polarity, with the goal of optimizing the utilization of medicinal plants as pharmacological resources.

Methods

Sample preparation. A total of 248 dried medicinal plant samples were obtained from various traditional medicinal suppliers in Korea, with associated metadata available as an xlsx file on Figshare[15]. The medicinal parts used varied across the samples, including aerial parts, barks, bulbs, flowers, fruits, galls (insect- or pathogen-induced abnormal plant tissue), heartwood, leaves, pistils, pollen, resin, roots, sap, sclerotium, seeds, stems, thorns, tubers, twigs, and whole plants. All samples were prepared in dried form and ground into coarse powders using a blender, stored in the fridge before the sample extraction.

Sample extraction. To evaluate the extraction efficiency across a range of solvent polarities, three solvents generally regarded as safe—100% water, 50% ethanol, and 100% ethanol— were selected. Food grade ethanol ($\geq$95%, Samchun Chemicals, Korea) was used, and distilled water was used as the aqueous solvent. For extraction, 1 g of each powdered sample was accurately weighed and mixed with 30 mL of the respective solvent containing the internal standard (IS1: 1 µM sulfamethazine). The mixture was subjected to ultrasonic extraction at 25 °C for 3 hours. Following extraction, the solution was filtered using filter paper to remove solid residues. An aliquot (500 µL) of the clear filtrate was dried using a speed vacuum concentrator. The dried extracts were reconstituted in 50% methanol containing the internal standard (IS2: 1 µM sulfadimethoxine for metabolomic analyses) to a final concentration of 500 ppm. The reconstituted extracts were filtered through a regenerated cellulose (RC) syringe filter (0.22 µm, 13 mm, ABLUO®, GVS, Italy) before analysis.

UHPLC-MS analytical methods. The liquid chromatography system used in this study was a Vanquish Flex UHPLC system (Thermo Fisher Scientific, Waltham, MA, USA). Chromatographic separation was achieved using an ACQUITY UPLC BEH C18 column (50 × 2.1 mm, 1.7 µm; Waters, USA). The mobile phases consisted of (A) distilled water with 0.1% formic acid and (B) acetonitrile with 0.1% formic acid. The gradient program began with 10% (B) held for 0.5 minutes, followed by a linear increase to 90% (B) over 14.5 minutes, which was then maintained for 2.5 minutes. Subsequently, the mobile phase composition was rapidly returned to 10% (B) and held for 2.5 minutes. The injection volume was 5 µL, and the flow rate was set to 0.3 mL/min. The autosampler and column temperatures were maintained at 4 and 25 °C, respectively. The UHPLC system was coupled to a Thermo Fisher Orbitrap Exploris120 mass spectrometer (Bremen, Germany) equipped with a Heated Electrospray Ionization (H-ESI) source. Data acquisition was performed in both positive and negative ion modes. In the positive ion mode, an ionization voltage of 3500 V was applied, while in the negative ion mode, an ionization voltage of 2500 V was used. The gas flows were set to sheath gas: 50 Arb, auxiliary gas: 10 Arb, and sweep gas: 1 Arb. The ion transfer tube and vaporizer temperatures were maintained at 325 °C and 350 °C, respectively. Full-scan MS data were acquired in a data dependent acquisition (DDA) mode, with a scan range of 50–1500 m/z. For MS/MS analysis, higher-energy collisional dissociation (HCD) was used with stepped collision energies of 15%, 30%, and 60%.

Data processing and analysis. Raw data obtained from UHPLC-MS/MS analysis were converted to the mzML format using MSConvert software (ver. 3.0.2)[16]. Subsequently, data processing for the feature extraction was conducted using MZmine (ver. 3.9.0)[17]. The parameters of MZmine are available at https://github.com/heejyang/MediHerb. Within the MZmine software, the noise thresholds for MS1 and MS2 feature extraction were set at 1.0×10^4 (5.0×10^4 for the negative mode) and 2.0×10^3 (2.0×10^3 for the negative mode), respectively. The ADAP chromatogram builder was used to generate the chromatogram with the following parameters: a minimum group size of 7 scans, a minimum group intensity of 1.5×10^5, and a minimum absolute intensity of 5.0×10^5, with an *m/z* tolerance of 5 ppm. Chromatogram smoothing was performed using the Loess smoothing algorithm. Chromatographic deconvolution was carried out using the local minimum feature resolver, with a minimum search range of 0.05 min. For MS1 deconvolution, the minimum absolute height was set at 5.0×10^5, the peak area coefficient was 1.50, and a minimum scan of 5 scan points was required. Isotope grouping was performed using the 13 C isotope filter module, applying an m/z and RT tolerance of 3.5 ppm and 0.01 minutes, respectively. The peaks for each sample were aligned based on *m/z* and RT tolerances of 5 ppm and 0.08 min, with weights set at 1:1 for m/z and RT using the Join aligner feature. To minimize cases where features split into multiple nodes due to different adducts, the ion identity networking method was applied with m/z tolerance of 3.5 ppm[18]. For the positive mode, the adducts included $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, and $[M+2H]^+$, while no specific adducts except for $[M\text{-}H]^-$ were set for the negative mode. The peaks processed from each sample were filtered to the peaks list present in a blank (50% MeOH or 100% MeOH), which was prepared in the same

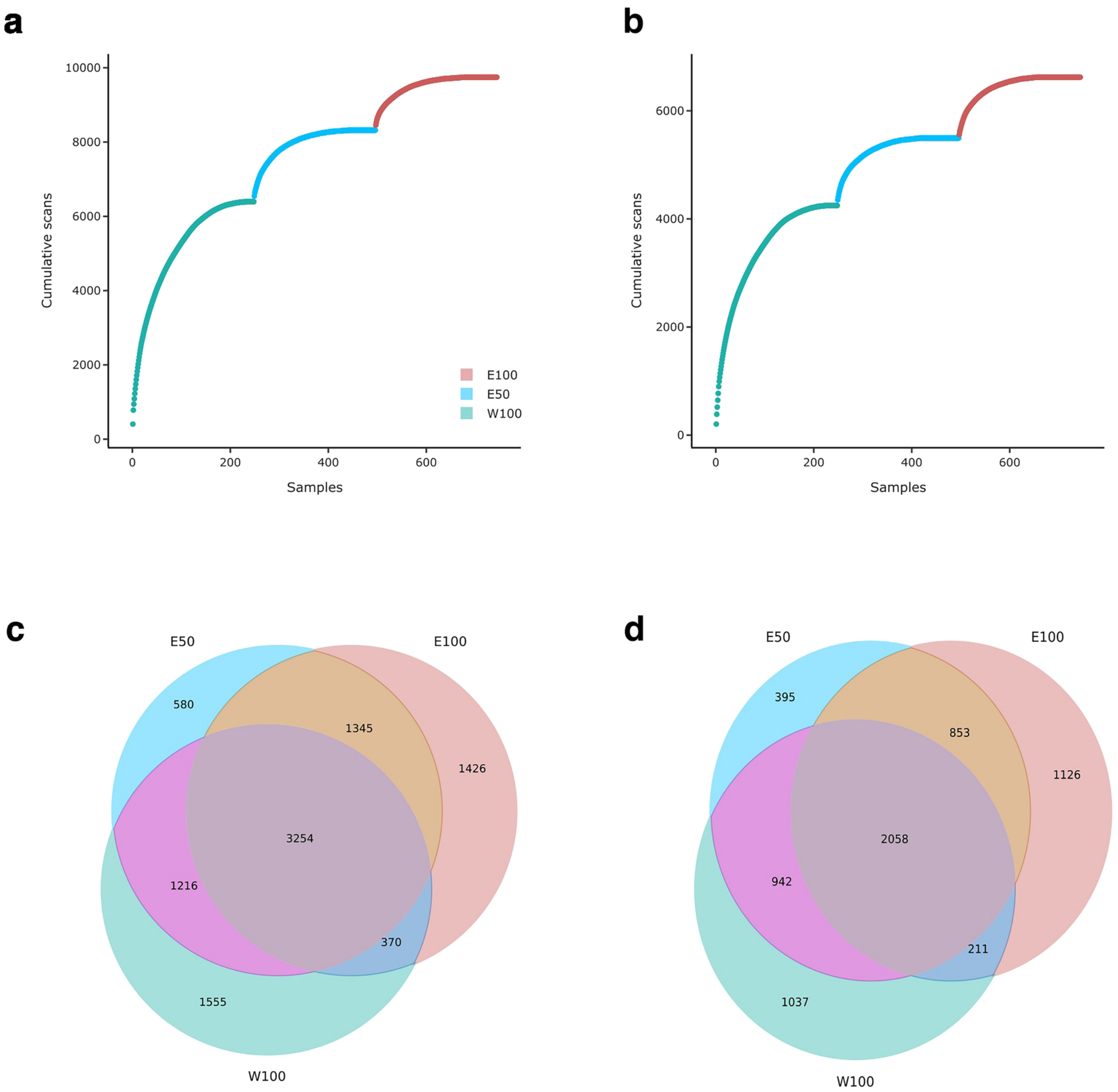

Fig. 1 Influence on molecular diversity by the polarity of extraction solvents. (**a**) The cumulative curve for scans of unique MS/MS spectra added by each solvent in the positive ion mode. (**b**) The cumulative curve for scans of unique MS/MS spectra added by each solvent in the negative ion mode. (**c**) Venn diagram for MS/MS spectral scans with top 10% intensity from three extraction solvents in the positive ion mode. (**d**) Venn diagram for MS/MS spectral scans with top 10% intensity from three extraction solvents in the negative ion mode.

method as the sample, using the feature list blank subtraction feature, the remaining peaks were analyzed using the following molecular networking workflow. A filtered list containing the *m/z* and their intensities of 63,944 scans in the positive mode and 42,481 in the negative mode, respectively, was exported as mgf for MS/MS spectra and csv files containing the peak area for each scan. For the molecular networking, the processed peak lists were uploaded to the GNPS2 platform (https://gnps2.org) to construct feature-based molecular networks (FBMN)[19]. MS1 and MS2 fragment ion tolerances were set at 0.005 Da each. A cosine score threshold of 0.7 and a minimum of six MS2 fragment matches were required to construct the molecular network, and a cosine score threshold of 0.6 and a minimum of four MS2 fragment matches were required to perform spectral searches against GNPS2 libraries to account for the variability of peaks from different mass instruments. The jobs can be accessed at the following links: https://gnps2.org/status?task=4e0677e342004091be96abc416a54eed for the positive mode and https://gnps2.org/status?task=5c8289f1ac5b4c3085071d2263349b2f for the negative mode. The resulting molecular networks were visualized using Cytoscape software (version 3.10.1)[20] and available at Figshare[15].

The in silico tools within the SIRIUS software (version 5.8.6)[21] were utilized for structural annotation of each node. Briefly, we performed MS2 annotation for features using a 5 ppm tolerance and Orbitrap profile mode for the SIRIUS module, and 10 candidates were stored for the next analysis. The Zodiac function at 0.95 of edge threshold was used to obtain ranked molecular formula predictions in SIRIUS software. CSI:FingerID

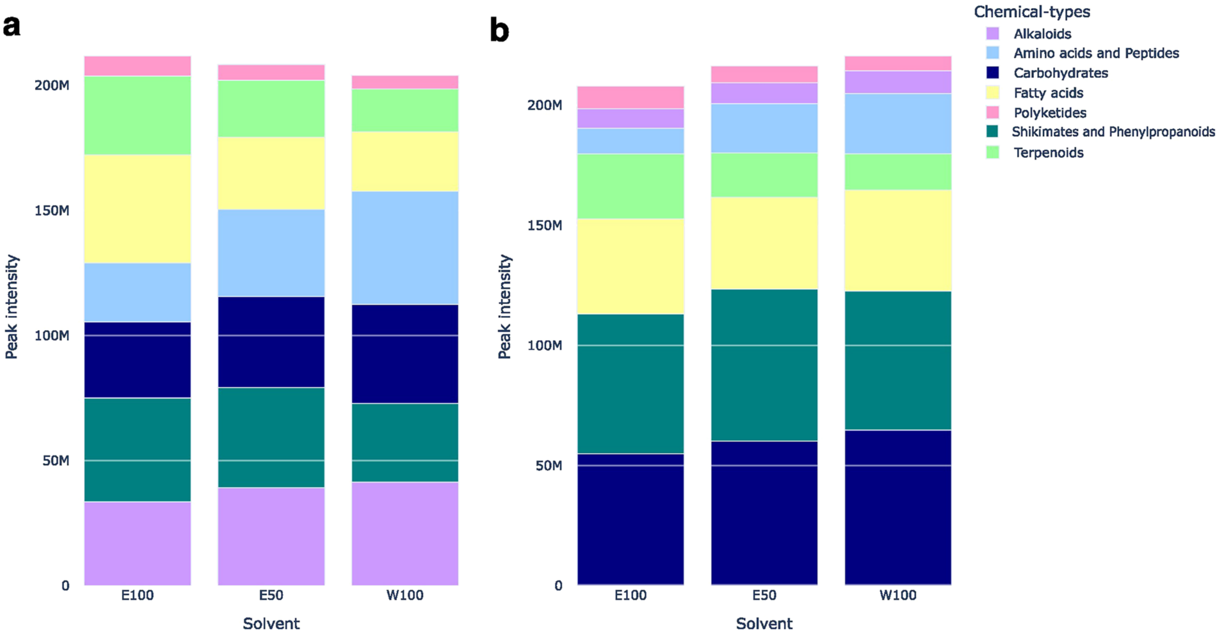

Fig. 2 Stacked bar plot for the chemical classification of pathway level ontology determined by NPClassifier software from the positive (**a**) and negative ion (**b**) modes.

was applied to predict the structural fingerprint by setting $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+2H]^+$ as the adducts for the positive mode and $[M\text{-}H]^-$ for the negative mode. and apply database search of compounds (all databases except in silico databases)[22]. Finally, the chemical class was finally obtained using CANOPUS[23]. The results from NPClassifier function, such as the levels of pathway, superclass, and class, were used as the structural annotation information.

Data Records

The data files are available at the MassiVE data repository portal (https://massive.ucsd.edu/). A total of 1,488 raw LC-MS/MS files, prepared by the solvent with three different polarities and measured by two ionization modes from 248 medicinal plants, are accessible via the MassIVE repository (MSV000097043) (https://doi.org/10.25345/C5GM8215S)[24]. The features for the scans processed using MZmine software and the chemical taxonomic data from NPClassifier modules in SIRIUS software for the positive and negative modes are accessible as an xlsx file at Figshare[15].

Technical Validation

Selection and preparation of representative medicinal plant materials.

Medicinal plants with both market accessibility and pharmaceutical relevance were prioritized. A comprehensive list of natural medicinal materials was obtained from the Korean Pharmacopoeia, managed by the Ministry of Food and Drug Safety (MFDS) of the Republic of Korea[25]. From this list, non-plant-derived materials, such as animal and mineral sources, were excluded, resulting in the selection of 248 plant-derived samples and the metadata are accessible as an xlsx file at Figshare[15]. The samples are primarily cultivated or collected in Korea and China, and are commonly distributed in dried form, with only a few additional processing, such as steaming. This systematic selection ensured consistency and reproducibility for downstream analyses.

Chromatographic and mass spectrometric analysis.

To capture the broad range of metabolites, chromatographic separations were performed under a gradient of solvents with varying polarities, from highly polar to nonpolar conditions. Mass spectrometric data were acquired in both positive and negative ion modes to ensure comprehensive metabolite detection. Data processing, including noise filtering and peak alignment, was carried out using MZmine software, which facilitated the generation of high-quality peak data for further analysis. Sulfadimethoxine was used as internal standards at predefined concentrations in the extraction solvents for retention time checks and normalization for metabolomic analyses. The processed data were normalized using sum normalization and converted into.mgf files for deposition and further analysis.

Considering their applicability in the industry, food-grade ethanol was selected as the non-polar solvent and three extraction solvents were chosen: 100% water, 50% ethanol, and 100% ethanol. These solvents cover a broad polarity range, enabling the extraction of diverse metabolites. A total of 744 samples were prepared using these solvents to ensure robust and comprehensive profiling of plant-derived compounds. 63,944 and 42,481 MS/MS scans were obtained in the positive and the negative modes, respectively, from 744 samples. The size of the datasets of total 1,488 files is larger than the previously reported datasets due to differences in equipment performance and variations in solvent conditions[26]. The accumulation curves, as visualized by unique scans corresponding

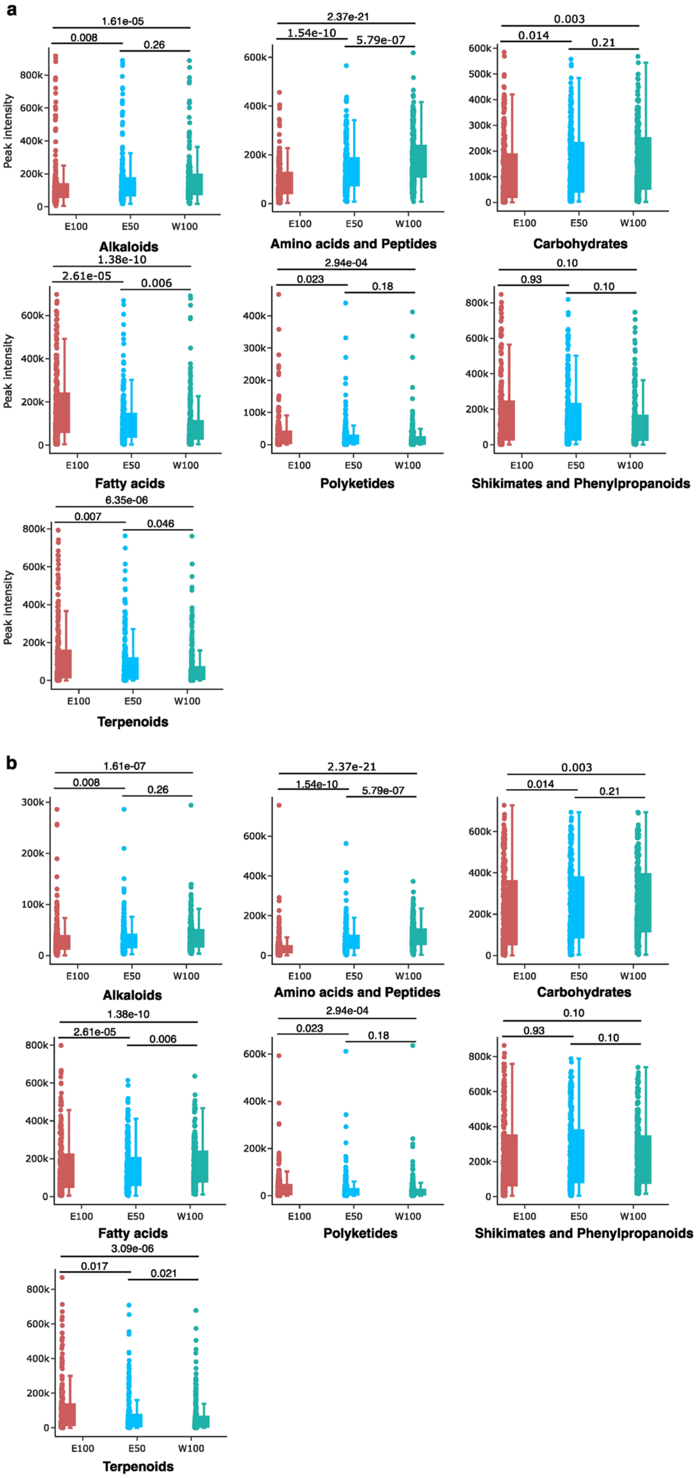

Fig. 3 Box-and-whisker plots of superclass level ontology determined by NPClassifier software from the positive (**a**) and negative ion (**b**) modes. A non-parametric Kruskal-Wallis test followed by Wilcoxon was performed, and all p values were corrected for multiple comparisons using the Benjamini-Hochberg correction.

to compounds extracted by the three solvent systems, revealed that solvent polarity significantly influenced the chemical profiles among the samples (Fig. 1a,b, and Supplementary Figures S1). We observed that the change in solvent polarity affected the distribution of nodes with different peak intensities. For example, Venn diagrams of a total of 9,746 scans with top 10% intensities among the total scans revealed that 33.4% (3,254 scans) existed simultaneously in three solvents, while 16.0% (1,555 scans) in 100% water, 6.0% (580 scans) in 50% ethanol, and 14.6% (1,426 scans) in 100% ethanol were exclusively detected in the one solvent system in the positive mode (Fig. 1c). In the case of the negative mode, 31.1% (2,058 scans) were in three solvent, and 15.7%

(1,037 scans), 6.0% (395 scans), and 17% (1,126 scans) were only in 100% water, 50% ethanol, and 100% ethanol, respectively (Fig. 1d). A similar trend was observed in both ion modes. Notably, 50% ethanol solvent extracted the least number of unique scans due to its intermediate properties between the other two solvents. Nevertheless, these results indicate that three solvent systems have a significant impact on molecular diversity.

Molecular diversity analysis using in silico molecular annotation. The availability of LC/MS libraries, although still limited, is steadily increasing. For structural annotation, in silico tools within the SIRIUS platform were utilized[27]. NPClassifier and ClassyFire modules in SIRIUS are widely used approaches for chemical classification studies[28,29]. NPClassifier, specifically designed for natural product classification, provided more consistent and reliable classification results[28]. These classifications revealed the compound distribution at the pathway level, influenced by solvent polarity. The two ionization modes exhibited divergent chemical distributions (Fig. 2). In the positive ion mode, nitrogen-containing compounds, such as "Alkaloids" and "Amino acids and Peptides", exhibited the stronger intensities. Whereas "Carbohydrates" and "Shikimates and Phenylpropanoids"-type compounds with a few hydroxyl groups in their functional groups dominated in the negative ion mode. Next, we compared the distribution of each chemical type to investigate differences in the extraction efficiency of compounds within each solvent system (Fig. 3). The results demonstrated a statistically significant difference observed in the positive ion mode in comparison to the negative mode. When comparing the extraction efficiency of each chemical type, "Alkaloids" and "Amino acids and Peptides", which contain nitrogen-containing compounds, exhibited higher peak intensities as the solvent polarity increased in both ionic modes, while "Fatty acids" and "Terpenoids" demonstrated better extraction efficiency at lower polarity.

Validation in silico annotation results using the molecular networking and public libraries. To validate the in silico annotation results, the scans from both ion modes were applied to the molecular networking for the investigation of the spectral similarities between the scans and to publicly available libraries (Supplementary Figure S2). In the molecular networks, the clusters of scans with similar MS/MS patterns were grouped together in the same color. Additionally, the in silico annotation results can be validated through the comparison of spectral data within public mass spectral libraries. Among a total of 7,925 scans in positive ion mode and 4,271 scans in negative ion mode annotated with SMILES information using the library repository in the publicly opened GNPS spectra library, the chemical classes of 7,461 and 4,009 scans were successfully annotated by NPClassifier module, respectively, and are accessible as an xlsx file at Figshare[15].

Among those scans, the accuracy of the chemical classification results derived from the in silico annotation method using only MS/MS patterns were 75.6%, 60.8% and 40.7% for "pathway level", "superclass level" and "class level" in the positive mode, and 77.0%, 62.0% and 47.5%in the negative mode, respectively. This finding indicated that the in silico annotation results in the dataset can be reliably used in other studies. Consequently, the raw MS spectral data of the dataset produced in this study, along with the annotated chemical taxonomic data of each scan, will serve as a reliable resource for the study of specialized metabolites in plants.

Code availability

The software version and parameters described in the study were found in the Methods section. The custom codes for dataset visualization is available at https://github.com/heejyang/MediHerb.

Received: 24 February 2025; Accepted: 22 May 2025;
Published online: 28 May 2025

References

1. Lei, X. *et al*. Status and thoughts of Chinese patent medicines seeking approval in the US market. *Chin. J. Integr. Med.* **20**, 403–408 (2014).
2. Wu, C. *et al*. Scientific and Regulatory Approach to Botanical Drug Development: A U.S. FDA Perspective. *J. Nat. Prod.* **83**, 552–562 (2020).
3. Heo, Y.-A. Birch Bark Extract: A Review in Epidermolysis Bullosa. *Drugs* **83**, 1309–1314 (2023).
4. Sitnikov, D. G., Monnin, C. S. & Vuckovic, D. Systematic Assessment of Seven Solvent and Solid-Phase Extraction Methods for Metabolomics Analysis of Human Plasma by LC-MS. *Sci Rep* **6**, 38885 (2016).
5. ICH Q3C (R9) Guideline on impurities: guideline for residual solvents_Step 5. *European Medicines Agency* (2024).
6. Zhou, Z. *et al*. Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat Commun* **13**, 6656 (2022).
7. Danzi, F. *et al*. To metabolomics and beyond: a technological portfolio to investigate cancer metabolism. *Sig Transduct Target Ther* **8**, 137 (2023).
8. Da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. USA.* **112**, 12549–12550 (2015).
9. 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).
10. Wang, J. *et al*. Extraction and purification of total flavonoids from Zanthoxylum planispinum Var. Dintanensis leaves and effect of altitude on total flavonoids content. *Sci Rep* **15**, 7080 (2025).
11. Nouioura, G. *et al*. Comprehensive analysis of different solvent extracts of Ferula communis L. fruit reveals phenolic compounds and their biological properties via *in vitro* and in silico assays. *Sci Rep* **14**, 8325 (2024).
12. Tourabi, M. *et al*. Efficacy of various extracting solvents on phytochemical composition, and biological properties of Mentha longifolia L. leaf extracts. *Sci Rep* **13**, 18028 (2023).
13. Ghaffar, N. & Perveen, A. Solvent polarity effects on extraction yield, phenolic content, and antioxidant properties of Malvaceae family seeds: a comparative study. *New Zealand Journal of Botany* 1–11 https://doi.org/10.1080/0028825X.2024.2392705 (2024).
14. Wakeel, A., Jan, S. A., Ullah, I., Shinwari, Z. K. & Xu, M. Solvent polarity mediates phytochemical yield and antioxidant capacity of *Isatis tinctoria*. *PeerJ* **7**, e7857 (2019).
15. Yang, H. *et al*. Mass Spectral Data of Primary and Secondary Metabolites Changes in Medicinal Plants by Solvent Polarity. 325090106 Bytes https://doi.org/10.6084/M9.FIGSHARE.29062577.V1 (2025).

16. Chambers, M. C. *et al*. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918–920 (2012).
17. Schmid, R. *et al*. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat Biotechnol* **41**, 447–449 (2023).
18. Schmid, R. *et al*. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. *Nat Commun* **12**, 3832 (2021).
19. Nothias, L.-F. *et al*. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* **17**, 905–908 (2020).
20. Shannon, P. *et al*. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).
21. Ludwig, M. *et al*. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nat Mach Intell* **2**, 629–641 (2020).
22. Hoffmann, M. A. *et al*. High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol* **40**, 411–421 (2022).
23. Dührkop, K. *et al*. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* **39**, 462–471 (2021).
24. Cho, C. & Yang, H. Mass Spectral Data of Primary and Secondary Metabolites Changes in Medicinal Plants by Solvent Polarity. https://doi.org/10.25345/C5GM8215S (2025).
25. Administration MFDS (Korea Ministry of Food and Drug Safety). *The Korean Pharmacopoeia 12th Ed*. (MFDS 2019).
26. Kang, K. B. *et al*. Mass spectrometry data on specialized metabolome of medicinal plants used in East Asian traditional medicine. *Sci Data* **9**, 528 (2022).
27. Dührkop, K. *et al*. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* **16**, 299–302 (2019).
28. Kim, H. W. *et al*. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **84**, 2795–2807 (2021).
29. Djoumbou Feunang, Y. *et al*. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8** (2016).

Acknowledgements

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) grant (Project No. RS-2021-NR061647) funded by Ministry of Education, Science and Technology, Republic of Korea and Cooperative Research Program for Agriculture Science and Technology Development (Project No. RS-2022-RD010287) funded by Rural Development Administration, Republic of Korea.

Author contributions

J.P. and J.L. contributed to M.S. data acquisition, C.C., D.L. and W.S. contributed to plant samples preparation, J.P. and H.Y. contributed to paper writing. H.Y. developed the concept.

Competing interests

H.Y. is a co-founder and holds equity in Bionsight, Inc. with prior approval by Kangwon National University.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05262-8.

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