



OPEN

DATA DESCRIPTOR

# Mass spectrometry data on specialized metabolome of medicinal plants used in East Asian traditional medicine

Kyo Bin Kang<sup>1</sup>✉, Eunah Jeong<sup>1</sup>, Seungju Son<sup>1</sup>, Eunjin Lee<sup>1</sup>, Seungjin Lee<sup>2</sup>, SeongYeon Choi<sup>3</sup>, Hyun Woo Kim<sup>4</sup>, Heejung Yang<sup>3</sup> & Sang Hee Shim<sup>2</sup>✉

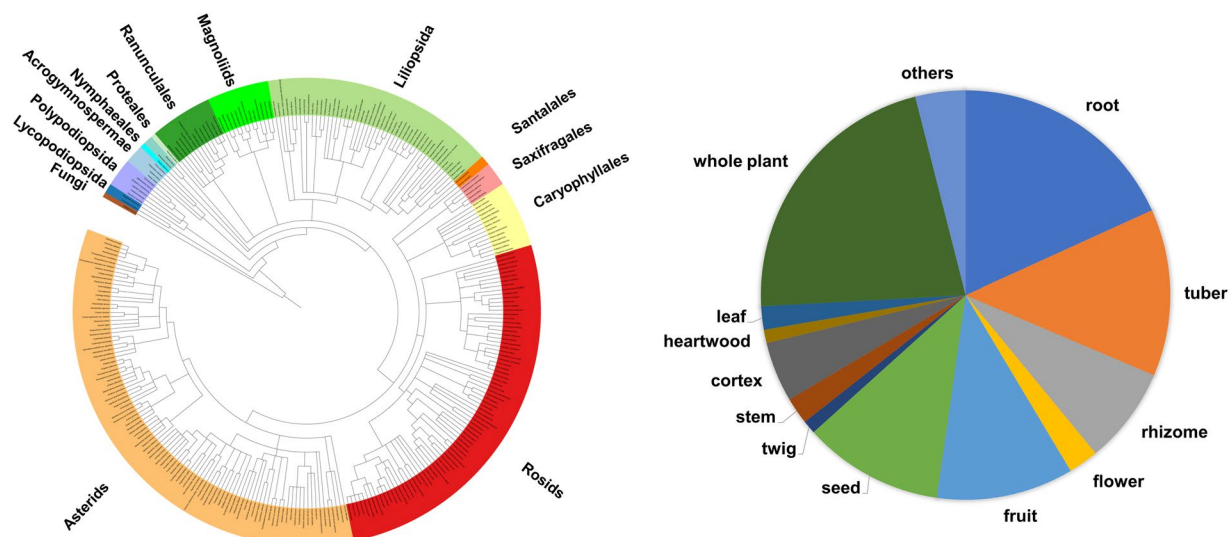
Traditional East Asian medicine not only serves as a potential source of drug discovery, but also plays an important role in the healthcare systems of Korea, China, and Japan. Tandem mass spectrometry (MS/MS)-based untargeted metabolomics is a key methodology for high-throughput analysis of the complex chemical compositions of medicinal plants used in traditional East Asian medicine. This Data Descriptor documents the deposition to a public repository of a re-analyzable raw LC-MS/MS dataset of 337 medicinal plants listed in the Korean Pharmacopeia, in addition to a reference spectral library of 223 phytochemicals isolated from medicinal plants. Enhanced by recently developed repository-level data analysis pipelines, this information can serve as a reference dataset for MS/MS-based untargeted metabolomic analysis of plant specialized metabolites.

## Background & Summary

Most cultures worldwide use plants to treat diseases. The integration of experimental knowledge on medicinal plant usage with theories or beliefs about health and illness is termed traditional medicine. Traditional East Asian traditional medicine is known to have originated approximately 3,000 years ago in China. It was introduced to Korea and Japan from China in the 6th century, with Buddhism and Chinese culture<sup>1</sup>. Since then, it has been widely used following a long history. The detailed practices are not exactly same in three countries due to several reasons, for example, the usage of different species in the same genus due to the different climate; however, they have strongly influenced each other. Traditional East Asian medicine still plays an essential role in public health care in East Asian countries; currently, standardized herbal formulae are manufactured by pharmaceutical companies and used as parts of the modern medical systems in Korea, China, and Japan. Medicinal plants used in traditional medicine are one of the most important sources of drug discovery, where artemisinin, an antimalarial agent from *Artemisia annua*, is the most representative case.

High-throughput analysis of samples with complex chemical compositions plays a key role in the investigation and modernization of traditional East Asian medicine. Tandem mass spectrometry (MS/MS), especially in combination with liquid chromatography (LC), is the analytical method most commonly used to analyze medicinal plants<sup>2,3</sup>. MS/MS-based untargeted metabolomics has been used to assess the quality of medicinal plants and related dietary and pharmaceutical products, and has also been utilized in structure-based bioactive compound discovery<sup>4-7</sup>. Despite the increased use of this technique, only a few reliable and controlled datasets of MS/MS data for medicinal plants have been deposited in public repositories. With the expansion of untargeted metabolomics to multiple fields, the importance of publicly available data is increasing. The successive launches of MASST and ReDU symbolize the increasing need for public datasets in MS/MS-based untargeted metabolomics research<sup>8,9</sup>. MASST enables the search for a single spectrum by comparison with all publicly available raw

<sup>1</sup>Research Institute of Pharmaceutical Sciences, College of Pharmacy, Sookmyung Women's University, Seoul, 04310, Korea. <sup>2</sup>Natural Products Research Institute, College of Pharmacy, Seoul National University, Seoul, 08826, Korea. <sup>3</sup>College of Pharmacy, Kangwon National University, Chuncheon, 24341, Korea. <sup>4</sup>College of Pharmacy and Integrated Research Institute for Drug Development, Dongguk University, Goyang, 10326, Korea. ✉e-mail: [kbkang@sookmyung.ac.kr](mailto:kbkang@sookmyung.ac.kr); [sanghee\\_shim@snu.ac.kr](mailto:sanghee_shim@snu.ac.kr)



**Fig. 1** Taxonomic coverage (left) and used plant parts (right; ‘others’ includes calyx, carpel, gall, husk, peel, pericarp, pollen, and resin) of the 337 medicinal plant samples. The phylogenetic tree was constructed using the PhyloT phylogenetic tree generator, based on NCBI taxonomy<sup>39</sup>. Taxa of different levels are marked in the phylogenetic tree to enhance readability.

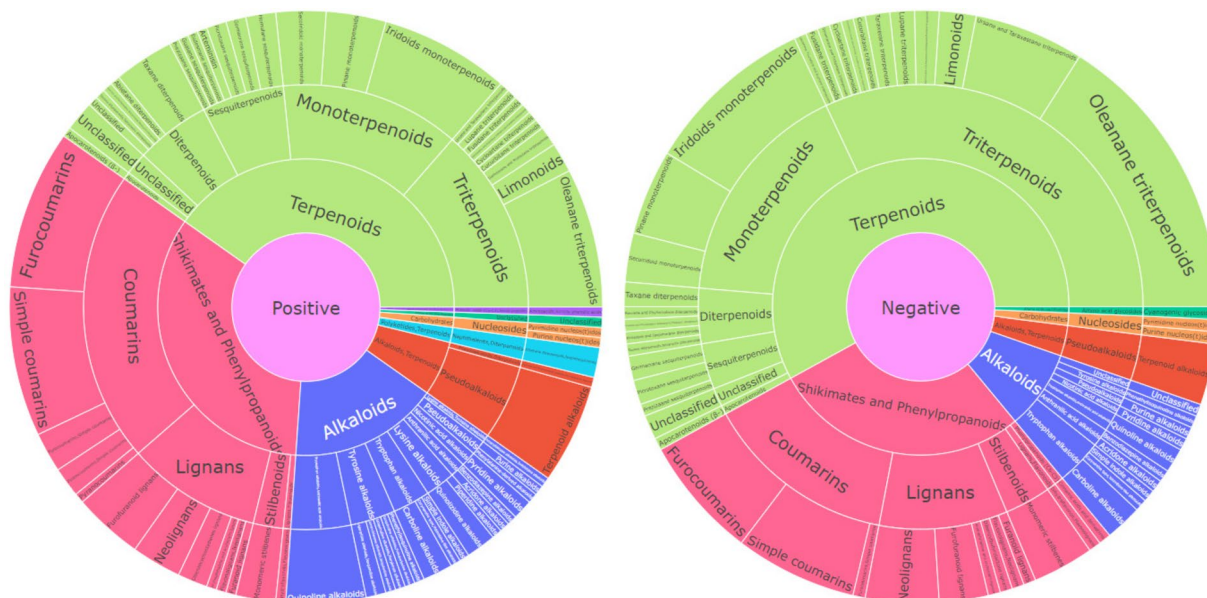
data, whereas ReDU enables the reuse of deposited datasets for repository-level analyses or co-analysis with the user’s own experimental data.

This Data Descriptor documents publicly available and re-analyzable raw LC-MS/MS dataset of 337 medicinal plants on the MassIVE raw data repository, which is linked with the Global Natural Product Social Molecular Networking platform (GNPS, <https://gnps.ucsd.edu>)<sup>10</sup>. This dataset is referred to as the KP337 dataset in this Data Descriptor, as most of the medicinal plants in the dataset are listed in the Korean Pharmacopeia (KP). The data do not cover the entire set of medicinal plants enlisted in KP, but cover most of the commonly used plants. The taxonomic coverage of the plants and plant parts used in Traditional East Asian medicine is summarized in Fig. 1. The KP337 dataset consists of raw LC-MS/MS data acquired in both positive and negative ion mode, and metadata formatted for compatibility with ReDU<sup>9</sup>. Thus, this dataset enables data re-usage, such as comparative analysis or propagation of spectral annotation based on spectral similarity<sup>11</sup>. Recently, a part of this dataset, specifically relating to various flavonoid C-glycosides, was utilized to establish the GNPS nearest neighbor suspect spectral library<sup>12</sup>. This case demonstrates the applicability of public datasets for propagation of spectral annotations. The KP337 dataset can also be applied to a MASST search of MS/MS spectra, which suggests the possible occurrence of queried molecules in medicinal plants. In natural product discovery projects, known compounds are often ignored and not reported. However, novel occurrences of known chemicals can provide insights into the medicinal or biological properties of medicinal plants, where the present dataset can contribute to such findings via MASST. Additionally, the occurrence data of known or unknown chemicals can enhance reference data-driven analysis, which was suggested as an alternative workflow for MS/MS-based untargeted metabolomics<sup>13</sup>.

We also report our efforts to establish a MS/MS spectral library of bioactive compounds obtained from medicinal plants that are used in traditional East Asian medicine. Although many phytochemicals have been previously found in medicinal plants, most of them vary in the historical collections of natural product chemistry laboratories and their MS/MS spectra have not been reported. Benchmarking recent efforts leading to the monoterpene indole alkaloid database (MIADB), and isoquinoline alkaloids and other annonaceous metabolites database (IQAMDB), which are spectral libraries built with compounds from historical collections of various natural product chemistry laboratories<sup>14,15</sup>, we established an MS/MS spectral library using 223 pure phytochemicals obtained from the legacy compound library of the Natural Products Research Institute, Seoul National University (SSK Legacy Library, named after Sam Sik Kang, who compiled the library over the course of 30 years). MS/MS spectra were acquired for all ionized molecules in positive (ESI+) and negative (ESI-) ion modes, which yielded 184 positive and 152 negative ion mode spectra. Compounds with low ionization efficiencies in each ionization mode were excluded. This spectral library will accelerate the annotation of phytochemicals in future metabolomic studies of medicinal plants. The chemical ontology of the phytochemicals included in the spectral library was estimated using the NPClassifier<sup>16</sup> and is summarized in Fig. 2.

## Methods

**Sample preparation.** The methanolic extracts of 337 medicinal plants were acquired from the Korea Plant Extract Bank of the Korea Research Institute of Bioscience and Biotechnology (Cheongju, Korea), where authentic samples were extracted with MeOH via sonication for 3 h. The extracts were dissolved in MeOH at 2 mg/mL for LC-MS/MS analysis. The collection of 223 compounds previously isolated from various medicinal plants by Prof. Sam Sik Kang (Seoul National University) has been maintained in Sang Hee Shim’s laboratory (Seoul



**Fig. 2** Chemical ontology of phytochemicals giving rise to positive (left) and negative (right) ion mode spectra.

National University), coupled with NMR spectra for structural identification. The compounds were dissolved in 50% aqueous MeOH at 100  $\mu\text{g}/\text{mL}$  concentration for LC-MS/MS analysis.

**LC-MS/MS data acquisition.** To compile the plant extract dataset, LC-MS/MS data were acquired using a Waters Acquity UPLC system (Waters Co., Milford, MA, USA) coupled to a Xevo G2 Q-TOF mass spectrometer (Waters MS Technologies, Manchester, UK) equipped with an electrospray ionization interface (ESI). To compose the spectral library, LC-MS/MS data were acquired using a Waters Acquity I-Class UPLC system linked to a Waters VION IMS Q-TOF MS equipped with an ESI interface. Chromatographic separation was performed using a Waters BEH  $\text{C}_{18}$  column ( $50 \times 2.1 \text{ mm}$ ,  $1.7 \mu\text{m}$ ), which was eluted with a mixture of water (solvent A) and MeCN (solvent B), both acidified with 0.1% formic acid, at a flow rate of 0.3 mL/min, with a linear gradient of 5–95% B (0–14 min) followed by 3 min washing with 100% B and 3 min reconditioning with 5% B successively. The samples (2.0  $\mu\text{L}$  injection volume) were analyzed in data-dependent acquisition (DDA) mode consisting of full MS survey scans in the  $m/z$  100–2000 Da range (scan time: 150 ms) followed by MS/MS scans for the three most intense ions ( $m/z$  100–2000 Da; scan time: 100 ms). The collision energy gradient was set as 20–100 eV. Protonated and deprotonated ions of leucine enkephalin ( $m/z$  556.2771  $[\text{M} + \text{H}]^+$  in the positive ion mode and  $m/z$  554.2615  $[\text{M} - \text{H}]^-$  in the negative ion mode, respectively) were measured in every 0.1 min as the lock mass to ensure mass accuracy and reproducibility.

**Public dataset deposition.** All the raw LC-MS/MS data files were converted from the Waters.Raw format to the open-sourced.mzML format using the MSConvert tool of ProteoWizard<sup>17</sup>. Sample metadata were prepared to ensure compatibility with ReDU. Additional sample information which are not included in the ReDU format, such as the plant part analyzed and taxonomic ontology, are included in additional metadata files.

**Molecular networking analysis.** Molecular networks were created using the classical molecular networking workflow on the GNPS web platform<sup>10</sup>. The networks were then created, where edges were filtered to have a cosine score above 0.65 and more than 5 matched peaks, with the precursor and fragment ion mass tolerance of 0.02 Da. The library spectra were searched to find annotation with the same score and matched peaks.

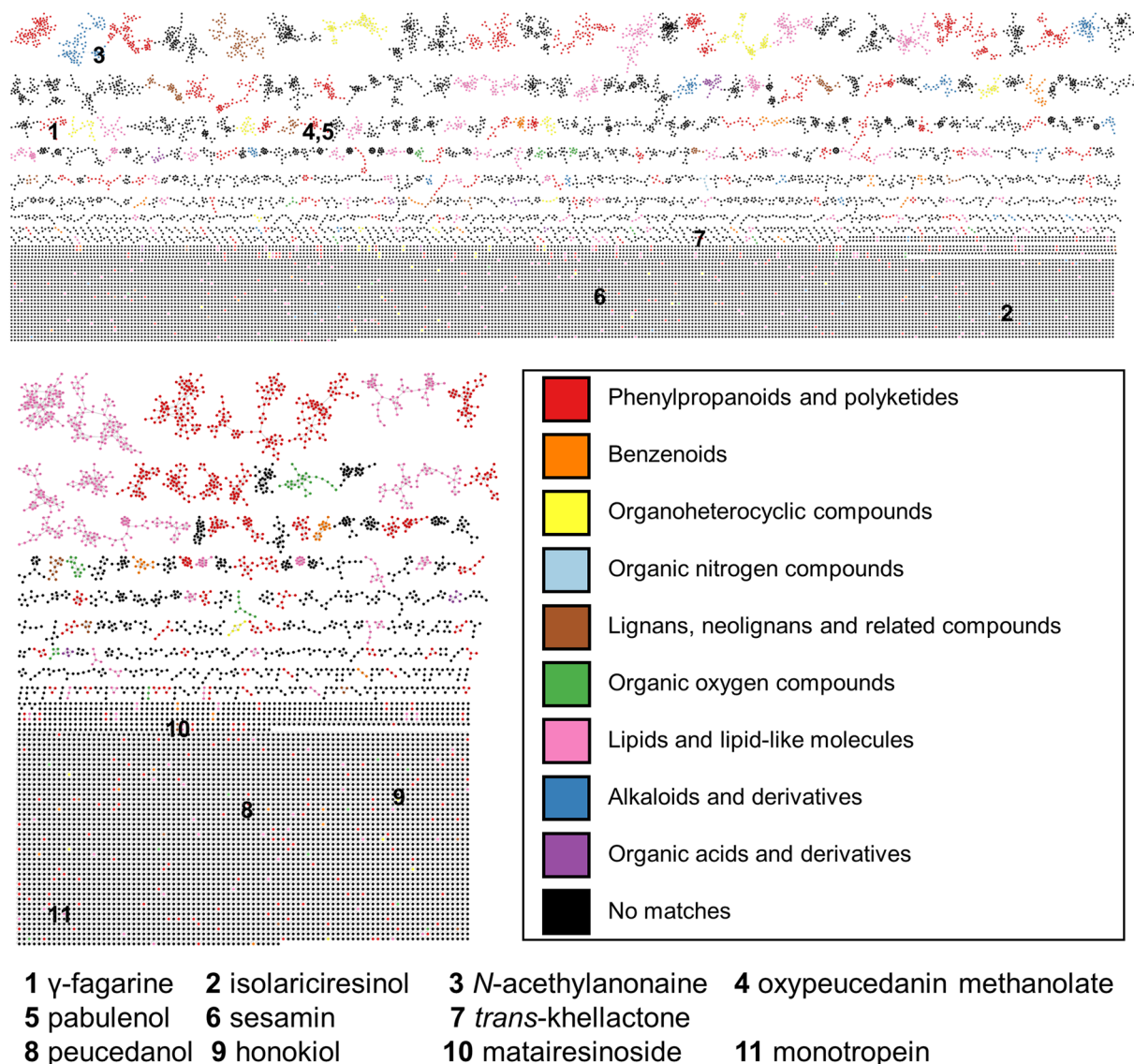
The molecular network with the positive ion mode data can be accessed via: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=94bd6547c84341ddaaff2e4599247871>

The positive ion mode MolNetEnhancer network is accessible via: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f896fc4740694c3fa8308dabfd2ff3c3>.

The molecular network with the negative ion mode data can be accessed via: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d4004a9da4c84ed2b509a83813ebbea1>.

The negative ion mode MolNetEnhancer network is accessible via: <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=301d52ce05244646824e6a96f7990bc>.

**Spectral library constitution.** The acquired LC-MS/MS files were opened with Waters UNIFI software. The list of MS features was automatically degenerated by UNIFI. MS features with hypothetical  $m/z$  values corresponding to commonly observed adducts ( $[\text{M} + \text{H}]^+$ ,  $[\text{M} + \text{Na}]^+$ ,  $[2\text{M} + \text{H}]^{2+}$ ,  $[\text{M} - \text{H}]^-$ ,  $[\text{M} + \text{HCOOH} - \text{H}]^-$ , and  $[2\text{M} - \text{H}]^-$ ) were found, after which the MS/MS spectra were manually examined. The MS/MS scans related



**Fig. 3** The molecular networks of the 337 medicinal plant samples, acquired in the positive (up) and negative (down) ion mode. The colours of spectral nodes denote class annotations given by the MolNetEnhancer workflow. Spectral nodes matched to the SSK legacy spectral library are represented by numbering.

to the standard compounds were exported as.mgf files. The metadata used to establish the spectral library, including SMILES and InChI identifiers of the structures, is provided as Supplementary Table 1.

### Data Records

Raw LC-MS/MS data of the 337 medicinal plants are accessible via MassIVE (<https://massive.ucsd.edu>) with the accession number MSV000086161<sup>18</sup>. All the data can be re- or co-analyzed via ReDU<sup>9</sup>. MS/MS spectra of the 223 phytochemicals (184 in positive ion mode and 152 in negative ion mode) from Dr. Sam Sik Kang's legacy chemical library have unique accession numbers from CCMSLIB00010007697 to CCMSLIB00010008032 in the spectral library of GNPS<sup>10</sup>, which is accessible via: <https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=GNPS-SAM-SIK-KANG-LEGACY-LIBRARY>).

### Technical Validation

**Molecular networking-based overview of the KP337 dataset.** Classical molecular networking analyses were performed with the datasets acquired in positive and negative ion mode<sup>18</sup>, respectively, to provide an overview of the specialized metabolome of the analyzed medicinal plants. The positive ion mode data yielded a molecular network consisting of 16,533 spectral nodes, while the negative ion mode data gave a network of 6,570 spectral nodes. The MolNetEnhancer workflow assigned class annotations to molecular families based on the spectral matching-based annotation<sup>19</sup>, the resulting molecular networks are shown in Fig. 3, and the class annotations are summarized in Table 1. In both networks, phenylpropanoids and polyketides account for the largest portion of the annotated molecular networks. This may be due to the high polyphenol content of the medicinal

ClassyFire Superclass	Positive ion mode	Negative ion mode
Phenylpropanoids and polyketides	1274 (7.71%)	799 (12.16%)
Benzenoids	106 (0.64%)	44 (0.67%)
Organoheterocyclic compounds	266 (1.61%)	12 (0.18%)
Organic nitrogen compounds	6 (0.04%)	0 (0.00%)
Lignans, neolignans, and related compounds	221 (1.34%)	29 (0.44%)
Organic oxygen compounds	59 (0.36%)	67 (1.02%)
Lipids and lipid-like molecules	765 (4.63%)	602 (9.16%)
Alkaloids and derivatives	284 (1.72%)	0 (0.00%)
Organic acids and derivatives	64 (0.39%)	12 (0.18%)
No matches	13488 (81.58%)	5005 (76.18%)
Sum	16533	6570

**Table 1.** Numbers of spectral nodes for each superclass-level annotation provided by the MolNetEnhancer workflow.

Compound	Ion mode	Cosine	Occurrence	Comment
$\gamma$ -fagarine	positive	0.98	<i>Phellodendron amurense</i> (cortex)	Described from <i>P. amurense</i> <sup>21</sup>
			<i>Dictamnus albus</i> (cortex)	Described from <i>D. albus</i> <sup>22</sup>
isolariciresinol	positive	0.85	<i>Patrinia scabiosifolia</i> (root)	Described from other <i>Patrinia</i> species <sup>27</sup>
			<i>Rubia argyi</i> (root)	Described from <i>R. argyi</i> <sup>26</sup>
<i>N</i> -acetylanonaine	positive	0.71	<i>Nelumbo nucifera</i> (whole/root)	Analogues are described from <i>N. nucifera</i> <sup>31</sup>
oxypeucedanin methanolate	positive	0.81	<i>Angelica dahurica</i> (root)	Described from <i>A. dahurica</i> <sup>24,25</sup>
pabulenol	positive	0.84	<i>Angelica dahurica</i> (root)	Described from <i>A. dahurica</i> <sup>24,25</sup>
sesamin	positive	0.79	<i>Asarum heterotropoides</i> (root)	Described from <i>A. heterotropoides</i> <sup>23</sup>
<i>trans</i> -khellactone	positive	0.76	<i>Glehnia littoralis</i> (root)	Analogues are described from <i>G. littoralis</i> <sup>29</sup>
peucedanol	negative	0.85	<i>Peucedanum japonicum</i> (root)	Described from <i>P. japonicum</i> <sup>32</sup>
			<i>Peucedanum praeruptorum</i> (root)	Described from <i>P. praeruptorum</i> <sup>33</sup>
			<i>Glehnia littoralis</i> (root)	Analogues are described from <i>G. littoralis</i> <sup>30</sup>
honokiol	negative	0.93	<i>Magnolia obovata</i> (cortex)	Described from <i>M. obovata</i> <sup>34</sup>
matairesinoside	negative	0.66	<i>Trachelospermum asiaticum</i> (whole)	Described from <i>T. asiaticum</i> <sup>35</sup>
			<i>Arctium lappa</i> (seed)	Described from <i>A. lappa</i> <sup>36</sup>
			<i>Patrinia scabiosifolia</i> (root)	—
			<i>Forsythia suspensa</i> (fruit)	Described from <i>F. suspensa</i> <sup>37</sup>
Monotropein	negative	0.67	<i>Pyrola japonica</i> (whole)	Described from <i>P. japonica</i> <sup>38</sup>
			<i>Gynochthodes officinalis</i> (root)	—

**Table 2.** Spectral matches between KP337 dataset and KSS legacy library.

plants. Organic nitrogen compounds and alkaloids and derivatives were only observed in the positive ion mode data-based network because of the basicity of nitrogen-containing compounds. For the positive and negative ion mode networks, 81.6% (13,488 of 16,533) and 76.2% (5,005 of 6,570) of the spectral nodes were respectively unannotated. This can seem to be too low, but it needs to be denoted that the class of each molecular family was annotated only based on the spectral matching. The number of publicly available reference spectra are still much lower than the number of known phytochemicals; thus, application of *in silico* spectral annotation methods would increase the annotation rate, as we demonstrated in a previous study on specialized metabolome of the family Rhamnaceae<sup>20</sup>.

**Spectroscopic validation of the phytochemicals.** The structures of the purified phytochemicals were confirmed by manual inspection of the nuclear magnetic resonance (NMR) and MS spectra.

**Dereplication of the KP medicinal plants data against the spectral library.** For further validation of the spectral library, we re-established the molecular networks of the KP337 dataset<sup>18</sup> using the SSK legacy spectral library, together with all the spectral libraries available in GNPS. Consequently, 11 compounds (7 in positive and 4 in negative ion mode) were matched as the candidates with the highest scores (Table 2). Most of the sample occurrences in the matched spectra correlated with the previous reports of the matched molecules from taxonomically same or close species, which supported the reliability of the dereplication result.  $\gamma$ -Fagarine was isolated from *Phellodendron amurense*<sup>21</sup> and *Dictamnus albus*<sup>22</sup>, while sesamin was found in *Asarum heterotropoides*<sup>23</sup>, and oxypeucedanin methanolate and pabulenol were isolated from *Angelica dahurica*<sup>24,25</sup>. Isolariciresinol, which was reported from *Rubia argyi*<sup>26</sup>, has not previously been reported in *Patrinia scabiosifolia*; however, it was reported from *P. scabra*, another species of the genus<sup>27</sup>. Due to the possible conservation of biosynthesis in close taxa,

species in the same genus or family often contain the same or similar specialized metabolites<sup>28</sup>. Thus, the occurrence of isolacricresnol in *P. scabiosifolia* can be supported by the occurrence of the same compound in *P. scabra*. Spectral matching suggested the occurrence of *trans*-khellactone and peucedanol in *Glehnia littoralis*, but neither of these two compounds have been reported in the plant. However, six *O*-acyl derivatives of *cis*-khellactone were reported from *G. littoralis*<sup>29</sup>. Along with the spectral matching results, this suggests the occurrence of *cis*-khellactone in *G. littoralis*, as the *cis/trans* isomers cannot be distinguished by MS/MS analysis. Similarly, oxypeucedanol has not been reported in *G. littoralis*, but multiple oxypeucedanol glycosides were reported from *G. littoralis*, which supports the occurrence of the aglycone in this plant<sup>30</sup>. These cases simultaneously highlight the applicability and value of the dataset and spectral library introduced in this Data Descriptor; the coverage of spectral matching was expanded, and the occurrence of previously known compounds can be easily estimated by searching the spectra against the dataset.

### Code availability

The data processing workflow for establishing the spectral library is available in GNPS.

Received: 30 June 2022; Accepted: 23 August 2022;

Published online: 27 August 2022

### References

- Park, H.-L. *et al.* Traditional medicine in China, Korea, and Japan: a brief introduction and comparison. *Evid. Based. Complement. Alternat. Med.* **2012**, 429103 (2012).
- Kind, T. & Fiehn, O. Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry. *Phytochem. Lett.* **21**, 313–319 (2017).
- Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H. & Bertrand, S. Accelerating metabolite identification in natural product research: toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. *Anal. Chem.* **91**, 704–742 (2019).
- Ernst, M., Silva, D. B., Silva, R. R., Vêncio, R. Z. N. & Lopes, N. P. Mass spectrometry in plant metabolomics strategies: from analytical platforms to data acquisition and processing. *Nat. Prod. Rep.* **31**, 784–806 (2014).
- Wolfender, J.-L., Marti, G., Thomas, A. & Bertrand, S. Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* **1382**, 136–164 (2015).
- Allard, P.-M. *et al.* Pharmacognosy in the digital era: shifting to contextualized metabolomics. *Curr. Opin. Biotechnol.* **54**, 57–64 (2018).
- Wolfender, J.-L., Litaudon, M., Touboul, D. & Queiroz, E. F. Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery. *Nat. Prod. Rep.* **36**, 855–868 (2019).
- Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
- Jarmusch, A. K. *et al.* ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
- 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).
- Jarmusch, S. A., van der Hooft, J. J. J., Dorrestein, P. C. & Jarmusch, A. K. Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Nat. Prod. Rep.* **38**, 2066–2082 (2021).
- Bittremieux, W. *et al.* Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. Preprint at <https://www.biorxiv.org/content/10.1101/2022.05.15.490691v1> (2022).
- Gauglitz, J. N. *et al.* Enhancing untargeted metabolomics using metadata-based source annotation. *Nat. Biotechnol.* <https://doi.org/10.1038/s41587-022-01368-1> (2022).
- Fox Ramos, A. E. *et al.* Collected mass spectrometry data on monoterpene indole alkaloids from natural product chemistry research. *Sci. Data* **6**, 15 (2019).
- Agnès, S. A. *et al.* Implementation of a MS/MS database for isoquinoline alkaloids and other annonaceous metabolites. *Sci. Data* **9**, 270 (2022).
- Kim, H. W. *et al.* NPClassifier: A deep neural network-based structural classification tool for natural products. *J. Nat. Prod.* **84**, 2795–2807 (2021).
- Adusumilli, R. & Mallick, P. Data conversion with ProteoWizard msConvert. *Methods Mol. Biol.* **1550**, 339–368 (2017).
- Yang, H., Kang, K. B. & Sung, S. H. GNPS - LC-MS/MS data from 337 Medicinal Plants listed in Korean Pharmacopeia. *MassIVE* <https://doi.org/10.25345/C5SB50> (2022).
- Ernst, M. *et al.* MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **9**, 144 (2019).
- Kang, K. B. *et al.* Comprehensive mass spectrometry-guided phenotyping of plant specialized metabolites reveals metabolic diversity in the cosmopolitan plant family Rhamnaceae. *Plant J.* **98**, 1134–1144 (2019).
- Min, Y. D. *et al.* Isolation of limonoids and alkaloids from *Phellodendron amurense* and their multidrug resistance (MDR) reversal activity. *Arch. Pharm. Res.* **30**, 58–63 (2007).
- Nam, K.-W., Je, K.-H., Shin, Y.-J., Kang, S. S. & Mar, W. Inhibitory effects of furoquinoline alkaloids from *Melicope confusa* and *Dictamnus albus* against human phosphodiesterase 5 (hPDE5A) *in vitro*. *Arch. Pharm. Res.* **28**, 675–679 (2005).
- Li, C.-Y., Chow, T. J. & Wu, T.-S. The epimerization of sesamin and asarinin. *J. Nat. Prod.* **68**, 1622–1624 (2005).
- Ban, H. S. *et al.* Inhibitory effects of furanocoumarins isolated from the roots of *Angelica dahurica* on prostaglandin E2 production. *Planta Med.* **69**, 408–412 (2003).
- Lee, Y. Y., Lee, S., Jin, J. L. & Yun-Choi, H. S. Platelet anti-aggregatory effects of coumarins from the roots of *Angelica genuflexa* and *A. gigas*. *Arch. Pharm. Res.* **26**, 723–726 (2003).
- Han, B.-H., Park, M.-K. & Park, Y.-H. A Lignan from *Rubia akane*. *Arch. Pharm. Res.* **13**, 289–291 (1990).
- Li, T.-Z. *et al.* Studies on the lignans from *Patrinia scabra*. *Yao Xue Xue Bao* **38**, 520–522 (2003).
- Rutz, A. *et al.* Taxonomically informed scoring enhances confidence in natural products annotation. *Front. Plant Sci.* **10**, 1329 (2019).
- Lee, J. W. *et al.* Pyranocoumarins from *Glehnia littoralis* inhibit the LPS-induced NO production in macrophage RAW 264.7 cells. *Bioorg. Med. Chem. Lett.* **24**, 2717–2719 (2014).
- Kitajima, J., Okamura, C., Ishikawa, T. & Tanaka, Y. Coumarin glycosides of *Glehnia littoralis* root and rhizoma. *Chem. Pharm. Bull.* **46**, 1404–1407 (1998).
- Lin, R.-J. *et al.* Anthelmintic activities of aporphine from *Nelumbo nucifera* Gaertn. cv. *Rosa-plena* against *Hymenolepis nana*. *Int. J. Mol. Sci.* **15**, 3624–3639 (2014).
- Chen, I. S. *et al.* Coumarins and antiplatelet aggregation constituents from Formosan *Peucedanum japonicum*. *Phytochemistry* **41**, 525–530 (1996).

33. Chen, L. L. *et al.* Tissue-specific metabolite profiling on the different parts of bolting and unbolting *Peucedanum praeruptorum* Dunn (Qianhu) by laser microdissection combined with UPLC-Q/TOF-MS and HPLC-DAD. *Molecules* **24** (2019).
34. Kawai, T., Kinoshita, K., Koyama, K. & Takahashi, K. Anti-emetic principles of *Magnolia obovata* bark and *Zingiber officinale* rhizome. *Planta Med.* **60**, 17–20 (1994).
35. Abe, F. & Yamauchi, T. Lignans from *Trachelospermum asiaticum* (Tracheolospermum. II). *Chem. Pharm. Bull.* **34**, 4340–4345 (1986).
36. Szokol-Borsodi, L., Sólyomváry, A., Molnár-Perl, I. & Boldizsár, I. Optimum yields of dibenzylbutyrolactone-type lignans from *Cynareae* fruits, during their ripening, germination and enzymatic hydrolysis processes, determined by on-line chromatographic methods. *Phytochem. Anal.* **23**, 598–603 (2012).
37. Guo, H., Liu, A.-H., Ye, M., Yang, M. & Guo, D.-A. Characterization of phenolic compounds in the fruits of *Forsythia suspensa* by high-performance liquid chromatography coupled with electrospray ionization tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **21**, 715–729 (2007).
38. Kim, J. S. *et al.* Phenolic glycosides from *Pyrola japonica*. *Chem. Pharm. Bull.* **52**, 714–717 (2004).
39. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

## Acknowledgements

This research was supported by National Research Foundation of Korea (NRF) grants funded by the Korean Government (MSIT) (No. NRF-2020R1C1C1004046, 2021R1A2C1004958, 2021K1A3A1A21038059, 2022R1A4A3022401, and 2022R1A5A2021216), a Sookmyung Women's University Research Grant (1-2203-2014), and a Ministry of Food and Drug Safety grant in 2022 (21173MFDS561).

## Author contributions

K.B.K. and S.H.S. conceived the project. E.J. and S.S. acquired and processed the LC-MS/MS data of 223 phytochemicals. E.L. prepared the ReDU-compatible metadata for the medicinal plant dataset. K.B.K., S.Y.C., H.W.K., and H.Y. acquired the LC-MS/MS data for the medicinal plant extracts. S.L. and S.H.S. maintained and contributed compounds. K.B.K. performed technical validation and wrote the draft of the manuscript. All authors read and commented on the manuscript.

## 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/s41597-022-01662-2>.

**Correspondence** and requests for materials should be addressed to K.B.K. or S.H.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022