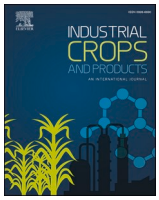
Industrial Crops & Products 231 (2025) 121159

Contents lists available at ScienceDirect

Industrial Crops & Products

journal homepage: www.elsevier.com/locate/indcrop



Clustering of chemical profiles of *Centella asiatica* cultivars, grown in greenhouses, allows grouping of metabolites with similar production trends

Md Nure Alam [a,b,1], Luke Marney [a,b,1], Liping Yang [a,b], Jaewoo Choi [a,b], Natasha Cerruti [a,c], Natascha Techen [a,d], Samuel Bassett [b], James Smith [e], Kevin Brown [a,f,g], Kadine Cabey [a,h], Ramya Viswanathan [a,h], Sumanaa Rajagopal [a,h], Amala Soumyanath [a,h], Jan F. Stevens [a,f,i], Claudia S. Maier [a,b,i,*]

[a] *Botanical Dietary Supplements Research Center (BENFRA), Oregon Health & Science University, Portland, OR 97239, USA*
[b] *Department of Chemistry, Oregon State University, Corvallis, OR 97331, USA*
[c] *Oregon's Wild Harvest, Redmond, OR 97756, United States*
[d] *National Center for Natural Products Research, University of Mississippi, University, MS 38677, USA*
[e] *School of Food Science & Nutrition, University of Leeds, Leeds LS2 9JT, United Kingdom*
[f] *Department of Pharmaceutical Sciences, Oregon State University, Corvallis, OR 97331, USA*
[g] *Chemical, Biological, and Environmental Engineering, Oregon State University, Corvallis, OR 97331, USA*
[h] *Department of Neurology, Oregon Health & Science University, Portland, OR 97239, USA*
[i] *Linus Pauling Institute, Oregon State University, Corvallis, OR 97331, USA*

ARTICLE INFO

Keywords:
Centella asiatica
Apiaceae
Metabolomics
Clustering
Bayesian hierarchical clustering
Caffeoylquinic acids
Triterpenoids

ABSTRACT

Centella asiatica (L.) Urban (also known as "gotu kola") is a perennial plant, used in traditional medicine for promoting resilience to central nervous system (CNS) disorders. *C. asiatica* is a tropical medicinal herb from the *Apiaceae* family and is native to Southeast Asian countries. The chemical composition and contaminant profile of commercial *C. asiatica* is variable. The goal of this study was to guide the future cultivation of organically grown *C. asiatica* for obtaining optimized plant materials for pre-clinical studies and clinical trials. Optimized plant materials in this case are defined as producing similar amounts of biologically active components as previously studied material. In this study, *C. asiatica* cultivars were grown in Central Oregon and their phytochemical compositions were examined. Four different cultivars were grown in climate-controlled greenhouses over three different vegetative propagation periods. Aerial parts of the plant were collected at four different harvest times: 8, 10, 12, and 14 weeks from growth initiation. The phytochemical composition of each cultivar was analyzed by liquid chromatography high-resolution tandem mass spectrometry (LC-HRMS/MS). Global metabolomic profiles allowed cultivar-specific compositional differences to be distinguished and production trends of phytochemical constituents to be analyzed using multinomial Bayesian hierarchical clustering and Self-Organizing Maps. Production trends of known bioactive phytoconstituents are reported here and will inform cultivation and harvest strategies to obtain *C. asiatica* materials of desired composition for preclinical and clinical studies. The computational methods for analyzing cultivar-specific and time-course dependent metabolomic profiles can be applied to other medicinal plant cultivation efforts to optimize cultivation and harvest practices.

1. Introduction

The therapeutic use of plants by humans reaches back approximately 60,000 years (Yuan et al., 2016). Traditional Chinese medicine, Ayurveda, Kampo, traditional Korean medicine, and Unani are different cultural medicinal practices that use herbal products for the treatment of diseases (Yuan et al., 2016). *Centella asiatica* (L.) Urban is a perennial plant belonging to the *Apiaceae* family reputed to have numerous health benefits in cosmetology, hepatoprotection, and neuroprotection (Orhan, 2012, Bylka et al., 2013, Intararuchikul et al., 2019). *C. asiatica* contains several classes of bioactive phytochemicals, including pentacyclic triterpenes (TTs), caffeoylquinic acids (CQAs), and flavonoids (Gray et al.,

* Corresponding author at: Department of Chemistry, Oregon State University, Corvallis, OR 97331, USA.
E-mail address: claudia.maier@oregonstate.edu (C.S. Maier).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.indcrop.2025.121159
Received 20 February 2025; Received in revised form 25 April 2025; Accepted 5 May 2025
Available online 10 May 2025

2018). CQAs and related secondary metabolites (di-CQAs, tri-CQAs, tetra-CQAs) are reported to have antioxidant and antiviral properties as well as to improve cognition and memory (Gray et al., 2018, Matthews et al., 2020, Gray et al., 2024). The biosynthetic routes leading to di-caffeoylquinic acid and derivatives include the shikimic acid and phenylpropanoid pathways (Magaña et al., 2021). Pentacyclic triterpenoids (TTs) from *C. asiatica*, commonly referred to as centelloids and produced via the isoprenoid biosynthesis pathway, have been reported to be neuroactive (Aharoni et al., 2005, James and Dubery, 2009, Soumyanath et al., 2010, Gershenzon and Kreis, 2018, Wu et al., 2020). Several factors can affect the production of these phytochemicals, such as harvest season, cultivation practices, processing method, growing region, and genetic variation (Long et al., 2012, Rahajanirina et al., 2012, Alqahtani et al., 2015).

For clinical uses, a standardized growing and harvesting protocol is important to ensure the consistent quality of plant material (Wright et al., 2022). Standardization using marker compound fingerprinting is recommended by the World Health Organization (WHO), European Medicines Agency (EMEA), and US Food and Drug Administration (USFDA) (Gajbhiye et al., 2016). Herein, we investigate an untargeted metabolomics-based analysis with novel data analysis approaches, along with the absolute quantification of phytochemical markers, to assist in the choice of plant variety and harvest time for future production.

C. asiatica is an important biologically active botanical. Our previous investigations demonstrated that an aqueous extract of *C. asiatica* (CAW), that is prepared by refluxing an aqueous suspension of the aerial parts, attenuates cellular oxidative stress (Gray et al., 2017) and improves cognition in an *in vivo* aged mouse model (Gray et al., 2024), and provides protection against β-amyloid toxicity both *in vitro* (Gray et al., 2017), and in an *in vivo* transgenic mouse model (Matthews et al., 2020). In recent phase-I human pharmacokinetics trials of CAW, aglycone triterpenoids and mono- and di-CQAs were present and quantifiable in plasma and urine (Wright et al., 2023).

The goal of this study was to develop a comprehensive and integrated phytochemical analysis strategy to assist in the production of consistent plant material for preparing CAW material for use in future preclinical and clinical studies. In previous work, we used liquid chromatography high resolution tandem mass spectrometry (LC-HRMS/MS) to obtain chemical profiles of aqueous methanolic extracts of *C. asiatica* and identified or tentatively assigned ~100 *C. asiatica* metabolites (Magana et al., 2020). In this work, we developed a suite of novel analytical and data-driven approaches to differentiate *C. asiatica* plant varieties grown in climate-controlled greenhouses in Central Oregon, USA, which yield measurable quantities of neuroactive compounds for sustainable access to reproducible research materials.

We used dimensionality reduction and clustering techniques to evaluate variations in the chemical compositions of four *C. asiatica* cultivars. In addition to the commonly used principal component analysis (PCA) (Brown, 1991, Brandolini et al., 2006), we applied singular value decomposition (SVD) (Kalman, 2002), which is, in theory, the same as PCA, and the non-linear reduction technique of uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), demonstrating that the choice of algorithm and implementation has an effect on biochemical conclusions and that testing multiple algorithms ensures that conclusions are supported in a thorough manner. We also applied Independent Component Analysis (ICA) (Hyvärinen et al., 2001) to obtain an alternative picture of independent sources of variation and enhance the reliability of the conclusions drawn from our untargeted metabolomics data. We found that the genetic variations of the *C. asiatica* cultivars led to varied chemical compositions detectable by these algorithms. We used feature-based molecular networks (FBMN) to summarize the variation of the phytochemical constituents in a cultivar-dependent manner (Wang et al., 2016). To investigate the time course production of metabolites, a novel implementation of unsupervised multinomial Bayesian hierarchical clustering (BHC) (Savage et al., 2009) was performed alongside Self-Organizing Maps (SOM) (Kohonen, 1982, Wittek et al., 2017). This study provides a metabolomic-centric framework for the controlled cultivation of *C. asiatica* varieties to guide and optimize cultivar selection, growth conditions, and harvest times.

2. Materials and methods

2.1. Chemicals

LC-MS grade methanol, water, and formic acid were purchased from Fisher Scientific (Hampton, NH, USA). Twelve authentic non-deuterated standards were purchased from Cayman Chemical (Ann Arbor, MI, USA). The compounds 4-O-caffeoylquinic acid (cryptochlorogenic acid, 4-CQA), 5-O-caffeoylquinic acid (5-CQA), 1,3-dicaffeoylquinic acid (1,3-DiCQA), 1,5-dicaffeoylquinic acid (1,5-DiCQA), 3,4-dicaffeoylquinic acid (isochlorogenic acid B, 3,4-DiCQA), 4,5-dicaffeoylquinic acid (isochlorogenic acid C, 4,5-DiCQA), and madecassoside (MS) had a purity of $\geq$ 98 %. The compounds 3-O-caffeoylquinic acid (chlorogenic acid, 3-CQA), 3,5-dicaffeoylquinic acid (isochlorogenic acid A, 3,5-DiCQA), asiaticoside (AS), madecassic acid (MA), and asiatic acid (AA) had a purity of $\geq$ 95 %. Digoxin-d_3 was used as an internal standard and was purchased from Cayman Chemical (Ann Arbor, MI, USA) with purity of $\geq$ 99 %. Digoxin-d_3 is toxic if inhaled or ingested, caution should be taken while working with digoxin-d_3.

2.2. Greenhouse cultivation of C. asiatica

Cuttings of four cultivars of *Centella asiatica* were acquired from four different commercial sources and grown in climate-controlled greenhouses at Oregon's Wild Harvest (OWH), Bend, OR, USA. Cultivars were propagated at multiples times during the year (Table 1). The aerial parts of the plant material were harvested after 8, 10, 12 and 14 weeks (Table 2), dried in a stainless food dehydrator at 50 °C (STX International Dehydra 1200W-XLS) to 5–10 % moisture, and stored at -20 °C until analysis.

Voucher Samples: Voucher samples of dried aerial plant material from each cultivar are stored in the BENFRA laboratories at Oregon Health & Science University (cultivar/plant material code names: BEN-CA-6, 9, 10, 11 and 12).

2.3. DNA barcoding analysis of C. asiatica varieties

DNA was extracted from the aerial plant material of the cultivar samples using the DNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. To amplify the Internal Transcribed Spacer (ITS) genomic region a previously published protocol, Primer Basic Local

Table 1
Start dates of cultivation and first harvest date (week 8) of the four *C. asiatica* cultivars in three groups each, corresponding to three propagation periods.

Cultivar	Plant Material Code	Group	Start Date (Planted)	1st Harvest Date (Week 8)
Mountain Valley	CA−9	1	6/18/2021	8/13/2021
Mountain Valley	CA−9	2	8/12/2021	10/7/2021
Mountain Valley	CA−9	3	9/9/2021	11/4/2021
Hawaii	CA−10	1	3/23/2021	5/17/2021
Hawaii	CA−10	2	4/22/2021	6/16/2021
Hawaii	CA−10	3	6/18/2021	8/13/2021
White Cloud	CA−11	1	4/22/2021	6/16/2021
White Cloud	CA−11	2	6/18/2021	8/13/2021
White Cloud	CA−11	3	8/12/2021	10/7/2021
9EZ	CA−12	1	6/18/2021	8/13/2021
9EZ	CA−12	2	8/12/2021	10/7/2021
9EZ	CA−12	3	9/9/2021	11/4/2021

Table 2
Experimental design accounting for three propagation periods per cultivar, four harvest points for each propagation group (weeks 8, 10, 12 & 14) and 3 replicate samples at each harvest.

Variety code/ cultivar	Origin	Number of propagation periods (groups)	Harvest times (weeks)	Number of Biological replicates	Number of samples
CA9	Mountain Valley	3	8, 10, 12, 14	3	35[a]
CA10	Hawaii	3	8, 10, 12, 14	3	36
CA11	White Cloud	3	8, 10, 12, 14	3	36
CA12	9EZ	3	8, 10, 12, 14	3	36
CA-6*	Commercial Product	1	1	1	1

* CA-6 is a commercial plant material of *Centella asiatica* (mixed origin) acquired through OWH (batch number X090016) from the supplier Organic India (batch number UFU0070), and was used as reference material; a – one replicate lost during sample preparation.

Alignment Search Tool – BLAST, was used (Altschul et al., 1990, Ye et al., 2012). The designed primers consisted of a 5' adapter sequence (pJet) facilitating direct sequencing with universal pJet primers of single PCR products. Primers (Table S1) were synthesized by Integrated DNA Technologies (Coralville). Extracted DNA was diluted to 1:10 and subjected to amplify the ITS region with the designed primers. Amplification consisted of two rounds of PCR. The first PCR consisted of a 25-µL reaction mixture containing 2 µL of a 1:10 DNA dilution, 1 × PCR reaction buffer, 0.2 mM dNTP mixture, 0.2 µM of each forward and reverse primers CentITSF and CentITSR, 1.5 mM $MgCl_2$, and 1 U of Platinum Taq DNA Polymerase (Invitrogen). The first PCR program consisted of an initial denaturation step at 94 °C for 3 min, followed by 10 cycles at 94°C for 80 sec, 58°C for 25 sec and 72°C for 1 min. The second PCR consisted of a 25-µL reaction mixture as mentioned above using 1 µL of the first PCR as a template and pJet forward/reverse primers. The second PCR program consisted of 35 cycles with denaturation at 94°C for 30 sec, 65°C for 20 sec, and 72°C for 1 min, with a final extension at 72°C for 3 min. After amplification, aliquots of 10 µL were analyzed by electrophoresis on 1.5 % borate agarose gel and visualized under UV light. PCR products were compared to the molecular size standard 1 kb plus DNA ladder (Invitrogen). The resulting PCR products were cloned into Vector pCR4-TOPO TA (Invitrogen) and transferred into *E. coli* according to the manufacturer's instructions. Eight transformants per cloning event were subjected to colony PCR using M13F and M13R primers. The PCR products of four transformants per DNA were sequenced with M13 forward and reverse primers at Genewiz/Azenta. Derived sequences were trimmed and subjected to homology search against the National Center for Biotechnology Information (NCBI) nucleotide database using BLAST available in Geneious Prime 2023.2.1 (Biomatters. LTD) (Altschul et al., 1990). Sequence alignment is shown in Figure S1. Additionally, a neighbor-joining consensus tree was built with the Tamura-Nei Distance model without an outgroup, bootstrap 500 replicates, available in Geneious Prime 2023.2.1.

2.4. Preparation of plant material extracts

The dried plant material was processed to a fine powder in a blade grinder. A suspension was prepared containing 0.5 mg/mL or 0.05 mg/mL plant material powder by adding in 70 % aq. (v/v) MeOH containing 0.1 % formic acid and 1 µg/mL of the internal standard digoxin-d_3 Plant material debris was removed by centrifugation.

2.5. Preparation of cultivar-specific samples and pooled samples

Each cultivar was represented in three propagation periods and was harvested at four time points, each time with three biological replicates, resulting in 36 samples for each cultivar (Table 2). In total, there were 143 samples for four cultivars, due to the loss of one of the samples. For analysis purposes, the samples were divided into two categories:

a) Cultivar samples: For a given cultivar, samples came from three groups (1, 2, 3) corresponding to propagation periods, four harvest points (8, 10, 12, 14 weeks), three biological replicates (a, b, c). Thus, samples were labeled with the cultivar number (CA-9, 10, 11 or 12) followed by propagation group, harvest week, and replicate letter e.g. CA-9–1–8a, CA-9–1–8b, CA-9–1–8c.

b) Pooled samples: for each of the four cultivars, pool samples were created by mixing extracts from all harvest week samples of the three groups, and a full pool sample was created by mixing the four cultivar pools and an extract from the control sample CA-6 (a commercial *C. asiatica* with a mixed origin). Pool samples were created using extracts made at 0.5 and 0.05 mg/mL original plant material (Fig. 1).

i) CA-9-Pool (all replicates from all harvest weeks and all propagation periods)

ii) CA-10-Pool (all replicates from all harvest weeks and all propagation periods)

iii) CA-11-Pool (all replicates from all harvest weeks and all propagation periods)

iv) CA-12- Pool (all replicates from all harvest weeks and all propagation periods)

v) Full Pool: CA-9-Pool + CA-10-Pool + CA-11-Pool + CA-12-Pool + CA-6.

Pooled samples served three purposes, a) quality control, b) normalization, and c) analysis of the metabolite differences between cultivars. To maintain linearity of response for phytochemicals at different abundances, pool samples and cultivar samples were prepared at two concentrations: 0.5 mg/mL and 0.05 mg/mL.

2.6. Preparation of standard solutions

Each standard stock solution of reference CQA and TT compounds was prepared at 1 mg/mL concentration in 95 % ethanol. A mixture of standards was prepared by mixing each standard in 95 % ethanol at a final concentration of 50 µg/mL. A mixture of standards was produced by diluting the standard stock solutions and 70 % methanol containing 0.1 % formic acid and the internal standard digoxin-d_3. The final concentration of this solution is 10 µg/mL of each standard with 1 µg/mL digoxin-d_3 as internal standard. Calibration curves were obtained by making a series of concentrations 1, 5, 10, 25, 50, 100, 250, 500, 1000, 2500, 5000, 10,000 ng/mL for all standards in 70 % methanol containing 0.1 % formic acid and 1 µg/mL digoxin-d_3. An extracting solution (70 % methanol containing 0.1 % formic acid and 1 µg/mL digoxin-d_3) was used to extract phytochemicals from plant material samples. A

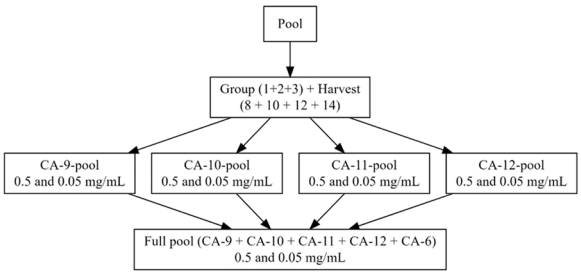


Fig. 1. Pool samples were prepared by combining timepoints from each cultivar as well as combining all cultivar pools into a Full pool sample. Samples were used for cultivar comparison as well as quality control during LC-HRMS/MS data acquisition.

simplified version of the entire workflow is shown in Fig. 2.

2.7. LC-HRMS/MS method

High-performance liquid chromatography (HPLC) was performed on a Shimadzu Nexera Ultra-HPLC system using an Intersil Phenyl-3 column (100 mm × 2.1 mm ID, 2 µm particle size, 100 Å pore size; GL Sciences, Torance, CA, USA). During gradient elution, the mobile phase consisted of water containing 0.1 % v/v formic acid as solvent A, and methanol containing 0.1 % v/v formic acid as solvent B. The chromatographic program was as follows: 0–0.1 min, 2 % B; 0.1–2.5 min, 2–15 % B; 2.5–3.0 min, 15–25 % B; 3.0–5.0 min, 25–35 % B; 5.0–6.0 min, 35–100 % B; 6.0–7.0 min, 100 % B return to 2 % B from 7.1 to 8.1 min. The column was held at 55 °C, and the flow rate was 0.8 mL/min. A 3 µL aliquot of each sample was injected for LC-HRMS/MS analysis. Data was collected using an untargeted data dependent acquisition (DDA) method on an AB SCIEX TripleTOF 5600 mass spectrometer equipped with a Turbo V ionization source. The desolvation gas temperature was kept constant at 550 °C. The instrument was operated in negative ionization mode. The MS1 and MS/MS scan ranged from *m/z* 70–1300 Table 3.

2.8. LC-MRM-MS method

An LC-MRM-MS method was used to obtain absolute quantitation of 12 phytochemical markers (Figure S2) on a Waters Xevo TQ-XS mass spectrometer coupled to a Waters Acquity UPLC I-Class system (Waters, Milford, MA). An Intertsil Phenyl-3 column (100 mm × 2.1 mm ID, 2 µm particle size, 100 Å pore size; GL Science, Torrance, CA, USA) was used to separate the CQAs and TTs. Gradient elution was identical to the LC-HRMS/MS method. The column temperature was held at 55 °C, with a flow rate of 0.8 mL/min. Injection volume for each sample was 1 µL. Desolvation and cone gas flow was maintained at 1000 L/h and 150 L/h respectively. Electrospray ionization was performed in negative mode. Spray voltage and desolvation gas temperature were held constant at 2300 V and 600 °C.

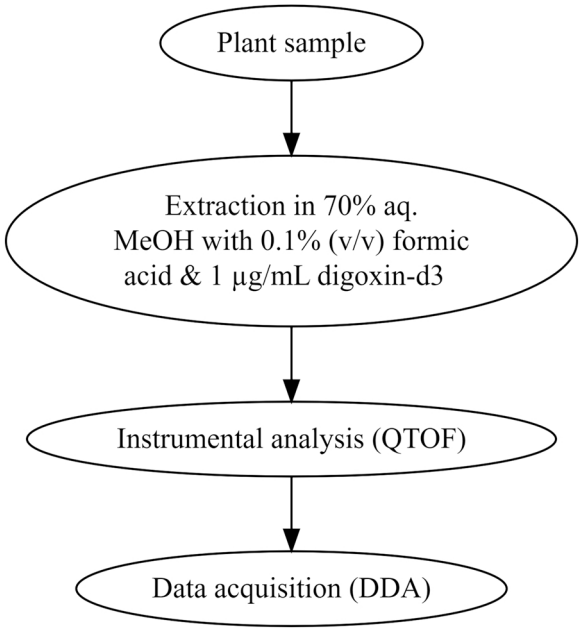


Fig. 2. Experimental design used for LC-HRMS/MS-based metabolomics of plant extracts (QTOF, quadrupole time-of-flight, DDA, data dependent acquisition).

Table 3
Feature *m/z* value, retention time, formula, and adduct information for the twelve marker compounds.

Name	*m/z*	RT (min)	Formula	Adduct
5-CQA	353.0778	1.4	$C_{16}H_{18}O_9$	[M - H]$^-$
4-CQA	353.0778	2.5	$C_{16}H_{18}O_9$	[M - H]$^-$
3-CQA	353.0778	2.6	$C_{16}H_{18}O_9$	[M - H]$^-$
1,3-DiCQA	515.1095	3.8	$C_{25}H_{24}O_{12}$	[M - H]$^-$
3,4-DiCQA	515.1095	4. 9	$C_{25}H_{24}O_{12}$	[M - H]$^-$
3,5-DiCQA	515.1095	5.0	$C_{25}H_{24}O_{12}$	[M - H]$^-$
1,5-DiCQA	515.1095	5.3	$C_{25}H_{24}O_{12}$	[M - H]$^-$
4,5-DiCQA	515.1095	5.6	$C_{25}H_{24}O_{12}$	[M - H]$^-$
MS	1019.495	6.2	$C_{48}H_{78}O_{20}$	[M + FA - H]$^-$
AS	1003.511	6.2	$C_{48}H_{78}O_{19}$	[M + FA - H]$^-$
MA	549.3321	6.38	$C_{30}H_{48}O_6$	[M + FA - H]$^-$
AA	533.3372	6.45	$C_{30}H_{48}O_5$	[M + FA - H]$^-$

2.9. Data analysis

For investigation of cultivar pool composition, a feature table consisting of retention time and abundances associated with untargeted features were obtained from 0.5 mg/mL and 0.05 mg/mL concentration data using Progenesis QI software. Principal component analysis (PCA), singular value decomposition (SVD), independent component analysis (ICA), and uniform manifold approximation and projection (UMAP) was performed with the feature table in R using the packages prcomp (Härdle et al., 2024), SVD (Kalman, 2002), ICA (Hyvärinen et al., 2001), and UMAP (McInnes et al., 2018). For time-course analysis, MultiQuant (Sciex) was used to integrate the peak areas of 12 marker compounds, the internal standard digoxin-d_3, and a set of 82 metabolites that were identified by accurate mass only (no retention time or fragmentation matching). These additional 82 masses are tentatively identified molecular features in *C. asiatica* that have been described previously (Magana et al., 2020). The method performance data for marker compounds and tentatively assigned molecular features are provided in the Supporting Information, Tables S2 to S5.

Time course data was initially split into early and late production subdivisions for clustering, but in the end utilization of a multinomial model with unsupervised Bayesian hierarchical clustering (BHC) (Savage et al., 2009) as well as self-organizing maps (SOM) (Kohonen, 1982, Wittek et al., 2017) allowed for data to be analyzed across the time range directly. Normalization was carried out with replicate pool samples with support vector regression (SVR) using the MetNormalizer R package (Shen et al., 2016). An overview of the data analysis approach is shown in Fig. 3.

Cosine similarity-based MS/MS clustering was used to generate molecular networks from both 0.50 mg/mL and 0.05 mg/mL data (0.05 mg/mL network not shown). Molecular networks are available publicly at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1786994c61db4463a93f7c6ddc113d21 on the Global Natural Products

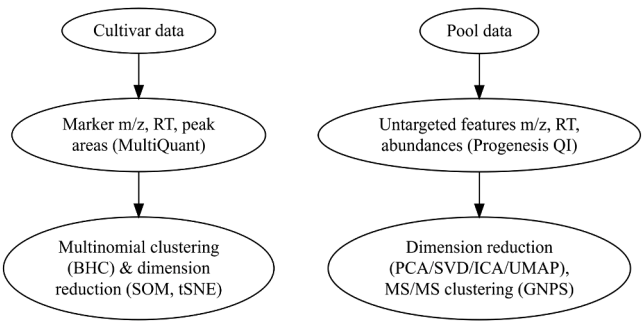


Fig. 3. Data processing workflows to investigate the compound variations observed in cultivars during different growth and harvest periods (left) and cultivar level distinction and MS/MS clustering in pool samples (right).

Social Molecular Networking (GNPS) platform (Wang et al., 2016). Edges were computed by cosine similarity at a 0.70 threshold with a minimum of six fragment ions matching. Cytoscape (version 3.9.1) was used to visualize molecular networks. Molecular structures were drawn using ChemDraw Professional 16.0.

3. Results and discussion

3.1. Phytochemical markers

Based on preclinical studies, we have assigned twelve consistently found compounds (Fig. 4) in *C. asiatica* extracts as bioactive phytochemical markers to facilitate the characterization of extracts and derived formulations (Yang et al., 2023).

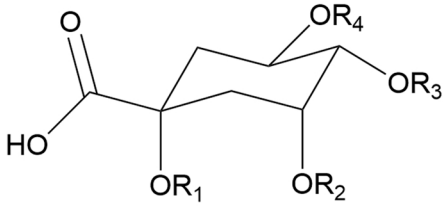
These phytochemical constituents were measured in four different *C. asiatica* cultivars grown in climate-controlled greenhouses in Central Oregon (Fig. 5) over a growth cycle of 8, 10, 12, and 14 weeks. Additional phytochemicals were measured by accurate mass and peak intensity as described previously (Magana et al., 2020).

3.2. DNA marker analysis confirms that all cultivars were from the genus *Centella*

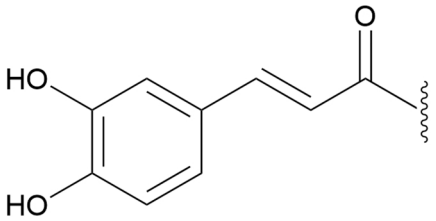
DNA marker analysis confirmed that all cultivars were from the genus *Centella*. The Internal Transcribed Spacer (ITS) marker region of each cultivar was determined and compared with available sequences in the NCBI database from the following three species: *C. asiatica, C capensis, and C. montana (NCBI)* (Fig. 6). Sequence alignment is shown in Figure S1. From the DNA marker analysis, all samples are identified as *Centella asiatica*.

3.3. Untargeted LC-HRMS/MS in conjunction with data reduction techniques reveals varietal differences in metabolite compositions

The four cultivars were grown and harvested at separate times of the year and analyzed using LC-HRMS/MS. Quality control (QC) pooled samples were created by combining samples from different harvest time points for a given cultivar and propagation period. QC samples were run twelve times during the LC-HRMS/MS analysis providing measurements

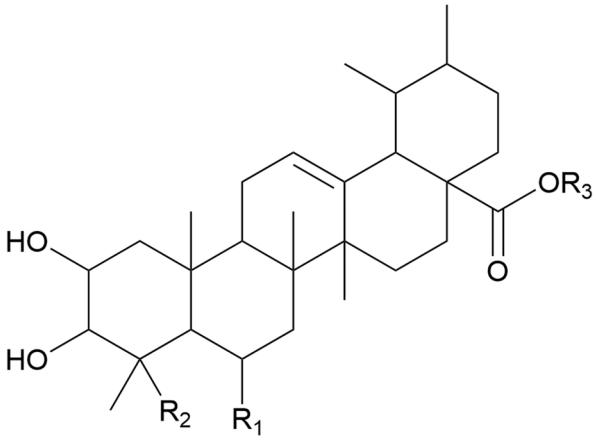


Quinic acid



Caffeoyl

	R_1	R_2	R_3	R_4
3-CQA (Chlorogenic acid)	H	caffeoyl	H	H
4-CQA (Cryptochlorogenic acid)	H	H	caffeoyl	H
5-CQA (Neochlorogenic acid)	H	H	H	caffeoyl
1,3-di-CQA	caffeoyl	caffeoyl	H	H
1,5-di-CQA	caffeoyl	H	H	caffeoyl
3,4-di-CQA	H	caffeoyl	caffeoyl	H
3,5-di-CQA	H	caffeoyl	H	caffeoyl
4,5-di-CQA	H	H	caffeoyl	caffeoyl



	R_1	R_2	R_3
Asiatic acid	H	CH_2OH	H
Madecassic acid	OH	CH_2OH	H
Asiaticoside	H	CH_2OH	Glucose-glucose-rhamnose
Madecassoside	OH	CH_2OH	Glucose-glucose-rhamnose

Fig. 4. Structures of mono-CQAs, di-CQAs, and triterpenoids used as phytochemical markers in this study (Chan et al. 2009, Orhan, 2012).

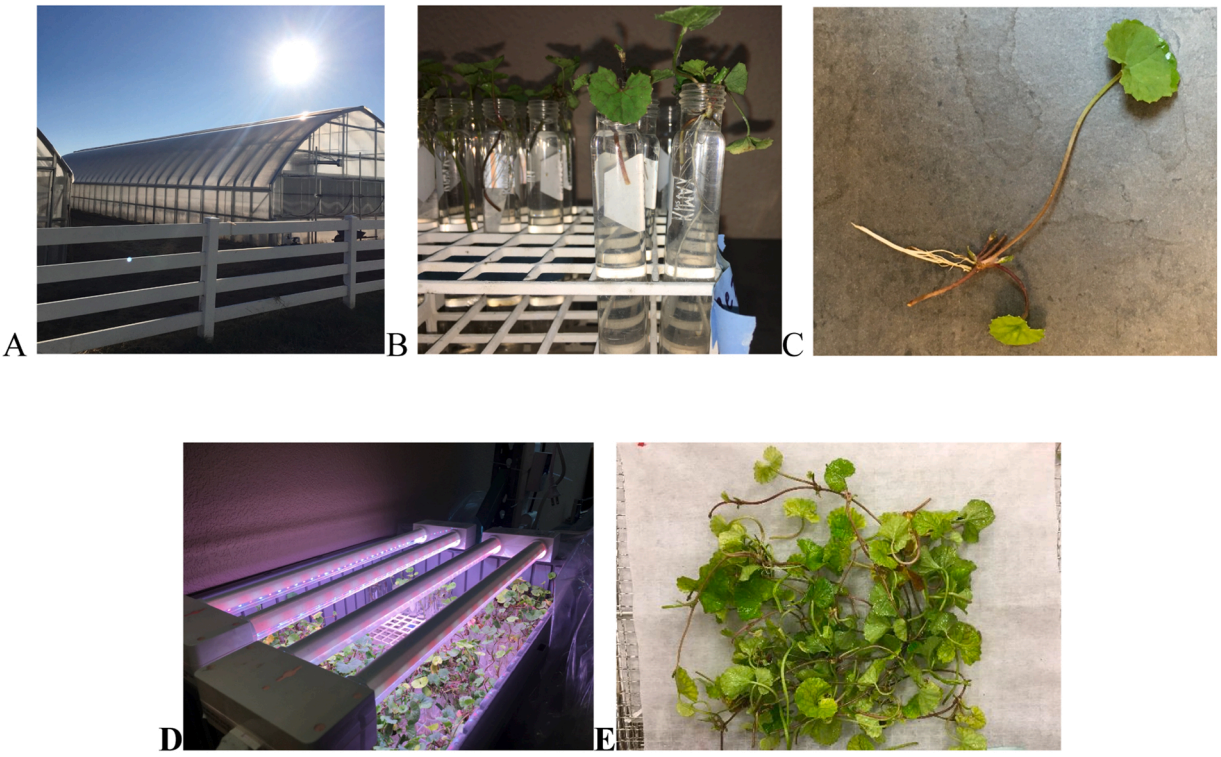


Fig. 5. *C. asiatica* cultivation pictures from Oregon's Wild Harvest in Redmond, Oregon. (A) Greenhouse used for *C. asiatica* cultivation: (B) Plant cuttings are rooted in distilled water with diluted fertilizer under growing lights at 22 ºC. (C) Roots started developing after one week. (D) Plantlets are potted and returned to the greenhouse after three weeks yielding (E) fresh plant samples.

for data normalization prior to untargeted analysis. Pool samples were analyzed by both LC-HRMS/MS and LC-MRM-MS (Supplemental figure S2, S3). Time-course samples were analyzed by LC-HRMS/MS. A representative 2D ion map (mass-to-charge vs. time (min)) is provided in Fig. 7.

Analysis of the untargeted LC-HRMS/MS data by dimensionality reduction techniques after normalization confirmed metabolic differences following the genetic origin of the plant cultivars. Initially, PCA on the untargeted pool samples showed that there were two distinct clusters separated based on data acquisition variation (Fig. 8**A**). This variation was attributed to analytical variation over the course of the four-day data acquisition. Analytical variation is caused by instrument drift that may arise from buffers, solvent systems, ion suppression, matrix effects and chromatographic column accumulation/degradation during long acquisitions. The systematic variation seen in Fig. 8**A** indicated that the ion signals of the two systematically grouped replicates is correctable by normalization.

Data pretreatment, such as transformation and normalization play a key role in metabolomics data analysis (van den Berg et al., 2006). Data transformation facilitates identifying underlying patterns and normalization allows for minimization of systematic variation (Misra, 2020). We used support vector regression (SVR) normalization to minimize the data variation (Fig. 8**B**). Sample-based SVR normalization considers the signal drift in QC as a representation of instrument drift over the injection period (Shen et al., 2016). During normalization, the full-pool sample was considered as a quality control (training dataset), and all others were considered as sample data (test set). After SVR normalization, the scores plot from PCA shows the differences in metabolite composition between the four different cultivar origins (CA-9, CA-10, CA-11, CA-12) and the commercial plant material sample (CA-6) (Fig. 8**B**).

Concentrations of the twelve marker phytochemicals were calculated using external calibration and internal standard normalization in both LC-HRMS/MS and LC-MRM-MS datasets (Table 4**)** and Figure S3). The parallel analysis of cultivar-pool and full-pool QC samples by LC-MRM-MS (Figure S2) did not show the same systematic variation seen in the LC-HRMS/MS data suggesting that the observed signal drift was related to the specific instrument used for analysis rather than sample degradation **(**Figure S2**)**. Accurate quantification and signal normalization ensured quality time-course data normalization and allowed additional benchmarking for the untargeted analysis. Relative standard deviation (RSD) before and after SVR normalization for the complete dataset are provided in the supplementary materials showing the effect of normalization on all data (Figure S4).

In all applied data reduction techniques (Fig. 9), cultivars were separated by multivariate dimensions. The cultivars CA-9 and CA-10 are consistently clustered closest to the previous standard material CA-6 and considered as future candidates for sourcing the plant materials for clinical trials, although other physical (i.e. non-chemical) growth parameters may need to be considered for selecting the optimum cultivar. A water extract of CA-6 was shown previously to promote resilience to stress or age-related neurological changes, as well as amelioration of age-related cognitive decline and anxiety in mice (Gray et al., 2017, Gray et al., 2024). Visual presentations of the PCA of the 0.5 mg/mL cultivar-pool and full-pool QC samples are shown in Figures S5-S6.

We note, when PCA is performed in R using prcomp, it relies on SVD under the hood to transform the data into a new coordinate system where the first few axes explain the most variance. SVD "uncovers" the underlying structure of the data by decomposing it into three matrices. PCA selects the top rows from the right singular vectors that capture the most variance. When we look at the resulting scores plot, it might appear slightly different than what one would expect if we were to apply SVD directly. In another sense, PCA is equivalent to running SVD with a specific choice of parameters and can result in slightly different scores plots.

Independent Component Analysis (ICA) aims to separate mixed signals into their original sources. ICA treats each data point as a mixture of several independent components and attempts to recover them using

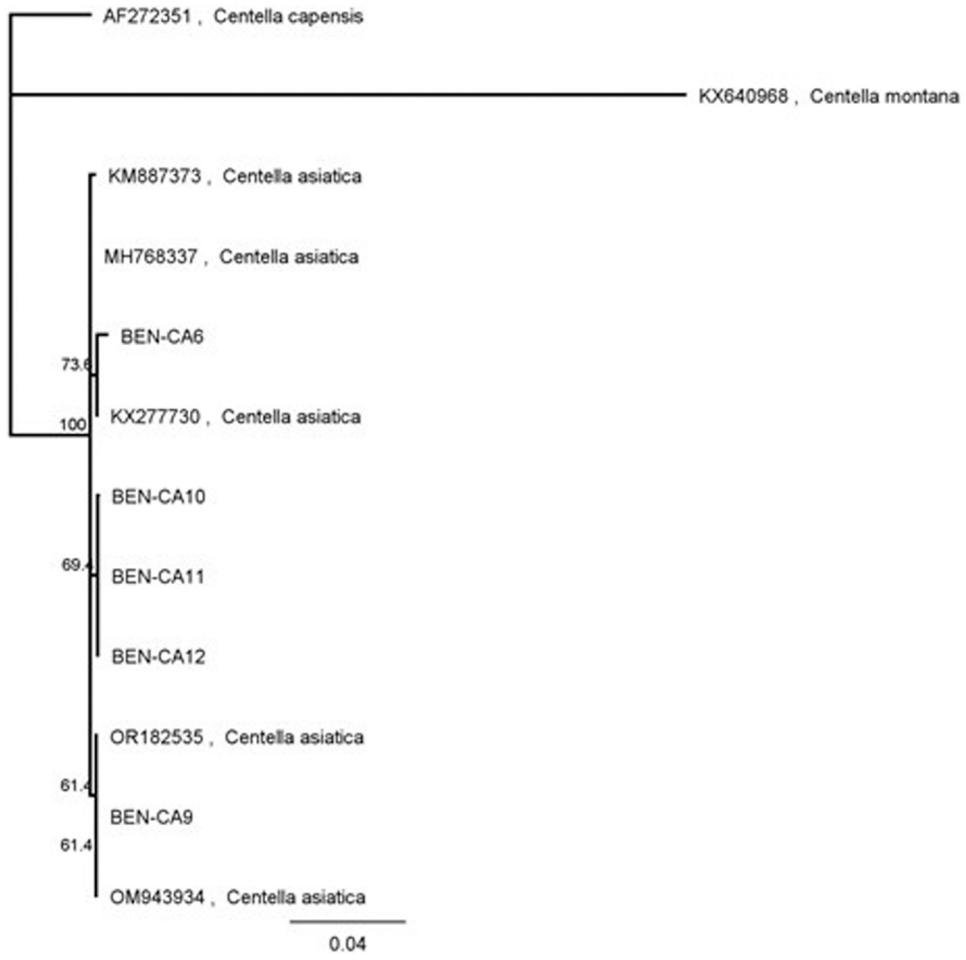


Fig. 6. Neighbor-joining tree with five hundred bootstraps using the Tamura-Nei genetic distance model without an outgroup for sample BEN-CA6, BEN-CA9, BEN-CA10, BEN-CA11, and BEN-CA12 confirmed as *Centella asiatica* when aligned with previously published sequences (KM887373, KX277730, MH768337, OM943934, and OR182535).

only non-linear transformations. While PCA focuses on explaining variance in the original data space, ICA seeks to extract underlying patterns from the mixture. UMAP (Uniform Manifold Approximation and Projection) uses a combination of PCA and t-SNE (t-distributed Stochastic Neighbor Embedding) to create a lower-dimensional representation while preserving local relationships between points in the high-dimensional space.

3.4. Compositional- and varietal-annotated molecular networks confirm presence of neuroactive metabolites in greenhouse-grown Centella cultivars

Cosine similarity-based MS/MS clustering assisted in identification of structurally similar features and their relative abundance in specific cultivars. Untargeted data are primarily used for hypothesis generation, global analysis, MS/MS correlation with databases, qualitative identification, and relative quantification (Schrimpe-Rutledge et al., 2016). To additionally reduce the complexity of metabolite-level information to specific metabolite changes, we performed cosine similarity-based MS/MS clustering on the 0.5 mg/mL cultivar-pooled data which allows for visual interpretation of relative abundances of specific known phytochemicals as well as unknown but spectrally related feature abundances (Fig. 10) (Bittremieux et al., 2022).

The GNPS network was analyzed and processed by Cytoscape to provide annotation according to cultivar information (Smoot et al., 2011). The network consists of 1800 nodes, 2016 edges, and 279 connected components.

Among all identified features, we labeled ten compounds using authentic standard retention times and masses. Other compounds in *C. asiatica* have been detailed previously (Magana et al., 2020). There are several possible reasons for the inability to annotate features: the chemical structure is available on the database, but there are no tandem MS/MS fingerprints available; database ion fragments can vary when data is collected with different instruments using different ionization modes (i.e., ESI+ or ESI-); multiple features arising from one metabolite due to formation of adducts (i.e., metal, solvent, ion pairing reagent), formation of multimer (Stefansson et al., 1996), in-source ion fragments (Xu et al., 2015), or from metabolic degradation products; in addition to other features that appear with unique retention times. Additionally, unknown features can also arise from three types of reactions during analysis: unimolecular reactions (racemization, rearrangement, elimination, and photolysis), reactions with oxidants, and other metabolites (Hanson et al., 2016). While their exact chemical name and structure may not be identified, they still contribute to our understanding of phytochemical production processes in the plant using untargeted metabolomics methods, helping to possibly extend our knowledge beyond the recommended guidelines from the WHO, EMEA, and USFDA for use in cultivation optimization of medicinal plants.

Here, mono- and di-caffeoylquinic acids are found in the same cluster except for 1,3-dicaffeoylquinic acid, which occurs as a single node. The triterpenoid glycosides, asiaticoside and madecassoside, are found in the same cluster with an additional five feature nodes. Aglycones occurred as single nodes, indicating there are fragment ion variations among glycosidic and aglycone triterpenoids. Using marker retention times, we identified hydrophobic aglycone triterpene asiatic and madecassic acids

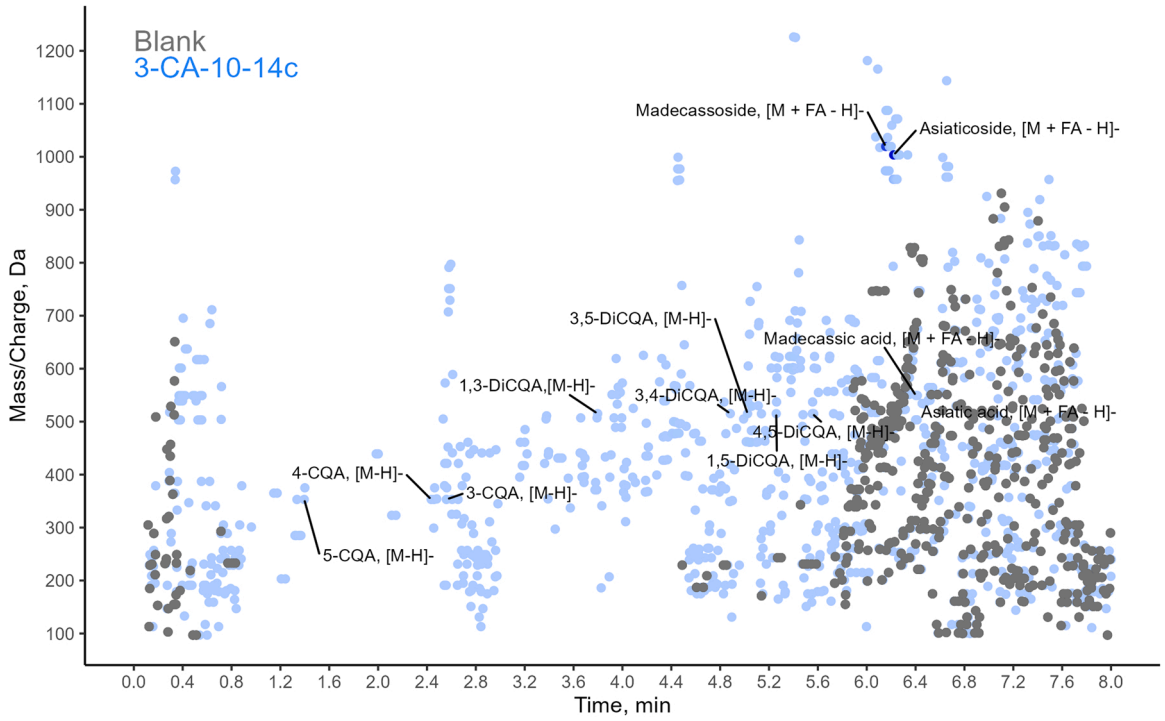


Fig. 7. 2D ion map showing the ion signals of the 12 markers in sample CA-10 (Hawaii). Blue dots represent features from 3-CA-10–14c sample, chosen as a representative sample, in contrast to dark grey dots representing a blank. Retention times were validated using authentic standards. Data obtained using the LC-HRMS/MS method described in 2.7.

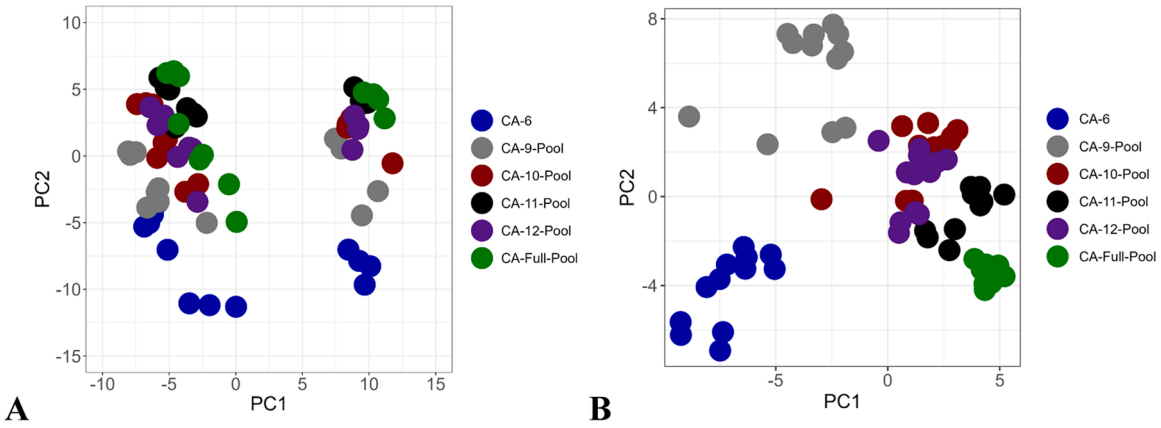


Fig. 8. Comparison of PCA score plots of (A) raw data and (B) support vector regression (SVR) normalized data of the 0.05 mg/mL cultivar-pooled samples and full-pool QC sample (12 replicates each). Data were natural log transformed and centered without scaling.

Table 4
Concentration of marker compounds (μg/mg plant material) in cultivar-pooled samples obtained by LC-HRMS/MS (using the 0.50 mg/mL extract data of 3 replicates).

Name	CA-6	CA-9	CA-10	CA-11	CA-12
5-CQA	0.066 ± 0.018	0.05 ± 0.01	0.05 ± 0.01	0.04 ± 0.01	0.05 ± 0.0
4-CQA	0.0050 ± 0.0012	0.0028 ± 0.0004	0.0030 ± 0.0006	0.0022 ± 0.0004	0.0026 ± 0.0006
3-CQA	2.02 ± 0.60	1.50 ± 0.32	1.00 ± 0.24	1.26 ± 0.44	1.90 ± 0.74
1,3-DiCQA	0.009 ± 0.002	0.0034 ± 0.0004	0.0058 ± 0.0012	0.0036 ± 0.0006	0.0032 ± 0.0004
3,4-DiCQA	0.11 ± 0.03	0.036 ± 0.008	0.072 ± 0.022	0.040 ± 0.022	0.044 ± 0.026
3,5-DiCQA	1.56 ± 0.50	0.96 ± 0.18	1.20 ± 0.30	0.96 ± 0.36	1.10 ± 0.42
1,5-DiCQA	1.54 ± 0.46	0.70 ± 0.14	0.64 ± 0.14	0.70 ± 0.26	0.96 ± 0.38
4,5-DiCQA	0.56 ± 0.16	0.13 ± 0.024	0.17 ± 0.04	0.132 ± 0.048	0.144 ± 0.058
MS	5.60 ± 1.44	5.66 ± 0.48	4.82 ± 0.56	5.22 ± 0.44	5.06 ± 0.74
AS	4.7 ± 1.14	9.34 ± 0.78	5.40 ± 0.52	6.62 ± 1.60	7.04 ± 1.64
MA	0.26 ± 0.08	0.074 ± 0.010	0.086 ± 0.018	0.078 ± 0.016	0.092 ± 0.018
AA	0.078 ± 0.028	0.028 ± 0.012	0.044 ± 0.008	0.042 ± 0.01	0.04 ± 0.006

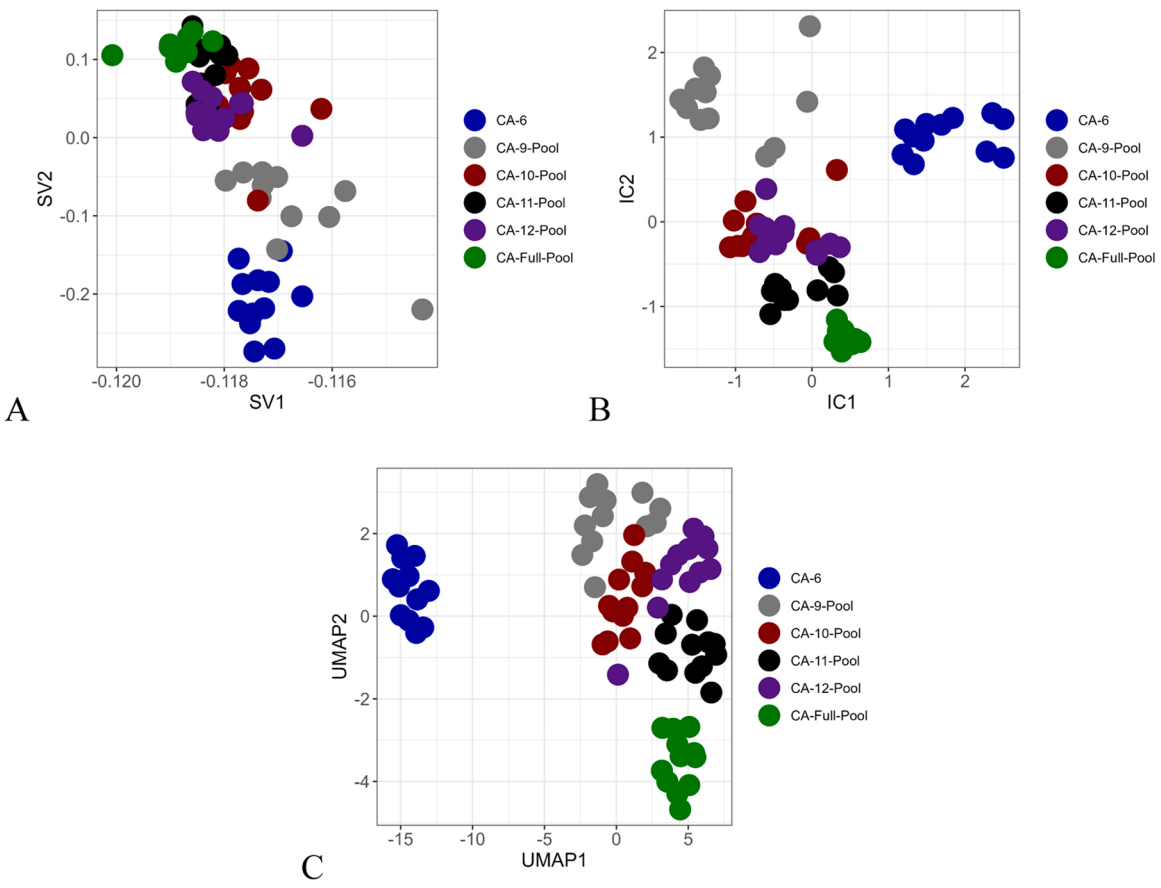


Fig. 9. LC-HRMS/MS data of the cultivar-pooled samples (0.05 mg/mL) analyzed by (A) singular value decomposition (SVD), (B) independent component analysis (ICA), and (C) uniform manifold approximation and projection (UMAP) after SVR normalization.

as single nodes. Additional centelloids, such as asiaticoside B, centellasaponins C and A, centellasapogenol A, isothankunic acid, madasiatic acid, and betulinic acid, have been reported in literature (James and Dubery, 2009). However, the identification of these triterpenoids remained a challenge due to lack of authentic standards. In addition, co-eluting molecular ions with overlapping retention times and MS/MS fragment ion patterns, as well as in-source ion fragmentation pose additional analytical challenges in providing proper annotation to unassigned features (Xu et al., 2015). Data is made available for subsequent mining and annotation through MetaboLights (MTBLS11735). Additionally, the MS/MS clustering molecular network from 0.05 mg/mL samples generated 1372 nodes 1494 edges with 214 connected components (network not shown, but available publicly at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1786994c61db4463a93f7c6ddc113d21).

The number of mass spectrometry features associated with each cultivar sample is in the thousands. MS/MS clustering-based molecular networking minimized data complexity by clustering structurally similar MS/MS data, allowing semi-quantitative comparison of feature abundances in four cultivars. MS/MS for each of the twelve-markers and their chromatographic elution profiles are shown in Figures S8-S19. We paired a focused look at a set of specialized metabolites that have been associated with the observed neurological bioactivity of *C. asiatica* extracts (Gray et al., 2017), namely three mono-CQAs, five di-CQA, and four TTs, for which authentic standards were available, with an untargeted approach to metabolomic phytochemical analysis.

3.5. Untargeted metabolomics and Bayesian hierarchical clustering reveals production time-course for selected bioactive compounds for each greenhouse-grown Centella cultivars

To understand metabolite changes in time, MS1 peak areas were integrated from a list of 94 compounds, including 12 marker compounds verified with authentic standards by retention time, accurate mass, and fragmentation spectra. Other peaks are consistently found features in our *C. asiatica* extracts (Magana et al., 2020). Feature peak areas were integrated using Sciex Multiquant software. The time-dependent profiles were analyzed by unsupervised non-parametric multinomial Dirichlet process Bayesian hierarchical clustering (BHC, Fig. 11). In traditional hierarchical clustering, each pair of objects is chosen based on pairwise distance difference (i.e., Euclidean, Manhattan) or pairwise norm-based similarity (e.g. Tanimoto, correlation, cosine), however, it does not support an optimal number of clusters for pruning trees and the choice of metric is arbitrary and complicated further for time series. BHC (R/BHC doi:10.18129/B9.bioc.BHC) is advantageous in this situation as it uses a Dirichlet process approach to merge clusters. The R/BHC algorithm uses the predictive distribution of test points to assign a point to an existing cluster in the tree from a prior probabilistic model of the data, the marginal likelihood.

In the final heatmap presentation from R/BHC, the most representative hierarchy informs the relation of mass spectrometry signals to the underlying distribution of signal. The contributions in the heat map are visualized from the distribution as a color scheme: green indicating the lower bound, black denoting the marginal likelihood, and red indicating the upper bound. Metabolites in the lower and upper bounds are considered silent contributions and do not contribute to the clustering. Specifically for this analysis, the resulting lower bound (green)

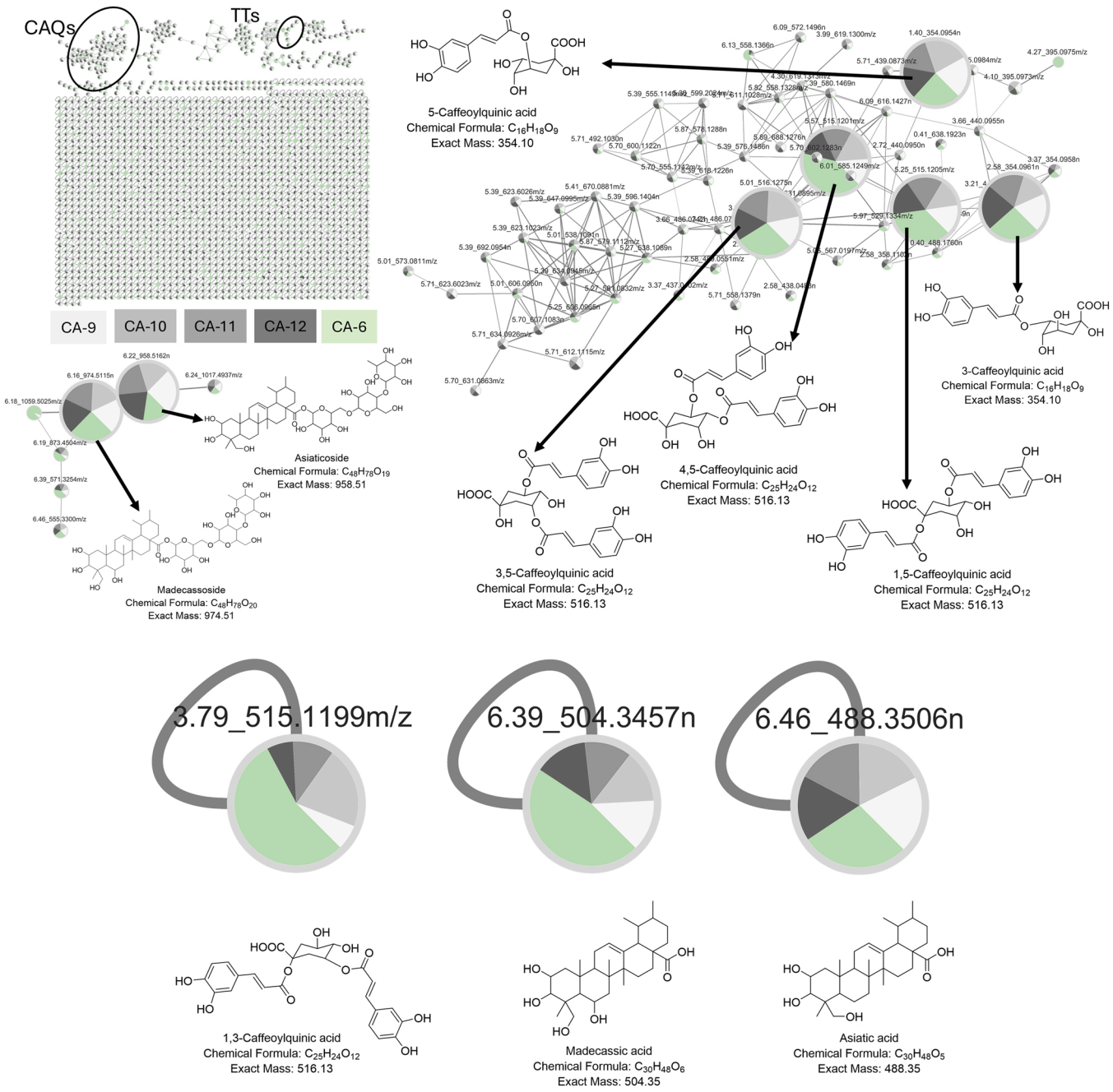


Fig. 10. Molecular network built using the four cultivar pool data (0.5 mg/mL extracts) with annotation of 10 marker compounds. The nodes are labeled as progenesis IDs (RT_Mass and *m/z* for measured mass and "n" for neutral mass where adducts of a feature have been combined.) Pie charts show the cultivar specific abundances. The edges are weighted to cosine similarity.

metabolites are more dispersed, while the upper bound metabolites are more tightly grouped. This visualization contributes to interpreting the results, identifying metabolites that exhibit similar trends. While the algorithm chooses the final number of clusters (Table 5**),** more confidence can be assigned to the final fusion at the top of the dendrogram distinguishing the upper bound (red) metabolites from the other two groups. For instance, madecassoside is quantified with baseline resolution and high concentration and has similar production trends with asiaticoside, which both are in the red upper bound and are represented in Cluster 1. Also in Cluster 1 are the metabolites asiatic acid and madecassic acid, which contribute to the final clustering yet are placed within the marginal likelihood and show an even more visually similar production in time. In comparison, 1,3-DiCQA is found in Cluster 5 within the lower bound (green) potentially due to its lower concentration and lower signal-to-noise in many samples and is grouped less confidently together with other metabolites.

Interestingly, four marker compounds 1,5-DiCQA, 3,5-DiCQA, asiaticoside, and madecassoside along with flavonoids and primary metabolites are in Cluster 1. Cluster 2 metabolites are grouped together centrally with high confidence too and all follow similar time-series concentration profiles. Representative concentration profiles for each likelihood bounds are shown in Fig. 12 indicating how metabolites that share similar production rates across different groups and harvest periods are grouped together. Interestingly, even those metabolites that are deemed in the lower-bound do follow similar concentration profiles across cultivation. Since the data is sorted by using a *Dirichlet* process that proceeds down a hierarchy, the outlying metabolites that do not fit into the previous distinct clusters can end up being similar noise terms

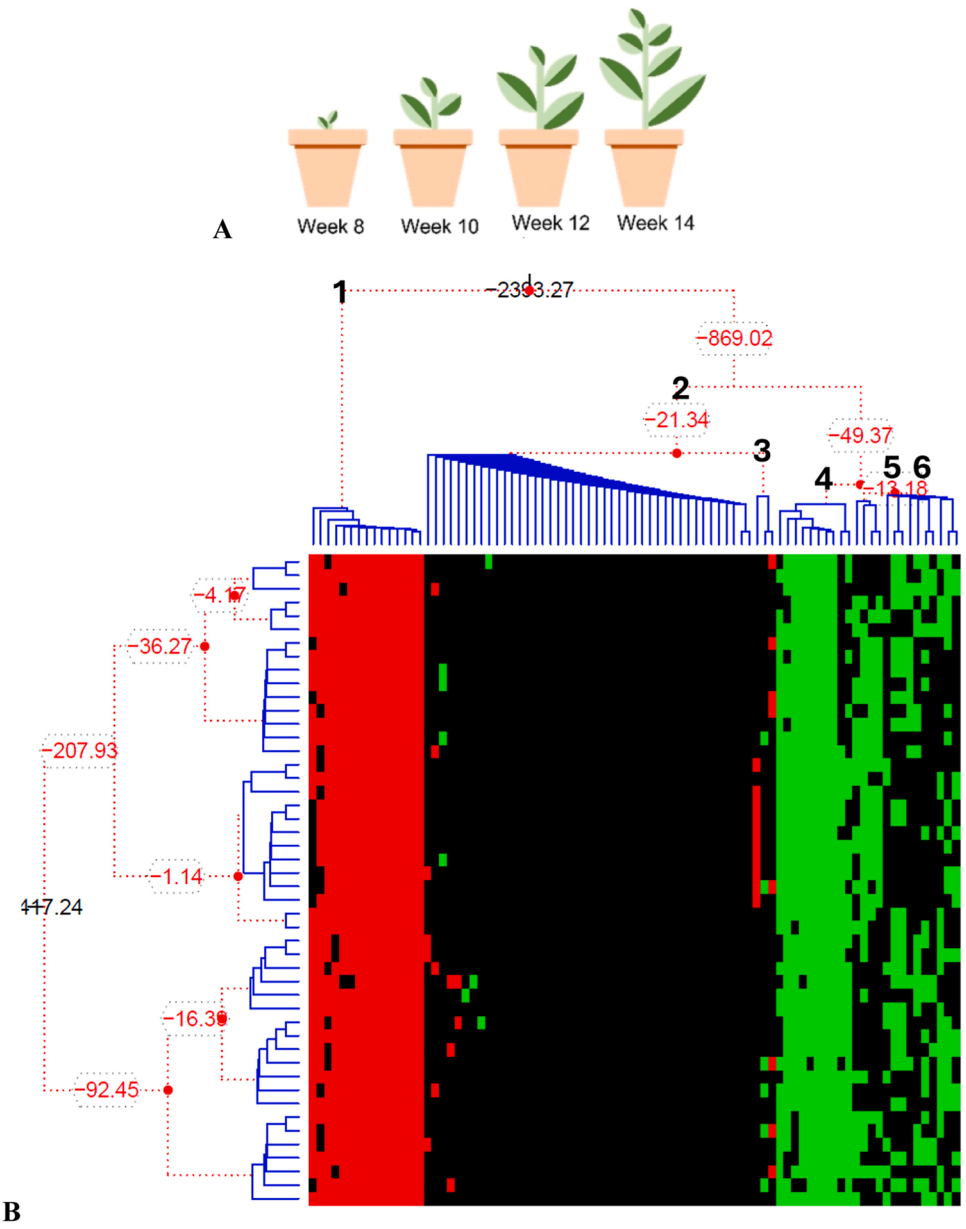


Fig. 11. Non-parametric multinomial clustering using BHC of (A) the time-course production in weeks of 94 metabolites, including 12 authentic marker compounds. Color coding of the heat map (B) is based on the marginal likelihood (black) of the time series of profiles (grouped on the y-axis) to reveal the quality of metabolites according to their final cluster membership (grouped on top on the x-axis). Green represents the lower bound and red represents the upper bound. Cluster numbers indicated on x-axis hierarchical tree and hyperparameter tuned values defined for tree branching are indicated in each branch, black numbers indicate a significant branching difference while red numbers are not as significant.

and may warrant further investigation.

The results here suggest that metabolites measured by an untargeted approach and with a distinct production cycle over time can indeed be clustered together. The utility of BHC is to find groups of metabolites with similar production trends and potentially aid in optimizing growth conditions for desired metabolite content in the future. Clustering occurs on both data axes. The sample clustering is shown in Figure S20. For reassurance, a bar plot of the sample amount was also analyzed (Figure S7) and the metabolite concentration changes were found to be unrelated to the analyzed sample amount by mass. A true change in concentration of these compounds occurs across different harvests and growth periods as measured by peak area. This implementation of BHC is highly novel in this phytochemical context.

3.6. Self-organizing maps (SOM) and t-distributed stochastic neighbor embedding (tSNE) confirm clustering of marker phytochemicals with similar time-course production across cultivar species and propagation groups

While peak-area information is useful to cluster and identify groups of untargeted metabolites that have similar time-course trends, specific concentration of metabolites through time is directly applicable to cultivation. Calibration curves normalized to the internal standard digoxin-d_3 provide accurate concentrations that can be used to identify trends (Figure S3). The concentration of 12 metabolite markers through time were analyzed by Self-Organizing Maps (SOM) (Kohonen, 1982) and t-distributed stochastic neighbor embedding (tSNE) (van der Maaten and Hinton, 2008) to cluster marker phytochemicals into similar time-course production across cultivar species and propagation groups.

Table 5
List of metabolites (tentatively assigned) (Magana et al. 2020) in the final clusters derived from BHC.

Cluster	Metabolite
1	3,5-Dihydroxyphenyl 1-O-(6-O-galloyl-beta-D-glucopyranoside), quercetin, mangiferin, 1,4-dicaffeoylquinic acid, tetradecanedioic acid, palmitic acid, malate, citric acid, asiatic acid, madecassic acid, asiaticoside, madecassoside, 1,5-dicaffeoylquinic acid, 3O-caffeoylquinic acid, 3,5-dicaffeoylquinic acid
2	tropic acid, succinate, soyacerebroside-I, quercetin−3-O-glucoside, pelargonidin−3-O-glucoside, naringin, isoferulic acid, glabraoside A, ginkgoic acid, furaneol−4,6-malonylglucoside, epicatechin, dihydroferulic acid, daucic acid, catechin, 5-O-caffeoylquinic acid, 16-hydroxypalmitic acid, kaempferol, 12-oxodihydrophytodienoic acid, 2-pyrrolidone−5-carboxylic acid, 3,4-dicaffeoylquinic acid, quercetin 3-(6”-acetylglucoside), succinoadenosine, rutin, adenine, L-ribulose, xanthurenic acid, uric acid, sambacin, phlorin, nomilinic acid 17-glucoside, L-arginine, kuwanon Y, isovalerylglucuronide, ginsenoyne K, gentiopicroside, folinic acid, dysolenticin B, deoxyfructosazine, 8-acetoxy−4’-methoxypinoresinol 4-glucoside, 5-methoxy-L-tryptophan, 5’-deoxy−5’-(methylsulfinyl)adenosine, 3,5-dihydroxy−2-methylphenyl-beta-D-glucopyranoside, tsangane L 3-glucoside
3	Stachyose, carlosic acid methyl ester, traumatic acid
4	guanosine, dihydroactinidiolide, N-acetyl L-glutamic acid, aesculin, 1-caffeoyl−5-feruloylquinic acid, enicoflavine, 4-O-caffeoylquinic acid, caprylic acid, 26-(2-Glucosyl−6-acetylglucosyl]−1,3,11,22-tetrahydroxyergosta−5,24-dien−26-oate, 3,4-dihydroxybenzaldehyde
5	1,3-dicaffeoylquinic acid, dihydrocaffeic acid, adenosine, pantothenic acid
6	caffeic acid, b-chlorogenin 3- [4” -(2”’-glucosyl−3”’-xylosylglucosyl) galactoside]e 3-hydroxy−2-oxo−3-phenylpropanoic-acid, succinyl L-proline, N-(1-deoxy−1-fructosyl) phenylalanine, 2-O-methyladenosine, linustatin, digalacturonate, kynurenic acid, N1, N5, N10, N14-tetra-trans-p-coumaroylspermine

The application of SOM to our metabolite dataset revealed distinct clusters of metabolites with similar time-course trends (Fig. 13). Specifically, the metabolites MS, AS, 3,5-DiCQA, and 3-CQA are clustered together (and are also close in the tSNE embedding) reflecting similar concentrations throughout propagation and harvest timing. Looking back to Fig. 12 we see the time-dependent changes of concentration for MS and AS. Other metabolites were clustered, but not as confidently. The metabolites AA, 3,4-DiCQA, 4-CQA, 1,5-DiCQA, and 4,5-DiCQA were grouped with a larger distance between neurons than the previous cluster, but still like each other. These findings underscore the utility of SOM in discerning and visualizing temporal patterns in metabolite data, offering significant insights into the metabolic changes throughout the harvest period. The utility may be that surrogate markers can be used for propagation and harvest timing, such as using AA to profile one cluster, 3,4-DiCQA for another, and 5-CQA for the third. But further research is likely needed to simplify the analysis in this way.

SOMs are a type of artificial neural network that is particularly effective for visualizing high-dimensional data. By mapping the high-dimensional metabolite data onto a two-dimensional grid, SOM clusters metabolites that exhibit similar temporal patterns. This is achieved through an iterative process where the map’s nodes, or neurons, adjust their weights to match the input data, thereby grouping similar data points closer together on the grid. Eventually, the grid will approximate the data distribution. In our specific case, metabolites with similar time-course trends are positioned in proximity on the SOM grid, facilitating the identification of patterns and trends that might not be immediately apparent in the raw data. Additionally, the topological height, represented by the color map in Fig. 13**A,** of each neuron in a unified distance matrix (U-matrix) representation of a self-organizing map shows the distance between neighboring neurons, a measure of how similar or dissimilar the neuron’s weight vectors are.

Compared to other clustering tools, SOM offers unique advantages and some limitations. One of the primary advantages of SOM is its ability to preserve the topological properties of the input data, meaning that the spatial relationships between data points are maintained in the output map. This makes SOM particularly useful for visualizing complex, high-dimensional datasets. Additionally, SOM provides a clear and intuitive visual representation of the clusters, which can be beneficial for interpreting the results.

SOM also has limitations, including the need to predefine the grid

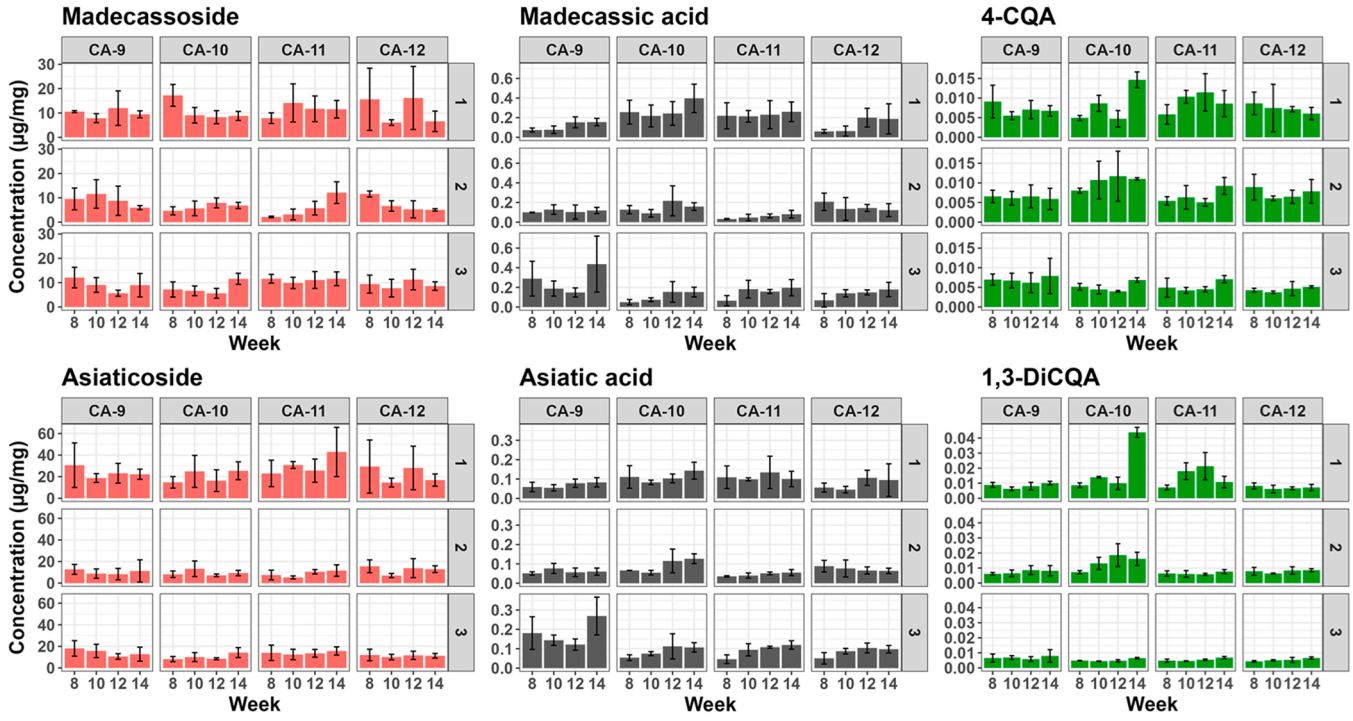


Fig. 12. Marker compound bar plots of madecassoside, asiaticoside, madecassic acid, asiatic acid, 4-CQA, and 1,3-DiCQA showing calculated concentrations over the time-course and how that relates to likelihood assignment of authenticated standard phytochemicals. Of note is the similarity of time-course data within each likelihood bounds and cluster.

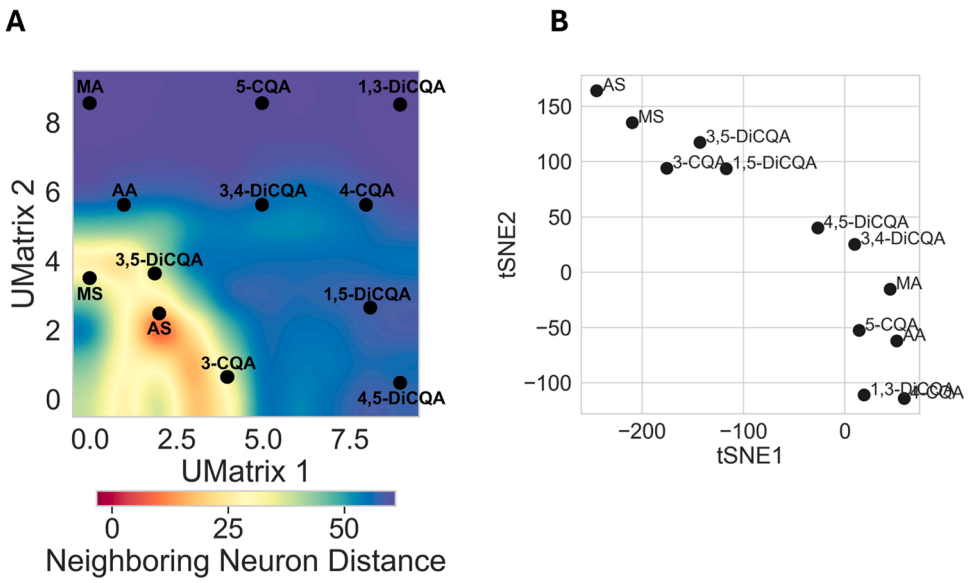


Fig. 13. Time course analysis by self-organizing maps (SOM) and t-distributed stochastic neighbor embedding (tSNE). (A) The unified distance matrix (U-matrix) representation of a self-organizing map calculated from the 12 phytochemical marker compounds throughout the time course experiment from four cultivars, three groups, and four harvests. (B) tSNE of the same dataset. In both cases, metabolites with similar time-course trends cluster together, represented by spatial arrangement in a two-dimensional space in both tSNE and SOM. SOM displays neighboring neuron distance indicating which metabolites cluster better (i.e. smaller distance equates to similarity).

size and topology, which can influence the results. Additionally, in both SOM and tSNE, peak area or concentration data is transformed onto a new two-dimensional representation that does not preserve the original peak abundance information, but rather acts as a simplification or monogram of the data. BHC does not do this. Unlike t-SNE, which is primarily used for dimensionality reduction and visualization and can be followed by clustering techniques like K-means clustering, SOM advantageously performs clustering and visualization simultaneously. SOM's ability to provide a detailed and visually interpretable clustering makes it a valuable tool for metabolite analysis. The combination of multiple techniques allows us to arrive at the most descriptive understanding of the cultivation of these plants and their phytochemicals.

The novel application of Bayesian hierarchical clustering (BHC) identified trends in production at different growth and harvest periods for 94 metabolites. These findings were corroborated by SOM and tSNE of the measured concentrations of 12 phytochemical markers. By investigating specific metabolite signals by cluster, we represent the metabolite changes through the growth cycle of each cultivar and can make future informed decisions about production and monitoring of these important phytochemicals.

Cultivation and propagation of different cultivars of *C. asiatica* followed by LC-HRMS/MS, LC-MRM-MS, data reduction techniques, GNPS, a novel application of BHC, as well as investigating metabolite production similarity by SOM and tSNE, all provide a comprehensive and capable strategy to investigate and describe the metabolomic composition of *C. asiatica* cultivars in time. We have found that the cycle of specific phytochemical production varies throughout the year which has corroborated previous findings that show the relationship of light intensity to the abundance of asiaticoside, madecassoside, flavonoids, and chlorogenic acid (Maulidiani et al., 2012). It remains unclear whether sunny periods are directly associated with varying concentrations from this study, since the plants are grown in greenhouses, yet many practical and specific insights were found for greenhouse cultivation. Thorough data has been collected and presented that link known metabolites, somewhat known metabolites, and tentatively known metabolites in the time-course production of phytochemicals from *C. asiatica*. These methods could also be applied to compare the effects of other cultivation variables such as temperature, light cycles, and soil additives.

4. Conclusions

This study provides a comprehensive investigation from cultivation-to-analysis for obtaining optimal preparation of plant materials for future clinical trials. There are 50,000–80,000 flowering plants reported as being used for medicinal purposes according to the International Union for Conservation of Nature and the World Wildlife Fund. The experimental, analytical, and computational approaches described here are applicable to other environmentally controlled plant cultivation studies and may contribute to global and specific analysis of phytochemical production monitoring.

CRediT authorship contribution statement

Alam, Md Nure: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation. **Bassett, Samuel:** Visualization, Investigation, Formal analysis. **Brown, Kevin:** Writing – review & editing, Writing – original draft, Visualization, Software, Data curation, Conceptualization. **Cabey, Kadine:** Methodology, Investigation, Formal analysis. **Cerruti, Natasha:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Choi, Jaewoo:** Writing – review & editing, Validation, Methodology, Formal analysis, Data curation. **Maier, Claudia S.:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Marney, Luke:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. Methodology, Investigation, Formal analysis. **Smith, James:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Data curation, Conceptualization. **Soumyanath, Amala:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Stevens, Jan F.:** Resources, Funding acquisition. **Techen, Natascha:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Viswanathan, Ramya:** Methodology, Investigation, Formal analysis. **Yang, Liping:** Writing – review & editing,

Validation, Methodology, Formal analysis, Data curation.

Supporting information

Supporting information of standard extracted ion chromatogram (XIC), MS/MS spectra for twelve markers, LC-MRM quantitation, calibration curves, as well as additional experimental details, materials and methods are provided.

Funding sources

This research was funded by NIH grants U19AT010829, S10OD026922, and S10RR027878.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.indcrop.2025.121159.

Data availability

Data will be made available on request.

References

Aharoni, A., Jongsma, M.A., Bouwmeester, H.J., 2005. Volatile science? Metabolic engineering of terpenoids in plants. Trends Plant Sci. 10 (12), 594–602. https://doi.org/10.1016/j.tplants.2005.10.005.

Alqahtani, A., Tongkao-On, W., Li, K.M., Razmovski-Naumovski, V., Chan, K., Li, G.Q., 2015. Seasonal variation of triterpenes and phenolic compounds in Australian Centella asiatica (L.) Urb. Phytochem. Anal. 26 (6), 436–443. https://doi.org/10.1002/pca.2578.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Bittremieux, W., Schmid, R., Huber, F., Van Der Hooft, J.J.J., Wang, M., Dorrestein, P.C., 2022. Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. J. Am. Soc. Mass Spectrom. 33 (9), 1733–1744. https://doi.org/10.1021/jasms.2c00153.

Brandolini, V., Coïsson, J.D., Tedeschi, P., Barile, D., Cereti, E., Maietti, A., Vecchiati, G., Martelli, A., Arlorio, M., 2006. Chemometrical characterization of four Italian rice varieties based on genetic and chemical analyses. J. Agric. Food Chem. 54 (26), 9985–9991. https://doi.org/10.1021/jf061799m.

Brown, J.S., 1991. Principal component and cluster analyses of cotton cultivar variability across the U.S. cotton belt. Crop Sci. 31 (4), 915–922. https://doi.org/10.2135/cropsci1991.0011183x003100040015x.

Bylka, W., Znajdek-Awizeń, P., Studzińska-Sroka, E., Brzezińska, M., 2013. Centella asiatica in cosmetology. Post. Dermatol. i Alergol. 30 (1), 46–49. https://doi.org/10.5114/pdia.2013.33378.

Chan, E.W.C., Lim, Y.Y., Ling, S.K., Tan, S.P., Lim, K.K., Khoo, M.G.H., 2009. Caffeoylquinic acids from leaves of Etlingera species (Zingiberaceae). Lwt 42 (5), 1026–1030. https://doi.org/10.1016/j.lwt.2009.01.003.

van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K., van der Werf, M.J., 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genom. 7, 1–15. https://doi.org/10.1186/1471-2164-7-142.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Gajbhiye, N.A., Makasana, J., Saha, A., Patel, I., Jat, R.S., 2016. LC-ESI-MS/MS method for simultaneous determination of triterpenoid glycosides and aglycones in centella asiatica L. Chromatographia 79 (11-12), 727–739. https://doi.org/10.1007/s10337-016-3089-x.

Gershenzon, J. and W. Kreis (2018). Biochemistry of Terpenoids: Monoterpenes, Sesquiterpenes, Diterpenes, Sterols, Cardiac Glycosides and Steroid Saponins.

Gray, N.E., Zweig, J.A., Matthews, D.G., Caruso, M., Quinn, J.F., Soumyanath, A., 2017. Centella asiatica attenuates mitochondrial dysfunction and oxidative stress in Aβ-exposed hippocampal neurons. Oxid. Med. Cell. Longev. 2017. https://doi.org/10.1155/2017/7023091.

Gray, N.E., Alcazar Magana, A., Lak, P., Wright, K.M., Quinn, J., Stevens, J.F., Maier, C. S., Soumyanath, A., 2018. Centella asiatica: phytochemistry and mechanisms of neuroprotection and cognitive enhancement. Phytochem. Rev. 17 (1), 161–194. https://doi.org/10.1007/s11101-017-9528-y.

Gray, N.E., Hack, W., Brandes, M.S., Zweig, J.A., Yang, L., Marney, L., Choi, J., Magana, A.A., Cerruti, N., McFerrin, J., Koike, S., Nguyen, T., Raber, J., Quinn, J.F., Maier, C.S., Soumyanath, A., 2024. Amelioration of age-related cognitive decline and anxiety in mice by Centella asiatica extract varies by sex, dose and mode of administration. Front. Aging 5. https://doi.org/10.3389/fragi.2024.1357922.

Hanson, A.D., Henry, C.S., Fiehn, O., De Crécy-Lagard, V., 2016. Metabolite damage and metabolite damage control in plants. Annu. Rev. Plant Biol. 67, 131–152. https://doi.org/10.1146/annurev-arplant-043015-111648.

Härdle, W., L. Simar and M. Fengler (2024). Applied multivariate statistical analysis, Springer Cham.

Hyvärinen, A., J. Karhunen and E. Oja (2001). Independent Component Analysis, John Wiley & Sons, Inc.

Intararuchikul, T., Teerapattarakan, N., Rodsiri, R., Tantisira, M., Wohlgemuth, G., Fiehn, O., Tansawat, R., 2019. Effects of Centella asiatica extract on antioxidant status and liver metabolome of rotenone-treated rats using GC–MS. Biomed. Chromatogr. 33 (2), 1–9. https://doi.org/10.1002/bmc.4395.

James, J.T., Dubery, I.A., 2009. Pentacyclic triterpenoids from the medicinal herb, Centella asiatica (L.) Urban. Molecules 14 (10), 3922–3941. https://doi.org/10.3390/molecules14103922.

Kalman, D., 2002. A singularly valuable decomposition: the SVD of a matrix. Coll. Math. J. 27 (1), 2–23. https://doi.org/10.1080/07468342.1996.11973744.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69. Doi 10.1007/Bf00337288.

Long, H.S., Stander, M.A., Van Wyk, B.E., 2012. Notes on the occurrence and significance of triterpenoids (asiaticoside and related compounds) and caffeoylquinic acids in Centella species. South Afr. J. Bot. 82, 53–59. https://doi.org/10.1016/j.sajb.2012.07.017.

Magana, A.A., Wright, K., Vaswani, A., Caruso, M., Reed, R.L., Bailey, C.F., Nguyen, T., Gray, N.E., Soumyanath, A., Quinn, J., Stevens, J.F., Maier, C.S., 2020. Integration of mass spectral fingerprinting analysis with precursor ion (MS1) quantification for the characterisation of botanical extracts: application to extracts of Centella asiatica (L.) Urban. Phytochem. Anal. 31 (6), 722–738. https://doi.org/10.1002/pca.2936.

Magaña, A.A., Kamimura, N., Soumyanath, A., Stevens, J.F., Maier, C.S., 2021. Caffeoylquinic acids: chemistry, biosynthesis, occurrence, analytical challenges, and bioactivity. Plant J. 107, 1299–1319. https://doi.org/10.1111/tpj.15390.

Matthews, D.G., Caruso, M., Magana, A.A., Wright, K.M., Maier, C.S., Stevens, J.F., Gray, N.E., Quinn, J.F., Soumyanath, A., 2020. Caffeoylquinic acids in centella asiatica reverse cognitive deficits in male 5XFAD Alzheimer's disease model mice. Nutrients 12 (11). DOI: ARTN 3488 10.3390/nu12113488.

Maulidiani, H., Khatib, A., Shaari, K., Abas, F., Shitan, M., Kneer, R., Neto, V., Lajis, N.H., 2012. Discrimination of three pegaga (Centella) varieties and determination of growth-lighting effects on metabolites content based on the chemometry of 1H nuclear magnetic resonance spectroscopy. J. Agric. Food Chem. 60 (1), 410–417. https://doi.org/10.1021/jf200270y.

McInnes, L., Healy, J., Saul, N., Großberger, L., 2018. UMAP: uniform manifold approximation and projection for dimension reduction. J. Open Source Softw. 3 (29), 861. https://doi.org/10.21105/joss.00861.

Misra, B.B., 2020. Data normalization strategies in metabolomics: current challenges, approaches, and tools. Eur. J. Mass Spectrom. 26 (3), 165–174. https://doi.org/10.1177/1469066720918446.

Orhan, I.E., 2012. Centella asiatica (L.) Urban: From traditional medicine to modern medicine with neuroprotective potential. Evid. -Based Complement. Altern. Med. 2012. https://doi.org/10.1155/2012/946259.

Rahajanirina, V., Ony, S., Raoseta, R., Roger, E., Razafindrazaka, H., Pirotais, S., Boucher, M., Danthu, P., 2012. The influence of certain taxonomic and environmental parameters on biomass production and triterpenoid content in the leaves of Centella asiatica (L.) Urb. from Madagascar. Chem. Biodivers. (2), 298–308. https://doi.org/10.1002/cbdv.201100073.

Savage, R.S., Heller, K., Xu, Y., Ghahramani, Z., Truman, W.M., Grant, M., Denby, K.J., Wild, D.L., 2009. R/BHC: fast Bayesian hierarchical clustering for microarray data. Bmc Bioinforma. 10. DOI: Artn 242 10.1186/1471-2105-10-242.

Schrimpe-Rutledge, A.C., Codreanu, S.G., Sherrod, S.D., McLean, J.A., 2016. Untargeted metabolomics strategies—challenges and emerging directions. J. Am. Soc. Mass Spectrom. 27 (12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y.

Shen, X., Gong, X., Cai, Y., Guo, Y., Tu, J., Li, H., Zhang, T., Wang, J., Xue, F., Zhu, Z.J., 2016. Normalization and integration of large-scale metabolomics data using support vector regression. Metabolomics 12 (5), 1–12. https://doi.org/10.1007/s11306-016-1026-5. ARTN 89.

Shen, X.T., Gong, X.Y., Cai, Y.P., Guo, Y., Tu, J., Li, H., Zhang, T., Wang, J.L., Xue, F.Z., Zhu, Z.J., 2016. Normalization and integration of large-scale metabolomics data using support vector regression. Metabolomics 12 (5). DOI: ARTN 89 10.1007/s11306-016-1026-5.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Ideker, T., 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27 (3), 431–432. https://doi.org/10.1093/bioinformatics/btq675.

Soumyanath, A., Zhong, Y.-P., Yu, X., Bourdette, D., Koop, D.R., Gold, S.A., Gold, B.G., 2010. Centella asiatica accelerates nerve regeneration upon oral administration and contains multiple active fractions increasing neurite elongation in-vitro. J. Pharm. Pharmacol. 57 (9), 1221–1229. https://doi.org/10.1211/jpp.57.9.0018.

Stefansson, M., Sjöberg, P.J.R.Per J.R., Markides, K., 1996. Regulation of multimer formation in electrospray mass spectrometry. Anal. Chem. 68 (10), 1792–1797. https://doi.org/10.1021/ac950980j.

Wang, M.X., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A.

V., Meehan, M.J., Liu, W.T., Criisemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C. C., Floros, D.J., Gavilan, R.G., Kleigrewe, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.C., Yang, Y.L., Humpf, H.U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A. M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya, C.A., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J.Q., Neupane, R., Gurr, J., Rodríguez, A.M.C., Lamsa, A., Zhang, C., Dorrestein, K., Duggan, B.M., Almaliti, J., Allard, P.M., Phapale, P., Nothias, L.F., Alexandrovr, T., Litaudon, M., Wolfender, J.L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W.Y., Liu, X.T., Zhang, L.X., Knight, R., Jensen, P.R., Palsson, B.O., Pogliano, K., Linington, R.G., Gutierrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorrestein, P.C., Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol. 34 (8), 828–837. https://doi.org/10.1038/nbt.3597.

Wittek, P., Gao, S.C., Lim, I.S., Zhao, L., 2017. somoclu: an efficient parallel library for self-organizing maps. J. Stat. Softw. 78 (9), 1–21. https://doi.org/10.18637/jss.v078.i09.

Wright, K.M., McFerrin, J., Alcázar Magaña, A., Roberts, J., Caruso, M., Kretzschmar, D., Stevens, J.F., Maier, C.S., Quinn, J.F., Soumyanath, A., 2022. Developing a rational, optimized product of centella asiatica for examination in clinical trials: real world challenges. Front. Nutr. 8 (January), 1–18. https://doi.org/10.3389/fnut.2021.799137.

Wright, K.M., Bollen, M., David, J., Mepham, B., Alcázar Magaña, A., McClure, C., Maier, C.S., Quinn, J.F., Soumyanath, A., 2023. Bioanalytical method validation and application to a phase 1, double-blind, randomized pharmacokinetic trial of a standardized Centella asiatica (L.) Urban water extract product in healthy older adults. Front. Pharmacol. 14 (August), 1–21. https://doi.org/10.3389/fphar.2023.1228030.

Wu, Z.W., Li, W.B., Zhou, J., Liu, X., Wang, L., Chen, B., Wang, M.K., Ji, L., Hu, W.C., Li, F., 2020. Oleanane-and ursane-type triterpene saponins from centella asiatica exhibit neuroprotective effects. J. Agric. Food Chem. 68 (26), 6977–6986. https://doi.org/10.1021/acs.jafc.0c01476.

Xu, Y.F., Lu, W., Rabinowitz, J.D., 2015. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. Anal. Chem. 87 (4), 2273–2281. https://doi.org/10.1021/ac504118y.

Yang, L., Marney, L., Magana, A.A., Choi, J., Wright, K., McFerrin, J., Gray, N.E., Soumyanath, A., Stevens, J.F., Maier, C.S., 2023. Quantification of caffeoylquinic acids and triterpenes as targeted bioactive compounds of Centella asiatica in extracts and formulations by liquid chromatography mass spectrometry, 100091-100091 J. Chromatogr. Open 4 (June). https://doi.org/10.1016/j.jcoa.2023.100091.

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T.L., 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinforma. 13, 134. https://doi.org/10.1186/1471-2105-13-134.

Yuan, H., Ma, Q., Ye, L., Piao, G., 2016. The traditional medicine and modern medicine from natural products. Molecules 21 (5). https://doi.org/10.3390/molecules21050559.