Untargeted Metabolomics Sheds Light on the Diversity of Major Classes of Secondary Metabolites in the Malpighiaceae Botanical Family

Helena Mannochio-Russo[1,2]*, Rafael F. de Almeida[3,4], Wilhan D. G. Nunes[5], Paula C. P. Bueno[6,7], Andrés M. Caraballo-Rodríguez[2], Anelize Bauermeister[2], Pieter C. Dorrestein[2] and Vanderlan S. Bolzani[1]*

[1]NuBBE, Department of Biochemistry and Organic Chemistry, Institute of Chemistry, São Paulo State University (UNESP), Araraquara, Brazil, [2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, San Diego, CA, United States, [3]Royal Botanical Gardens Kew, Science, Ecosystem Stewardship, Diversity and Livelihoods, Richmond, United Kingdom, [4]Department of Biological Sciences, Lamol Lab, Feira de Santana State University (UEFS), Feira de Santana, Brazil, [5]Federal Institute of Education, Science and Technology of Rondônia (IFRO), Ji-Paraná, Brazil, [6]Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany, [7]Institute of Chemistry, Federal University of Alfenas (UNIFAL), Alfenas, Brazil

Natural products produced by plants are one of the most investigated natural sources, which substantially contributed to the development of the natural products field. Even though these compounds are widely explored, the literature still lacks comprehensive investigations aiming to explore the evolution of secondary metabolites produced by plants, especially if classical methodologies are employed. The development of sensitive hyphenated techniques and computational tools for data processing has enabled the study of large datasets, being valuable assets for chemosystematic studies. Here, we describe a strategy for chemotaxonomic investigations using the Malpighiaceae botanical family as a model. Our workflow was based on MS/MS untargeted metabolomics, spectral searches, and recently described *in silico* classification tools, which were mapped into the latest molecular phylogeny accepted for this family. The metabolomic analysis revealed that different ionization modes and extraction protocols significantly impacted the chemical profiles, influencing the chemotaxonomic results. Spectral searches within public databases revealed several clades or genera-specific molecular families, being potential chemical markers for these taxa, while the *in silico* classification tools were able to expand the Malpighiaceae chemical space. The classes putatively annotated were used for ancestral character reconstructions, which recovered several classes of metabolites as homoplasies (i.e., non-exclusive) or synapomorphies (i.e., exclusive) for all sampled clades and genera. Our workflow combines several approaches to perform a comprehensive evolutionary chemical study. We expect it to be used on further chemotaxonomic investigations to expand chemical knowledge and reveal biological insights for compounds classes in different biological groups.

Keywords: chemotaxonomy, mass spectrometry, metabolite annotation, metabolomics, evolution, ancestral character reconstruction, systematics, malpighiales

INTRODUCTION

Plant metabolites have been widely explored since the 1930s, mainly aiming at an in-depth study of species with ethnopharmacological value, bioactive extracts and compounds, and new chemical structures (Newman and Cragg, 2020; Atanasov et al., 2021). Despite its invaluable importance, the establishment of phylogenetic diversification and distribution patterns of plant secondary metabolites is still in its early steps, and several plant families have not been deeply explored to date in this context.

The development of highly sensitive detection techniques, such as mass spectrometry (MS), allowed the investigation of plant extracts as a whole, providing a comprehensive overview of the metabolites biosynthesized by plant species (De Vos et al., 2007; Gemperline et al., 2016). In this context, chemosystematics studies employing MS, multivariate analyses, and bioinformatic tools for exploring large plant datasets have been successfully performed, raising valuable insights regarding the biosynthetic pathways involved in different phylogenetic groups (Gallon et al., 2018; Martucci et al., 2018; Ernst et al., 2019; Kang et al., 2019). It is now possible to support DNA-based phylogenetic studies at a molecular level and assist in elaborating evolutionary hypotheses based on natural products and metabolomics analyses (Schmitt and Barker, 2009). The rapid development of natural product bioinformatics tools and databases (Li and Gaquerel, 2021; Medema, 2021; Bauermeister et al., 2022) can be of great value to assist and accelerate more comprehensive chemosystematics studies aiming at taxa-specific metabolic pathways.

Even though evolutionary studies at a metabolite level are of great interest, certain factors can significantly impact the detection of metabolites and, consequently, the conclusions drawn. For instance, the extraction solvent used and the ionization mode in MS can prioritize one class of compounds over another (Floros et al., 2016; Creydt and Fischer, 2017), leading to biased results. Even though several studies compare different extraction protocols and ionization modes in metabolomics analyses, an investigation of these factors in chemosystematics studies is still poorly explored. Therefore, selecting a plant family with a broad diversity of classes of compunds identified, in addition to an extensive record of DNA-based phylogenetic studies, would be ideal for evaluating these variables, enabling a deeper chemotaxonomic investigation.

The Malpighiaceae plant family is an excellent example of both the high diversity of secondary metabolites produced by plant species, and a flowering plant family with all its genera sampled in DNA-based phylogenetic studies (Davis and Anderson, 2010; Mannochio-Russo et al., 2020). Malpighiaceae is one of the 36 families of flowering plants placed in Malpighiales by several phylogenetic studies based on chloroplast genes (Angiosperm Phylogeny Group, 1998, 2003, 2009; Angiosperm Phylogeny Group et al., 2016), being also one of the most important and diverse orders of angiosperms in tropical forests (Xi et al., 2012; Cai et al., 2019). This family currently comprises 74 genera and *ca.* 1,300 species, mostly confined to the American tropics, with Brazil being

its most representative country. Few genera and species reach the tropics of Africa, Asia, and Oceania (Davis and Anderson, 2010; de Almeida et al., 2021; de Almeida and van den Berg, 2021). Some Amazonian species of Malpighiaceae are traditionally known for their psychedelic or aphrodisiac properties, with several studies focusing on the chemical characterization of these species (Samoylenko et al., 2010; Queiroz et al., 2014). On the other hand, several extra-Amazonian Malpighiaceae species are long reported as toxic to cattle (i.e., *Amorimia*, and *Niedenzuella* spp.; Riet-Correa et al., 2012; Lee et al., 2014), causing significant economic losses in the Brazilian growing cattle industry. Additionally, *Malpighia* and *Byrsonima* spp. also stand out for the nutritional value of their fruits (Belwal et al., 2018; Neri-Numa et al., 2018).

In the past two decades, traditional intrafamilial classifications of several angiosperm families (i.e., subfamily and tribe ranks), solely based on macromorphology, were proven to be non-monophyletic (i.e., artificial, and not reflecting common ancestry; Angiosperm Phylogeny Group, 1998, 2003, 2009; Angiosperm Phylogeny Group et al., 2016). During this time, Malpighiaceae has gone through unprecedented changes in its traditional classification due to several DNA-based phylogenetic studies (Cameron et al., 2001; Davis et al., 2001; Davis and Anderson, 2010; de Almeida et al., 2017, 2018). Key morphological characters of its traditional classification system (i.e., fruit types) were recovered as highly homoplastic (i.e., non-exclusive; Cameron et al., 2001; Davis et al., 2001). The inevitable recognition of unforeseen relationships within Malpighiaceae brought to light deep taxonomic problems regarding the monophyly of several genera [e.g., *Banisteriopsis* C.R.Rob., *Mascagnia* (Bertero ex DC.) Bertero, *Stigmaphyllon* A.Juss., and *Tetrapterys* Cav.], tribes (e.g., just Gaudichaudieae Horan. was recovered as monophyletic), and all its subfamilies (e.g., Byrsonimoideae W.R.Anderson and Malpighioideae A.Juss.; Cameron et al., 2001; Davis et al., 2001; Davis and Anderson, 2010; de Almeida et al., 2017). Since then, different authors have gradually proposed new genera and combinations to accommodate these newly identified lineages (Anderson, 2006, 2011; Anderson and Davis, 2006; Davis et al., 2020; de Almeida and van den Berg, 2021). Even though some morphological characters were used to reconstruct the last generic phylogeny of Malpighiaceae, no morphological characters were ever recovered, circumscribed, or discussed for its major phylogenetic clades (de Almeida and van den Berg, 2021). As a result, its traditional classification was rejected, and 10 informal clades, without any morphological circumscription, were recognized in the most recent generic phylogeny for Malpighiaceae (Davis and Anderson, 2010). More recently, phylogenomic studies were performed with six Malpighiaceae species and strongly corroborated previous phylogenetic studies within this family (Menezes et al., 2018; Ramachandran et al., 2018; Cai et al., 2019; Jo et al., 2019; Gong et al., 2020). In this context, a deeper chemical investigation of the family and mapping its chemical traits in a phylogenetic tree that reflects the evolutionary relationships among organisms can be of great value in chemotaxonomic investigations (Schmitt and Barker, 2009). It can help predict metabolically interesting groups of organisms

to assist future studies of this taxon and give more support to the evolutionary hypotheses.

In this study, we present a new approach for chemosystematics studies, by combining natural products research with phylogenetic methods (Schmitt and Barker, 2009). We performed metabolomics analyses in combination with recently described *in silico* fragmentation predictors, chemical hierarchy analysis, and ancestral character reconstructions to map the presence/absence of the annotated metabolites in the most recent generic DNA-based phylogenetic tree of Malpighiaceae (**Figure 1**). For this, we evaluated a unique collection of Malpighiaceae samples, comprising 39 genera (out of 74) and 137 species from each of the major phylogenetic groups currently accepted for the family (collection distribution is shown in **Figure 2**). This study comprised representative samples from all the currently accepted phylogenetic clades of the family, which enabled us to obtain a comprehensive overview of the metabolites produced by this family (percentage of genera covered by each clade: 67, 50, 67, 100, 33, 60, 20, 59, 31, and 63% for clades A–J, respectively[1]). We evaluated (i) the impact of different extraction protocols and ionization modes in MS for chemotaxonomic investigations; (ii) the metabolites annotated based on spectral matches and *in silico* tools; and (iii) how the chemical diversity in Malpighiaceae evolved over the geological time in this family. With this enhanced approach, we were able to provide insights regarding the

---

[1]http://powo.science.kew.org/

complementarity of the different ionization modes, provide the first chemical information of several Malpighiaceae genera, and draw conclusions regarding the evolution of the classes of secondary metabolites annotated in the Malpighiaceae plant family.
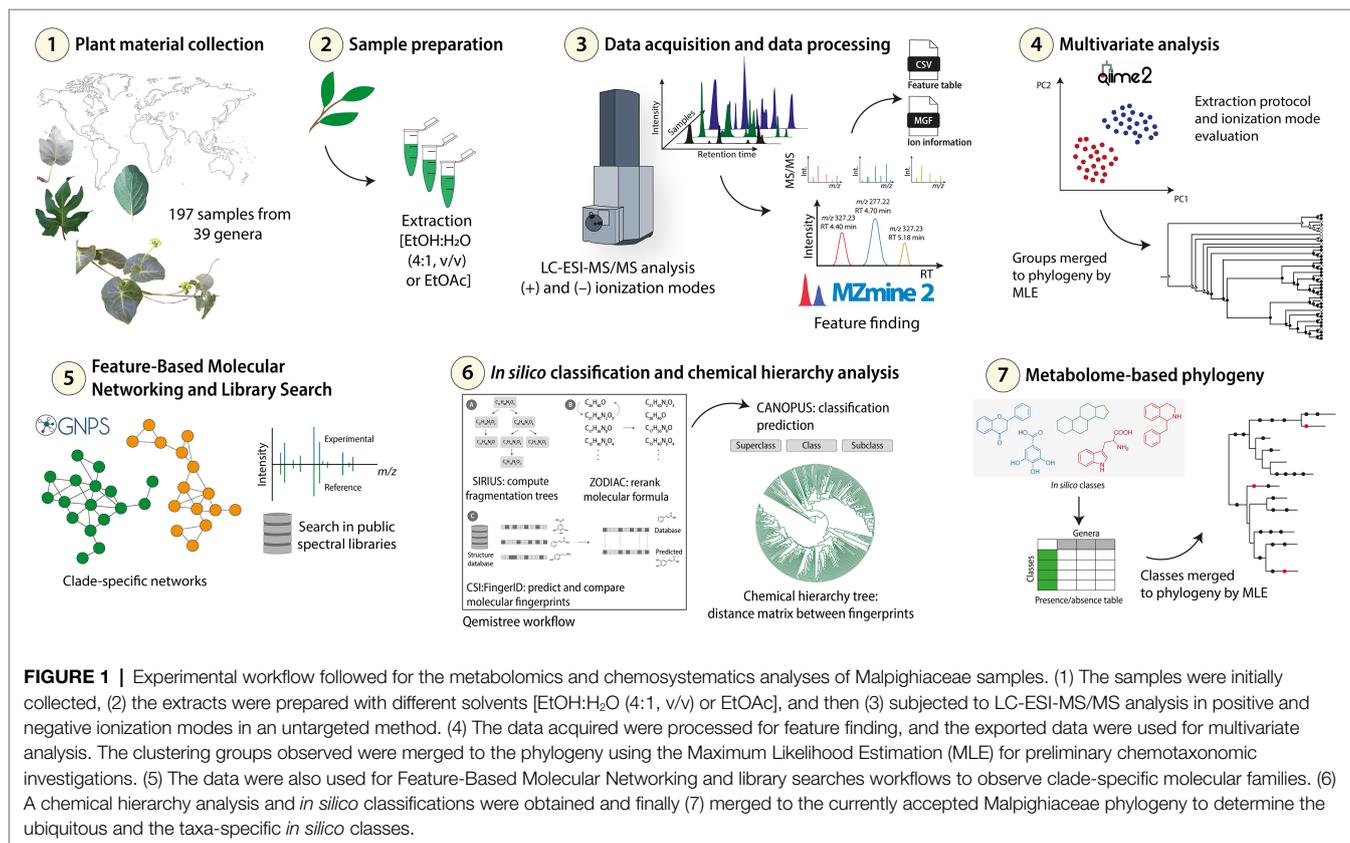
MATERIALS AND METHODS
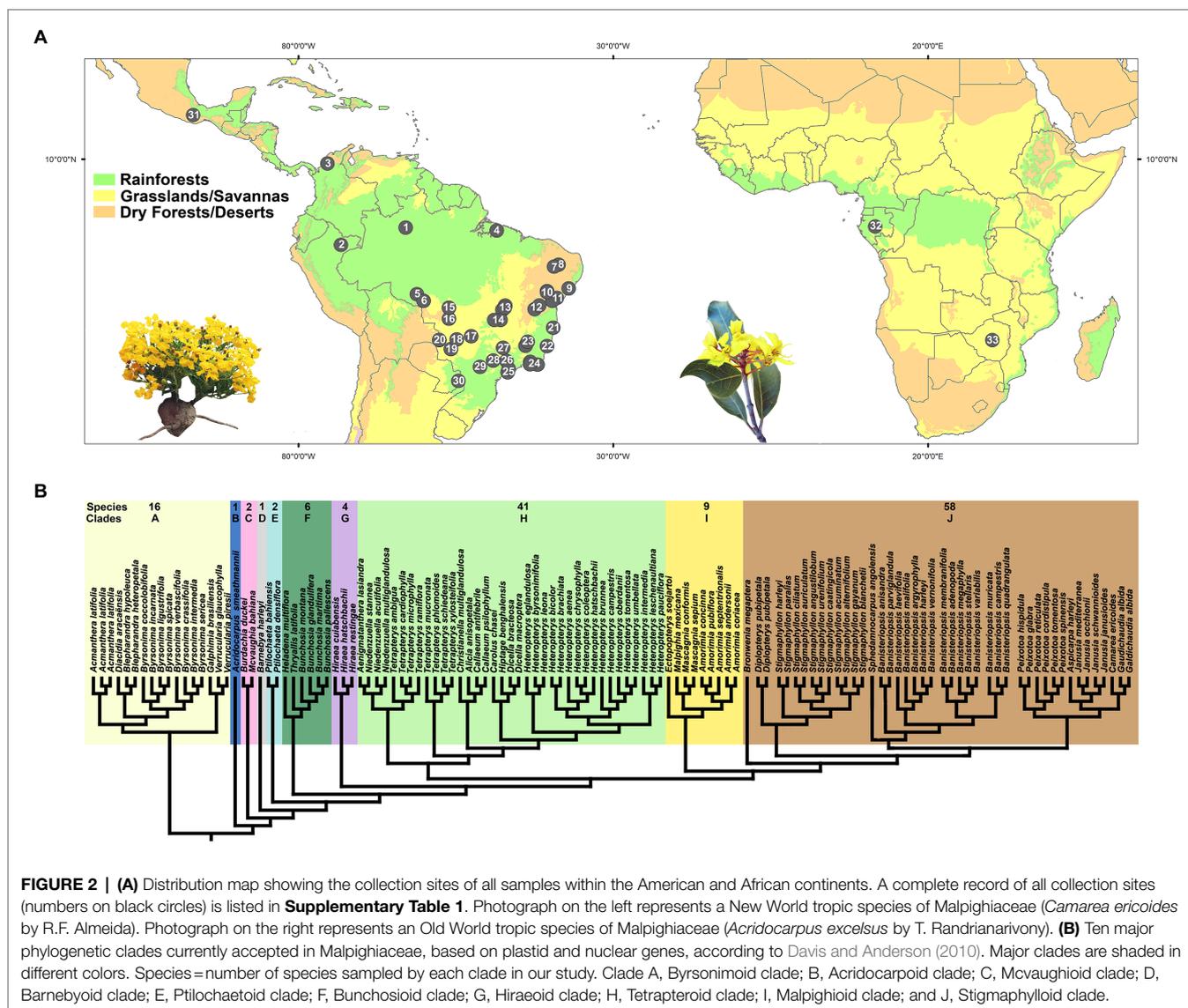
General Information

The ethyl alcohol (proof, for molecular biology) used for the extraction procedure was acquired from Sigma-Aldrich (St. Louis, United States). The ethyl acetate (HPLC grade) used for the extraction procedure was obtained by J.T. Baker (J.T. Baker-Avantor, Radnor, United States). Acetonitrile and water, both LC–MS grade, were obtained from Fisher Scientific (Fair Lawn, NJ, United States).

Collection of Plant Material

Most plant samples were collected by R.F. Almeida on field expeditions throughout Brazil from 2013 to 2017, or were retrieved from discarded fragmented samples used for DNA extraction in molecular studies (de Almeida et al., 2017, 2018; de Almeida and van den Berg, 2020, 2021). For information regarding all sampled specimens, see **Supplementary Table 1**. After each collection, the plant materials were dried in a desiccator containing silica at room temperature. Samples were then frozen in liquid nitrogen and grounded in a ball mill.

FIGURE 1 | Experimental workflow followed for the metabolomics and chemosystematics analyses of Malpighiaceae samples. (1) The samples were initially collected, (2) the extracts were prepared with different solvents [EtOH:H$_2$O (4:1, v/v) or EtOAc], and then (3) subjected to LC-ESI-MS/MS analysis in positive and negative ionization modes in an untargeted method. (4) The data acquired were processed for feature finding, and the exported data were used for multivariate analysis. The clustering groups observed were merged to the phylogeny using the Maximum Likelihood Estimation (MLE) for preliminary chemotaxonomic investigations. (5) The data were also used for Feature-Based Molecular Networking and library searches workflows to observe clade-specific molecular families. (6) A chemical hierarchy analysis and *in silico* classifications were obtained and finally (7) merged to the currently accepted Malpighiaceae phylogeny to determine the ubiquitous and the taxa-specific *in silico* classes.

FIGURE 2 | (A) Distribution map showing the collection sites of all samples within the American and African continents. A complete record of all collection sites (numbers on black circles) is listed in **Supplementary Table 1**. Photograph on the left represents a New World tropic species of Malpighiaceae (*Camarea ericoides* by R.F. Almeida). Photograph on the right represents an Old World tropic species of Malpighiaceae (*Acridocarpus excelsus* by T. Randrianarivony). **(B)** Ten major phylogenetic clades currently accepted in Malpighiaceae, based on plastid and nuclear genes, according to Davis and Anderson (2010). Major clades are shaded in different colors. Species = number of species sampled by each clade in our study. Clade A, Byrsonimoid clade; B, Acridocarpoid clade; C, Mcvaughioid clade; D, Barnebyoid clade; E, Ptilochaetoid clade; F, Bunchosioid clade; G, Hiraeoid clade; H, Tetrapteroid clade; I, Malpighioid clade; and J, Stigmaphylloid clade.

The samples were stored in a freezer at −20°C until the preparation of the extracts.

The authorization for conducting this study was conceded by the National System for Management of Genetic Heritage and Associated Traditional Knowledge (SISGEN), n° A6FDC2E.

Extraction Procedure

The plant sample materials were weighed and extracted with EtOH:H$_2$O (4:1, v/v) or EtOAc (100%) in a proportion of 20 mg of plant material to 1 mL of extraction solvent. The samples were homogenized in a Qiagen TissueLyzer II (Qiagen, Hilden, Germany) for 5 min at 25 MHz and extracted for additional 30 min at room temperature. The samples were centrifuged (5,000 g) for 15 min, and 300 μL of the supernatants were transferred to a 96-deep-well plate. The solvent was dried in a Labconco CentriVap (United States),

and the plates were sealed and stored at −80°C prior to analyses.

UHPLC-MS/MS Analysis

The extracts were initially resuspended in 200 μL MeOH:H$_2$O (4:1) containing sulfachloropyridazine (2 μM) as internal standard [to monitor sample injection during the Ultra High Performance Liquid Chromatography (UHPLC)-tandem Mass Spectrometry (MS/MS) data acquisition], and sonicated for 15 min. The plates were centrifuged for 10 min at 1,300 g, and the supernatants were then transferred to a new 96-well plate for metabolomics analyses.

The analyses were carried out with a Thermo Scientific UltiMate 3000 UHPLC system coupled to a Maxis Impact QTOF mass spectrometer (Bruker Daltonics, Germany), controlled by the Otof Control and Hystar software packages, and equipped with ESI source. The extracts were analyzed using a Kinetex 1.7 μm C18 reversed-phase UHPLC column (50 × 2.1 mm; Phenomenex,

Torrance, CA, United States), at 40°C, and an injection volume of 5 μL. The pump system consisted of water (A) and acetonitrile (B), both acidified with formic acid (0.1%, v/v), and the flow rate was set at 0.5 ml/min. The metabolites separation was achieved with 5% solvent B for 1 min, followed by a linear gradient from 5 to 100% in 5 min. The column was washed at 100% solvent B for 2 min, then returned to the initial 5% in 1 min, and the equilibration of the column was achieved for 1 min at 5% solvent B. The mass spectra were acquired in both positive and negative ionization modes, separately, in a mass range of 50–2,000 Da in data-dependent acquisition (DDA) mode. The parameters used for data acquisition were set as follows: nitrogen used as nebulizer gas with pressure at 2 bar, a capillary voltage of 4,200 V, ion source temperature of 200°C, dry gas flow at 9 L/min, and spectra rate acquisition of three spectra/s. The five most intense selected ions per spectrum were fragmented (MS/MS) using ramped collision-induced dissociation energy, ranging from 22 to 50 eV. MS/MS active exclusion was set after five spectra and released after 30 s.

The UHPLC–MS/MS data were deposited in the MassIVE Public GNPS dataset[2] (MSV000085119) and are publicly available.

MS/MS Data Pre-Processing

The raw data files (.d) were converted to .mzXML format using DataAnalysis software (Bruker) after lock mass correction using hexakis(1H,1H,2H-difluoroethoxy) phosphazene (Synquest Laboratories, Alachua, FL, United States), with m/z 622.029509. The quality of the analyses was evaluated considering the retention time and the m/z of a standard solution containing a mixture of six standards, which was analyzed after the completion of each row in a 96-well plate.

The .mzXML files were processed in MZmine2 (version 2.37.corr17.7_kai_merge2) for positive and negative ionization modes, separately. The parameters used for feature finding were as follows: mass detection (centroid, 1.0E3 and 1.0E1 for MS1 and MS2, respectively); chromatogram builder (minimum time span of 0.01 min, minimum height of 3.0E3, and m/z tolerance of 20 ppm); chromatogram deconvolution (baseline cut-off algorithm: minimum peak height: 1.0E3, peak duration range: 0.01–3 min; and baseline level: 1.0E3) with median m/z center calculation, m/z range for MS2 scan pairing of 0.02 Da and retention time (RT) range for MS2 scan pairing of 0.1 min; isotope peaks grouper (m/z tolerance set at 20 ppm, RT tolerance of 0.1 min, maximum charge of 3, and representative isotope set to most intense), join alignment (m/z tolerance of 20 ppm, weight for m/z and RT of 75 and 25, respectively, and RT tolerance of 0.1 min). A filter was applied in order to keep only the features with MS/MS spectra. This feature list was exported as a feature quantification table (.csv), as a MS2 spectral summary (.mgf), and with the SIRIUS export module (.mgf) for downstream analyses.

Feature-Based Molecular Networking

To investigate the metabolic profile of the dataset, the processed LC–MS/MS data (.mgf and .csv) were used to create a Feature-Based Molecular Network (FBMN) (Nothias et al., 2020) on the GNPS platform (Wang et al., 2016) with input files containing only the features detected in the hydroethanolic extracts. The data were filtered by removing all MS/MS fragment ions within +/− 17 Da of the precursor ion in order to remove possible residual precursor ions, which can sometimes be observed in MS/MS spectra acquired in QToF equipment. Additionally, MS/MS spectra were window filtered to select only the top six fragment ions in the +/− 50 Da window throughout the spectrum. Both the precursor ion and the MS/MS fragment ion tolerance were set to 0.02 Da. A molecular network was created, in which the edges were filtered to have a cosine score above 0.7 and at least four matched peaks. Similarly, the parameters for the library search (for comparison between the experimental and library spectra) were set to have a score above 0.7 and at least four matched peaks to assist in the metabolites annotation—level three according to the metabolomics standards initiative (MSI; Sumner et al., 2007). The FBMN jobs on GNPS can be found at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2c5f11403ac847a298e4d7866a491143 (positive mode) and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=501c16500476451f978311057266fbdf (negative mode).

The molecular network visualization was performed in Cytoscape (version 3.7.2, Cytoscape Consortium, San Diego, CA, United States; Shannon et al., 2003), in which the nodes correspond to ion features, while the edges between the nodes represent the MS/MS cosine scores calculated. Subnetworks in which the nodes were found in significant abundances in blanks were excluded from the Cytoscape visualization to avoid misinterpretations due to contaminants in the analyses. Sample type information was added to color the nodes as pie charts representing the relative abundance of the features across the samples (colors based on the phylogenetic clades A–J). Node size was scaled relative to the sum of the peak areas obtained in the samples in which the feature was detected. Compounds with the same MS/MS spectra, but with different retention times, were represented as separate nodes, indicating isomers.

Chemical Hierarchy Analysis

A chemical hierarchy analysis (Qemistree; Tripathi et al., 2021) was performed with the metabolites detected in the hydroethanolic extraction protocol. For this, we used the q2-qemistree qiime2 plugin,[3] in which the feature quantification table (.csv) and the file obtained from the SIRIUS (Dührkop et al., 2019) export module (.mgf) from MZmine were used as input. Briefly, the Qemistree workflow consists of applying SIRIUS (version 4.8.2) to the .mgf file (containing ion information), generating predicted molecular formulas for each feature. The predicted molecular formulas were reranked using ZODIAC (Ludwig et al., 2020), and the predicted molecular fingerprints were subsequently generated using fragmentation trees *via* CSI:FingerID (Dührkop et al., 2015). The chemical taxonomy of the predicted metabolite structures was obtained

---

[2] http://massive.ucsd.edu

[3] https://github.com/biocore/q2-qemistree

by CANOPUS (superclass, class, and subclass; Dührkop et al., 2021). The Euclidean pairwise distances between the molecular fingerprints were calculated, and the fingerprint vectors were hierarchically clustered to generate a tree representing the structural chemical relationships of this dataset. The tree was then pruned in order to keep only the fingerprints classified up to a superclass level. The trees were visualized interactively in EMPRESS (Cantrell et al., 2021) for data exploration, in which clade information was added as relative abundance stacked barcharts to each feature. The dendrogram obtained can be interactively visualized with the .qzv files found in the Github repository.[4]

Statistical Analysis

The feature table exported from MZmine was used to perform unsupervised analysis using Qiime 2 (version 2020.2; Bolyen et al., 2019) bioinformatics pipeline within a Jupyter notebook. The metabolomic profiles were compared using the Bray–Curtis distance metric for comparing different extraction protocols, and using Canberra metric to investigate the four subsets individually (two extraction solvents, and two ionization modes). The Principal Coordinates Analysis (PCoA) plots showing the top three principal coordinates were visualized using EMPeror (Vázquez-Baeza et al., 2013). Permutational multivariate ANOVA (PERMANOVA; Anderson, 2001) was also performed in Qiime 2 on metabolite distance matrices to test for clustering significance (with 999 permutations), and the F statistic was reported as a measure of effect size.

The *in silico* classes retrieved from the Qemistree workflow (Tripathi et al., 2021) and CANOPUS (Dührkop et al., 2021) were tested for differential enrichment in the most sampled clades in study (clades A, F–J). Initially, the relative abundances were summed for the class group, the data were normalized by arc-sine square root transformation, and the effect of the clades was tested using a simple ANOVA. The adjusted values of p from ANOVA were obtained from the Benjamini-Hochberg method. The pairwise differences between the clades were tested through a *post-hoc* Tukey test, and the magnitude of differential enrichment was calculated through the $\log_2$ fold change in mean relative abundance between clades. The subclasses statistically enriched for specific clades were then selected to build a heatmap.

Phylogenetic Analyses

In total, 39 genera (out of 74) and 139 species (out of 1,300) of Malpighiaceae were sampled, representing all of its 10 major clades recognized by recent molecular phylogenetic studies (**Supplementary Table 1**). Since our sampling focused on the diversity of genera and clades within the Malpighiaceae family, genera not sampled on the chemical analyses were not included in the molecular phylogenetic analysis. Sequences for the genes *matK*, *ndhF*, PHYC, and *rbcL* were retrieved from GenBank,[5] edited using Geneious (Kearse et al., 2012), and aligned using

---

[4]https://github.com/helenamrusso/Malpighiaceae_supplementary
[5]https://www.ncbi.nlm.nih.gov/genbank/

Muscle (Edgar, 2004), with subsequent adjustments in the preliminary matrices to ensure that the nitrogenous bases were correctly aligned. The complete data matrices are available at TreeBase (https://www.treebase.org/treebase-web/search/studySearch.html, accession number S11008).

All trees were rooted in the Byrsonimoid clade (clade A), which is considered the sister-group of the other Malpighiaceae clades according to Davis and Anderson (2010). Combined analysis of plastid + nuclear regions was performed using Bayesian inference and Maximum Likelihood criteria to reconstruct our phylogenetic hypotheses. Both model-based methods were conducted with a mixed substitution model (GTR + G + I) and unlinked parameters, using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003) and raxmlGUI2 (Edler et al., 2021). For the Bayesian inference, the Markov Chain Monte Carlo (MCMC) was run using two simultaneous independent runs with four chains each (one cold and three heated), saving one tree every 1,000 generations, for a total of 10 million generations. We excluded 20% of retained trees as "burn in", and checked for a stationary phase of likelihood, checking for ESS values higher than 200 for all parameters on Tracer 1.6 (Rambaut et al., 2014). The posterior probabilities (PP) of clades were based on the majority rule consensus, using the stored trees, and calculated with MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003).

Ancestral State Reconstruction

Chemical profiles were selected based on the clustering trends of the PCoA analysis for EtOH80 and EtOAc100% extracts, in both positive and negative ionization modes (four subsets). The subsets in which samples clustered into two groups (A and B) represented different chemical profiles. Character coding followed the recommendations of (Sereno, 2007) for morphological phylogenies. Primary homology hypotheses (de Pinna, 1991) were proposed for different chemical profiles, with a total of two subgroups of classes scored (i.e., presence and absence). All characters were optimized on the concatenated tree from the Bayesian inference using the maximum likelihood function with Mesquite 2.73 (Maddison and Maddison, 2006) using the mk1 algorithm.

After annotating the main classes of compounds present in the hydroethanolic extracts, we transformed the obtained quantitative matrices for the positive and negative ionization modes into qualitative matrices considering only the presence and absence of all annotated classes. Since we do not have all the analyzed species of this study sampled in the latest generic molecular phylogeny of Malpighiaceae, we summarized our results at the generic level using an arithmetic mean equation. After summarizing both matrices (retrieved from the positive and negative ionization modes), we compared all classes of secondary metabolites identified in each of them, merging their binary coding into a single column. By doing so, we obtained a total of 78 classes of secondary metabolites that were optimized into a summarized version of the phylogenetic tree presented by us in the previous step, including only the genera sampled in this study. Ancestral character reconstructions were performed using Maximum

Likelihood approaches into the software Mesquite 2.73 (Maddison and Maddison, 2006) and visualized in the software Winclada (Nixon, 1999) using the fast optimization method, which favors homoplasies in the analyses.
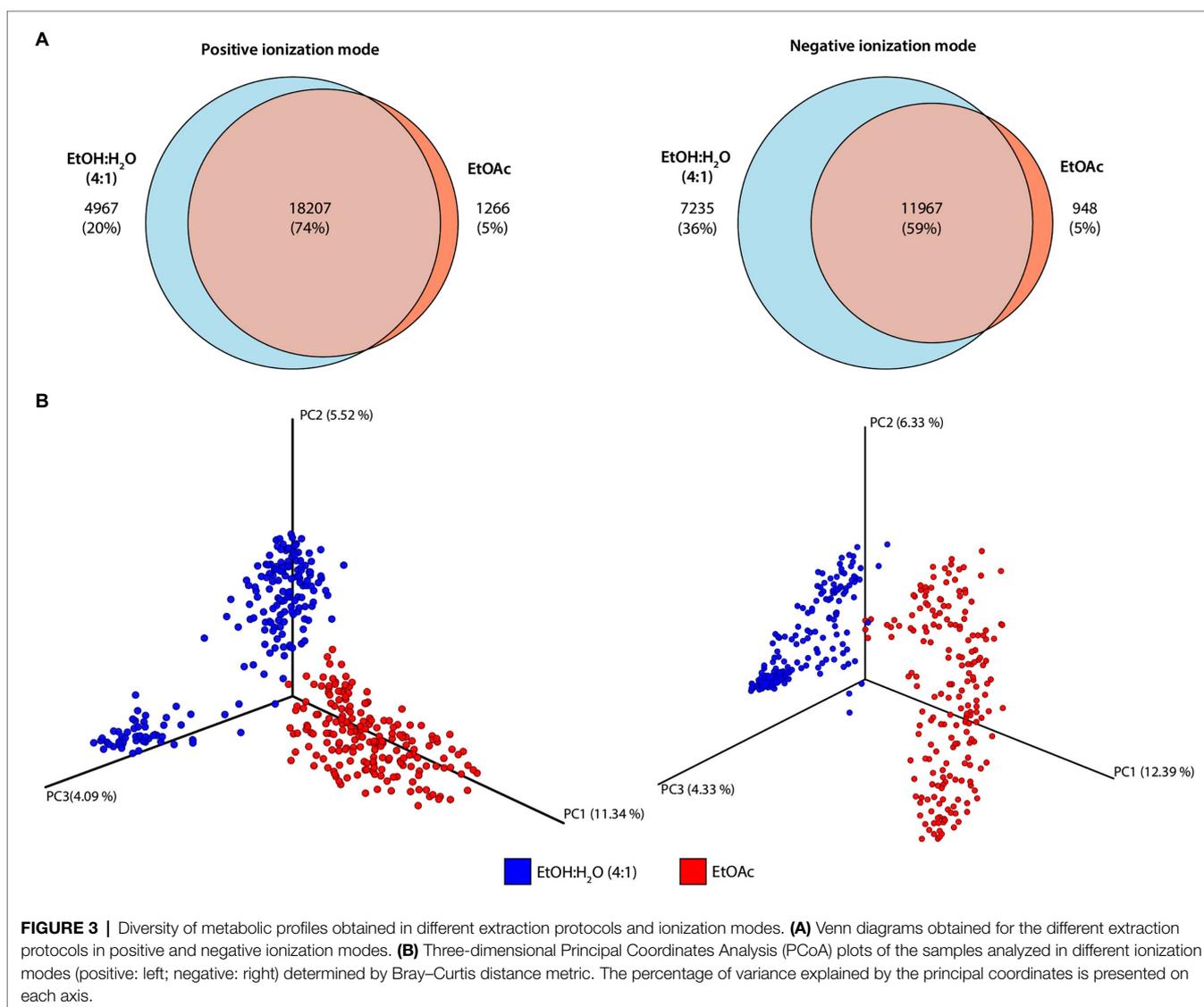
RESULTS AND DISCUSSION

Evaluation of Extraction Solvent and Ionization Mode in Malpighiaceae Chemical Diversity

The extracts (EtOH:H$_2$O, 4:1 v/v; EtOAc) obtained from the 197 Malpighiaceae samples were analyzed by LC–MS/MS in both positive and negative ionization modes, and the LC–MS data were processed in MZmine2. The feature finding step resulted in a total of 24,440 and 20,150 features detected in positive and negative ionization modes, respectively. Usually, most of the dereplication tools based on LC–MS/MS consider the fragmentation of [M+H]$^+$ or [M−H]$^-$ adducts. Generally,

molecules with pH greater than 7 (basic compounds) could be easily ionized in the positive mode making adducts with proton(s). The formation of deprotonated molecules is usually limited to compounds able to form acidic protons in the negative ionization mode. Although it is of high importance for the characterization of secondary metabolites, the signal intensity for data acquired in the negative mode is usually lower compared to the positive ionization mode. In addition, the use of positively charged ion fragmentation is more relevant due to the larger spectral library availability (Steckel and Schlosser, 2019; Wolfender et al., 2019).

The Venn diagrams obtained in positive and negative ionization modes for the different extraction solvents (**Figure 3A**) showed that, in the positive ionization mode, 74% of the features were shared between the two extraction protocols, while in the negative ionization mode, this number dropped to 59%, with the number of metabolites exclusively detected in the hydroethanolic extraction rising from 20 to 36%. Only about 5% of the features were detected exclusively with ethyl acetate as extraction solvent,

FIGURE 3 | Diversity of metabolic profiles obtained in different extraction protocols and ionization modes. **(A)** Venn diagrams obtained for the different extraction protocols in positive and negative ionization modes. **(B)** Three-dimensional Principal Coordinates Analysis (PCoA) plots of the samples analyzed in different ionization modes (positive: left; negative: right) determined by Bray–Curtis distance metric. The percentage of variance explained by the principal coordinates is presented on each axis.

regardless of the ionization mode used. These results are in accordance with previous reports since even though there is a common core metabolome, solvent-specific metabolites are likely to be observed (Crüsemann et al., 2017).

Usually, distinct classes of metabolites are extracted with solvents of different polarities (usually hydroalcoholic mixtures, methanol, ethyl acetate, or methylene chloride), which will allow the enrichment of specific classes of metabolites (such as flavonoids, coumarins, glucosides, alkaloids, diterpenes, and saponins, among others), depending on the solvent polarity (Wolfender et al., 2019; Pilon et al., 2020). In this way, different extraction solvents can be used to obtain a broader chemical diversity. Our results show that the diversity of solvent-specific metabolites will also vary depending on the ionization mode employed.

To observe the chemical space provided by the metabolomic profiles obtained by the different extraction protocols, PCoA of the positive and negative subsets were created (**Figure 3B**). In summary, differently than Principal Component Analysis (PCA) which measures correlations among the samples, PCoA analysis is used to calculate distances among them, and the way these distances are calculated can result in different clustering trends in the plots. When the Euclidean distance is used in PCoA analysis, the result will be the same as if PCA was employed (Mohammadi and Prasanna, 2003; Bauermeister et al., 2022). To evaluate the impact of different extraction protocols, PCoA plots obtained by Bray–Curtis distance metric showed that the two extraction solvents resulted in very different metabolomic profiles on both ionization modes (for positive ionization mode: PERMANOVA $F = 40.39$, $p = 0.001$; for negative ionization mode: PERMANOVA $F = 38.37$, $p = 0.001$). Even though most of the features detected are shared between the two extraction protocols, the compounds' relative abundance may significantly vary resulting in different metabolomic profiles.

In order to determine whether there would be clustering trends for each of the extracts obtained and in both ionization modes, PCoAs for these four subsets were also obtained. For the hydroethanolic extract, two main groups were formed in both ionization modes (**Supplementary Figure 1A**; for positive ionization mode: PERMANOVA $F = 28.59$, $p = 0.001$; for negative ionization mode: PERMANOVA $F = 43.92$, $p = 0.001$). For the ethyl acetate extract, a separation into two groups was obtained in the positive ionization mode (PERMANOVA $F = 18.71$, $p = 0.001$; **Supplementary Figure 2A**), while no group separation was observed in the negative ionization mode in any of the distance matrices tested (**Supplementary Figure 1B**), reinforcing that the extraction solvent and ionization mode influence data acquisition.

The clustering trends observed for the hydroethanolic extracts were optimized using the maximum likelihood criteria into the most recent molecular phylogenetic tree of Malpighiaceae (obtained from Davis and Anderson, 2010), resulting in the cladograms shown in **Supplementary Figure 1B**. These trees are graphical representations of how these different chemical profiles obtained from hydroethanolic extracts originated in specific clades (i.e., groups) through time (i.e., geological time). For the metabolites detected in the positive ionization mode, we can infer that the production of metabolites from group B originated in at least three separate geological times (firstly

in part of clade A, secondly in clade C, and thirdly in the most recent common ancestor of clade E shared with all remaining clades of Malpighiaceae; i.e., clades E–J). In other words, the samples from group B are, in general, evolutionary more recent than the ones from group A. Some specific genera (such as *Amorimia*, *Alicia*, and *Banisteriopsis*), which are more recent in terms of their origins in past geological times, also cluster together with group A. A possible explanation for this tendency is that these genera are mostly found in dry environments, just like some of the early-diverging genera of Malpighiaceae (Davis et al., 2014; de Almeida et al., 2021).

On the other hand, the metabolomic profile obtained in the negative ionization mode for the hydroethanolic extracts showed a different tendency. The different chemical profiles obtained from group B in the negative ionization mode were only recovered as homoplastic to clade J (i.e., *Bronwenia*, *Diplopterys*, *Stigmaphyllon*, *Sphedamnocarpus*, and *Peixotoa*), besides also independently originated in some genera from clades A, E, F, and H.

We also evaluated the metabolomic profile obtained for the ethyl acetate extracts in both ionization modes (**Supplementary Figures 2A,C**). Contrary to what we previously observed, the clustering trends observed in the positive ionization mode did not show any clear correlation with the phylogeny of the Malpighiaceae family (**Supplementary Figure 1B**), while the negative ionization mode did not show any clustering trends in the metrics tested. These differences in both extraction protocols could be explained due to the different polarity ranges covered by these two solvents. From these results, we observed that different extraction protocols and ionization modes in MS can significantly impact the results from multivariate analysis and chemosystematics investigations.
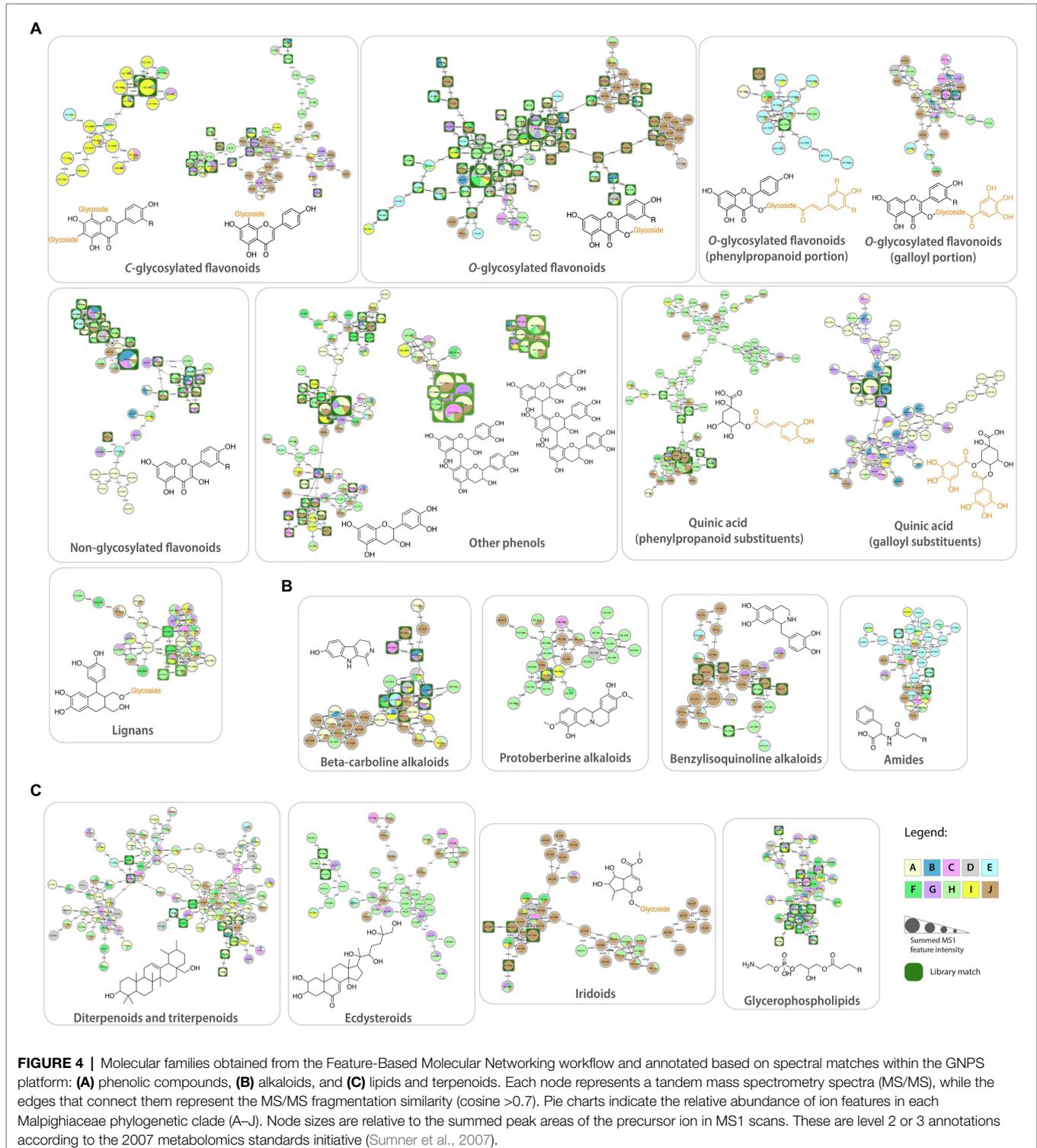
Molecular Families and Metabolite Annotation

Since the chemical profiles obtained for the hydroethanolic extracts showed a promising correlation with the evolution of Malpighiaceae, we made an in-depth investigation of the data obtained in both positive and negative ionization modes. The MS/MS library search performed in GNPS resulted in 1,070 spectral matches for the positive and 1,025 matches for the negative ionization modes, resulting in 4.6 and 5.3% of the detected chemical spaces, respectively. These matches were manually evaluated and compared with the literature, resulting in level 2 or 3 annotations according to the MSI (Sumner et al., 2007). In addition, these compounds were also searched in the *Dictionary of Natural Products*, LOTUS (Rutz et al., 2021), and Scifinder databases to inspect for previous reports in Malpighiaceae.

In addition to the library searches, the MS/MS data were visualized by molecular networking analysis. The molecular families constructed by such analysis represent the similarity of fragmentation patterns obtained by tandem mass spectrometry (MS/MS) analysis. In other words, structurally similar compounds will present similar chemical stability and functional groups, leading to similar fragmentation patterns (Yang et al., 2013). These molecular families consist of nodes (representing MS/MS spectra) and of edges connecting these nodes (representing

the cosine similarity between two nodes, which measure the relatedness in MS/MS spectra; Aron et al., 2020). In this way, depending on the cosine score set in the analysis, the connections between the nodes can be more, or less strict. All the molecular families discussed in this section can be found in the **Supplementary Material**.

The library matches retrieved from the analyses obtained in the positive ionization mode showed the presence of a high diversity of classes of compounds, including *C*-glycosylated and *O*-glycosylated flavonoids, lipids, alkaloids, quinic acid derivatives, amides, triterpenes, iridoids, and lignans (**Figure 4**). Some of these chemical classes were widely detected in all phylogenetic

FIGURE 4 | Molecular families obtained from the Feature-Based Molecular Networking workflow and annotated based on spectral matches within the GNPS platform: **(A)** phenolic compounds, **(B)** alkaloids, and **(C)** lipids and terpenoids. Each node represents a tandem mass spectrometry spectra (MS/MS), while the edges that connect them represent the MS/MS fragmentation similarity (cosine >0.7). Pie charts indicate the relative abundance of ion features in each Malpighiaceae phylogenetic clade (A–J). Node sizes are relative to the summed peak areas of the precursor ion in MS1 scans. These are level 2 or 3 annotations according to the 2007 metabolomics standards initiative (Sumner et al., 2007).

clades, while others were more specific to particular clades or even genera.

Several molecular families related to flavonoid compounds were obtained, highlighting the difference in the fragmentation pattern in MS/MS spectra among these subclasses with different substituents. For instance, the usual MS/MS spectra obtained for *O*-glycosylated flavonoids in electrospray ionization consist of the neutral losses of the glycosidic substituents, such as hexosides (162 Da), deoxyhexosides (146 Da), and pentosides (132 Da), eventually reaching the aglycone. On the other hand, *C*-glycosylated flavonoids present a very distinct fragmentation pathway, with many more fragments observed between the precursor ion and the aglycone. The loss of the glycosidic portion is usually reached by the consecutive losses of water molecules and 120 Da, characteristic of *C*-glycosylated flavonoids (Mannochio-Russo et al., 2020). Therefore, the distinction of these two molecular families in the molecular networking analysis is expected. In addition, compounds containing two *C*-glycosidic bounds, even more fragments are expected, and we can observe two molecular families for *C*-glycosylated flavonoids.

Flavonoids with two *C*-glycosylated portions were mainly detected in samples from clade I, especially from the genera *Amorimia* and *Mascagnia* (**Supplementary Figure 3**). Library matches from this molecular family included apigenin-di-*C*-hexoside-pentoside and luteolin-di-*C*-hexoside, being in accordance with the previous reports for the *Amorimia* genus (Mannochio-Russo et al., 2020). On the other hand, spectral matches to flavonoids containing only one *C*-glycosylated portion (**Supplementary Figure 4**) were less clade-specific, being detected in samples from clades A, G, H, I, and J.

In addition to the *C*-glycosylated flavonoids, three main molecular families were observed for *O*-glycosylated ones: (1) flavonoids containing only glycosides, (2) glycosides in addition to phenylpropanoids, and (3) glycosides in addition to galloyl portions. These are also explained by the difference in general MS/MS spectra of each of these groups, in which the flavonoids containing the galloyl portion present a characteristic fragment at *m/z* 153, and the ones containing phenylpropanoids will show the fragment ions regarding this substituent (*m/z* 147 for coumaric acid, for instance). Finally, *O*-glycosylated flavonoids containing only glycosides will only present neutral losses regarding each glycosidic portion.

Several matches to *O*-glycosylated flavonoids were observed in all Malpighiaceae clades (**Supplementary Figure 5**), which is expected since this class of compounds is considered ubiquitous to plant species (Buer et al., 2010). This network is primarily composed of quercetin and kaempferol derivatives bound to glycosidic portions. A small cluster in this molecular family corresponds to compounds mostly present in clade J, specifically from *Stigmaphyllon* species. These compounds were annotated as flavonoids with glucuronide, acetylated, and malonylated hexoside substituents, which have not been reported in the literature for the genus *Stigmaphyllon* to date.

O-glycosylated flavonoids containing sugars and phenylpropanoid portions as substituents were also observed (**Supplementary Figure 6**), mainly present in the *Ptilochaeta*

genus (clade E), which is in accordance with previous reports (Mannochio-Russo et al., 2020). In addition, sugars bound to galloyl portions (**Supplementary Figure 7**) were also detected as flavonoid substituents and widely found in all Malpighiaceae clades.

A network containing mainly non-glycosylated flavonoids (**Supplementary Figure 8**), widely distributed in all clades, was also observed, in which only the fragments relative to aglycones were detected in the MS/MS spectra. These compounds can represent, indeed, aglycones, or even in-source fragments of the *O*-glycosylated flavonoids. Additionally, a molecular family with spectral matches to catechin and afzelechin and their derivatives (**Supplementary Figure 9**) was also formed, containing mainly methoxylated portions and galloyl and sugar substituents.

Quinic acid derivatives were detected in higher amounts in clades H and J for phenylpropanoid substituents, and in clades A, B, and G for gallic acid substituents (**Supplementary Figures 10, 11**). Usually, the MS/MS patterns will contain an ion relative to the quinic acid moiety (usually with additional neutral loss of water), in addition to the fragment at *m/z* 153, relative to the galloyl portion, while other characteristic fragments will be observed for the phenylpropanoid derivatives (Mannochio-Russo et al., 2020). Previous reports described the presence of galloylquinic acids in *Byrsonima* (clade A) species (Fraige et al., 2018; Mannochio-Russo et al., 2020), while quinic acids containing phenylpropanoids substituents have been described for the genus *Heteropterys* (clade H; Huerta-Reyes et al., 2013; Paula-Freire et al., 2013).

Many lipid-like molecules presented library matches, mainly corresponding to the glycerophospholipids (**Supplementary Figure 12**), fatty acids, and fatty esters (**Supplementary Figure 13**) classes, besides jasmonic acid derivatives (**Supplementary Figure 14**), which were largely distributed in all Malpighiaceae clades. Glycerophospholipids also present key fragments in MS/MS analysis, such as the cleavage of the choline group, which can generate a fragment of the choline group itself (if positively charged) and the fragment of the long-chain portion (Ivanova et al., 2007). Lipids represent an important class of compounds, widely found in plants, with key roles in multiple signaling processes (Mamode Cassim et al., 2019). Along with sugars, such compounds are among the main constituents of the oil glands present in Malpighiaceae leaves and flowers (Possobom et al., 2010). Lipids are mainly produced by oil glands that play an important ecological role in Malpighiaceae, a botanical family which is mainly pollinated by oil collecting bees, and the oldest family characterized by oil-bee pollination (Anderson, 1990; Renner and Schaefer, 2010; Davis et al., 2014).

β-carboline alkaloids and other tryptophan derivatives were detected in samples from many clades (**Supplementary Figure 15**), including *Tetrapterys* and *Banisteriopsis* species, corroborating previous reports (Samoylenko et al., 2010; Queiroz et al., 2014). In general, β-carboline alkaloids present MS/MS spectra with fragments relative to the loss of a hydroxyl group (if present) and to the formation of a four-membered ring (*m/z* 184 and *m/z* 160, respectively, harmalol as an example). A small cluster into this molecular family showed the presence of compounds

without library matches with higher *m/z* in samples from clade A and J, mainly in *Janusia*, *Banisteriopsis*, and *Byrsonima* species, which can indicate the presence of glycosylated portions due to the mass differences observed between the nodes.

Networks containing library matches corresponding to isoquinoline, protoberberine, and benzylisoquinoline alkaloids were also observed (**Supplementary Figures 16–18**, respectively). Isoquinoline alkaloids were mainly observed in samples from clades G, H, and J. Berberine alkaloids usually show a retro Diels-Alder (RDA) reaction and a B-ring cleavage, forming fragments *m/z* 151 and *m/z* 178, respectively, in the case of scoulerine (Qing et al., 2020). On the other hand, benzylisoquinoline alkaloids will usually present some key fragments, such as the initial loss of the nitrogen atom as ammonia or as methylamine (in the case of methylated nitrogens), in addition to an "even electron"-type McLafferty rearrangement with a reversed charge distribution, and a fragment relative to the benzyl moiety. Therefore, these key fragmentation pathways will generate ions at *m/z* 299, *m/z* 192, and *m/z* 137, taking reticuline as an example (Schmidt et al., 2005). The molecular family representing protoberberine alkaloids showed nodes corresponding to compounds mostly detected in clades H and J, particularly in *Stigmaphyllon* and *Alicia* species. Similarly, benzylisoquinoline alkaloids were also mainly detected in samples from clades H and J, especially in *Stigmaphyllon* species. These classes of compounds have not been described to date for Malpighiaceae species and represent important traits for chemosystematics studies in Malpighiaceae due to their specificity to the above-mentioned clades, especially for protoberberine and benzylisoquinoline alkaloids.

Spectral matches corresponding to amides and polyamines were also observed in molecular families (as shown in **Supplementary Figures 19**, **20**, respectively). These compounds can generate ions relative to the *N*-cleavage of the amide bond (Barrère et al., 2014). Amides were mainly detected in the *Ptilochaeta* genus (clade E) as fatty amides and small dipeptides, while the polyamines were primarily observed in samples from clades E, H, I, and J, being this the first report of these classes of compounds for Malpighiaceae.

Several molecular families related to terpenoids were also observed, which are characterized for presenting several fragments in MS/MS spectra depending on the general skeleton of the molecule, such as RDA, McLafferty rearrangement, water losses, among others (Demarque et al., 2016). One of the largest networks obtained (**Supplementary Figure 21**) showed library matches to triterpenoids and their precursors. These compounds were also largely distributed among all Malpighiaceae clades, in accordance with previous reports. In fact, triterpenoids are the most described class of compounds from Malpighiaceae species to date, with numerous reports in the literature for *Acridocarpus*, *Byrsonima*, and *Galphimia* genera (Cao et al., 2004; Cardoso Taketa et al., 2004; Aguiar et al., 2005).

It is important to mention at this point that the molecular networking approach highlights the chemical similarity of compounds based on their fragmentation patterns, which may not necessarily reflect their biosynthetic origin. For instance, some diterpenoids are grouped in the same molecular family

as the triterpenoids, even though their biosynthesis differs from the plastidial MEP and the cytosolic mevalonate pathway. The connection between them can be observed since these two classes of compounds present similar MS/MS fragmentation patterns. A possible way to separate these compounds into two distinct molecular families would be to significantly increase the threshold for the cosine score similarity (set to 0.7 in this work). However, in the molecular networking analysis, the cosine value is set for the entire dataset, and thus, we used an intermediary value that was adequate for most of the families discussed in this work. Since different chemical classes present different key MS/MS fragments, it is very likely that while separating triterpenoids from diterpenoids, other molecular families can be fragmented into smaller networks, even though they can present significant correlations. In addition, it is important to emphasize that even though we can increase this cosine threshold, it is very likely that the diterpenoids and triterpenoids will not be perfectly separated since these two classes share several key MS/MS fragments.

A separate molecular family relative to ecdysteroids, another class of terpenoids, was also observed (**Supplementary Figure 22**), and mainly detected in samples from clade H, particularly in *Niedenzuella* species. These compounds are mainly characterized by the presence of consecutive neutral losses of water molecules, in addition to the hydrocarbon side chain cleavage (Lavrynenko et al., 2013; Mannochio-Russo et al., 2020). The presence of ecdysteroids in *Niedenzuella multiglandulosa* was recently described in the literature (Mannochio Russo et al., 2020; Mannochio-Russo et al., 2020), being the only reports of this class of compounds in the Malpighiaceae family. Our results showed a wide diversity of ecdysteroids in all six *Niedenzuella* species sampled, being possible chemical markers for the *Niedenzuella* genus. This represents about 40% of all the *Niedenzuella* species reported to date, and future studies with other *Niedenzuella* species should be performed to corroborate this hypothesis. Moreover, these steroids were also detected in *Hiraea*, *Tetrapterys* (including in the recently segregated *Glicophyllum* species, evolutionary close to *Niedenzuella*), and *Peixotoa* genera (clades G, H, and J, respectively).

Iridoids, another relevant class of terpenoids, were also annotated based on library matches, which are shown in **Supplementary Figure 23**. These compounds present key fragments relative to the neutral losses of the glycosidic portions, in addition to water losses and other possible substituents (Wu et al., 2010). The correspondent networks revealed that these iridoids are mainly present in samples from clade J, especially in *Stigmaphyllon* species, being in accordance with previous reports (Sainty et al., 1981; Davioud et al., 1985). To date, iridoids have only been reported for *Stigmaphyllon* genus in Malpighiaceae, and in this study, we observe that these compounds are also present in samples from other clades, in particular, *Heteropterys oberdanii* species (clade H). Furthermore, a molecular family with a spectral match to a secoiridoid (**Supplementary Figure 24**) was also observed, with part of the compounds widely distributed among all the clades, and part of them mainly detected in *Stigmaphyllon* species. Finally, molecular families relative to

neolignans and furofuranoid lignans (**Supplementary Figure 25**) were observed and widely distributed in Malpighiaceae clades, which have not been previously reported in this family to date.

The library matches obtained for the negative ionization mode and the molecular networks formed were also inspected, and these analyses showed mainly the same classes of compounds described for the positive ionization mode. Two molecular families stood out for presenting library matches to classes of compounds that were not observed in the positive ionization mode: proanthocyanidins and lignans (**Supplementary Figures 26, 27**). Proanthocyanidins dimers and trimers were observed in all Malpighiaceae clades, with characteristic fragmentation patterns (such as RDA), corroborating previous reports (Fraige et al., 2018; Mannochio-Russo et al., 2020). Lignans have already been reported for *Tetrapterys mucronata* (Queiroz et al., 2014), and here we observe that, in addition to *T. mucronata* species (clade H), these compounds are also present in other clades.

It is important to emphasize that the compounds discussed above represent only a part of the compounds detected in this study. In fact, about 87% of the molecular families (composed of two nodes or more) obtained in both ionization modes did not show any spectral match, including networks composed mainly or exclusively by nodes representing specific clades/genera. These numbers point to the possibility of undescribed natural products. On the other hand, even considering only the spectral matches, it was possible to obtain important information regarding the classes of compounds produced by these plant species. Another crucial point to be considered is that the sampling in this study comprises a larger number of samples from clades A, H, I, and J, and that these clades presented most of the spectral matches observed. It is possible that if a greater number of samples from other clades are included, more spectral matches can be retrieved to give more insights on other clades as well.

In silico Metabolite Annotation and Chemical Hierarchy Analysis

In order to amplify the chemical space from the Malpighiaceae dataset and have additional information about the classes of compounds detected, we used the Qemistree workflow combined with the CANOPUS classification tool for systematic compound class annotation (Dührkop et al., 2021; Tripathi et al., 2021). These *in silico* classifications consist of level 3 annotations according to the MSI (Sumner et al., 2007). In this way, it was possible to construct a chemical tree based on molecular fingerprints from MS/MS spectra and *in silico* classification tools. A total of 7,489 and 3,773 fingerprints were generated and classified at a superclass level for the positive and negative ionization modes, respectively. In this way, two chemical hierarchy trees were obtained, as shown in **Figure 5A**.

From the results obtained, it is evident that the ionization mode employed greatly influences the classifications obtained. At a superclass level, the ones most retrieved in the positive ionization modes were the "organic acids and derivatives," followed by "benzenoids" and "organoheterocyclic compounds." In contrast, most of the superclasses retrieved in the negative ionization

mode consisted of "lipids and lipid-like molecules," "organic acids and derivatives," and "benzenoids" (**Supplementary Figure 28A**). At a CANOPUS class level, the "carboxylic acids and derivatives" was the main recovered class in both ionization modes. Ion features classified as "benzene and substituted derivatives," and "azoles" were also observed several times for the positive ionization mode, while for the negative ionization mode, "organooxygen compounds" and "fatty acyls" were the second and third most abundant classifications.

These results confirm some of the conclusions drawn from the molecular networks, such as the presence of lipids and lipid-like molecules in several Malpighiaceae clades. The molecular networks and library searches in spectral libraries, combined with the *in silico* approaches based on structural databases allowed us to expand the Malpighiaceae chemical space. Several hypotheses raised from the molecular networks were corroborated with the *in silico* classifications, giving higher confidence in these results.
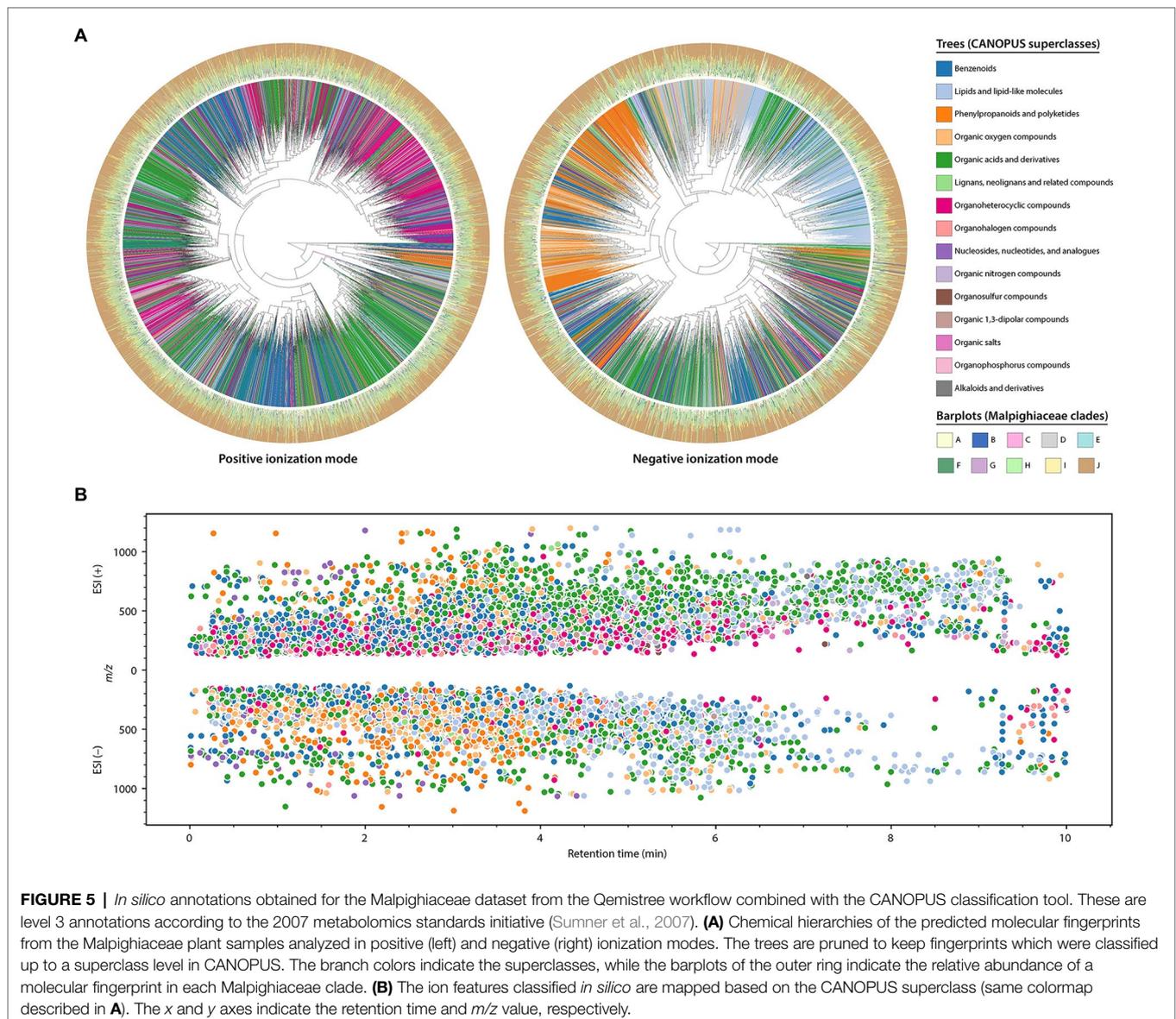
The classifications obtained in the different ionization modes are also shown distributed in the chromatographic run (**Figure 5B**). In addition to being possible to observe differences in classifications between ionization modes, the retention time ranges also vary for specific superclasses. For instance, in the negative ionization mode, organic oxygen compounds, and phenylpropanoids and polyketides elute from near the dead volume to approximately 4 min, while the ion features classified as lipids and lipid-like molecules present higher retention times.

The putative chemical classes retrieved as highly correlated with the most sampled clades (clades A, F, G, H, I, and J; ANOVA $p < 0.05$) were selected to build a heatmap (**Supplementary Figure 29**). The normalized distribution pattern of the different classes within the sampled genera showed that specific classes are significantly enriched in determined genera/clades, which corroborates our observations retrieved from the molecular networks. In addition, once again, the results obtained for the positive and negative ionization modes differ.

These results show that both ionization modes result in complementary chemical classifications, which is crucial for comprehensive chemotaxonomic investigations. In fact, the *in silico* tools used in this study relies on public spectral and structural databases, which are known to be more populated with data on the positive ionization mode (Tripathi et al., 2021). In this way, the results obtained for the negative ionization mode are less extensive compared to the positive ionization mode.
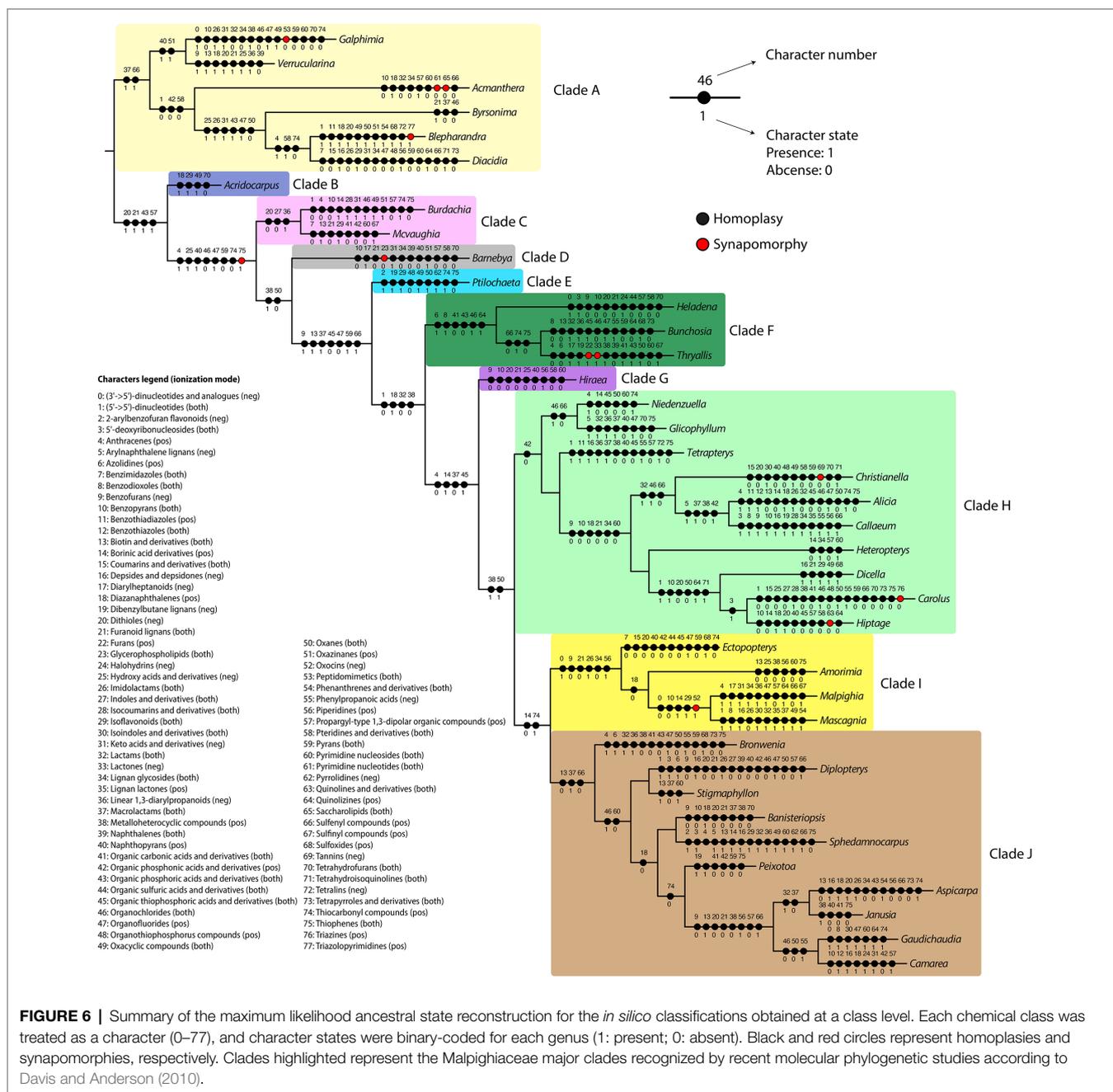
Ancestral Character Reconstructions for the Classes of Secondary Metabolites Annotated in Malpighiaceae

Ancestral character reconstruction analyzes have been increasingly encouraged in natural products studies since the early 2010s (Schmitt and Barker, 2009). It has been used in the chemistry of natural products to investigate chemical evolutionary relationships comprising different organisms, such as plants, fungi, and animals (Lumbsch et al., 2006; Bondoc et al., 2013; Allevato et al., 2019; Coley et al., 2019; Chen et al., 2020; Beaulieu et al., 2021). Phylogenetic methods have proven to be a promising approach to explore the evolution

FIGURE 5 | *In silico* annotations obtained for the Malpighiaceae dataset from the Qemistree workflow combined with the CANOPUS classification tool. These are level 3 annotations according to the 2007 metabolomics standards initiative (Sumner et al., 2007). **(A)** Chemical hierarchies of the predicted molecular fingerprints from the Malpighiaceae plant samples analyzed in positive (left) and negative (right) ionization modes. The trees are pruned to keep fingerprints which were classified up to a superclass level in CANOPUS. The branch colors indicate the superclasses, while the barplots of the outer ring indicate the relative abundance of a molecular fingerprint in each Malpighiaceae clade. **(B)** The ion features classified *in silico* are mapped based on the CANOPUS superclass (same colormap described in **A**). The *x* and *y* axes indicate the retention time and *m/z* value, respectively.

of chemical compounds in a specific genus or family of plants and other living organisms. This analysis consists of optimizing binary character states (presence/absence) into a DNA-based molecular phylogenetic tree by statistically testing using Maximum Likelihood Estimation, and recovering which character states (such as chemical classes) characterizes a given clade or taxonomic group within the molecular phylogenetic tree (i.e., genera or major clades recovered by DNA-based phylogenetic studies in Malpighiaceae). In this way, we can determine the statistical probability of the ancestor of a particular genus or clade to exclusively (i.e., synapomorphy) or non-exclusively (i.e., homoplasy) show the presence or absence of a specific chemical class. These exclusive or non-exclusive chemical classes can be used in future studies to circumscribe the analyzed taxa into a new classification, which additionally considers chemical information.

In the present study, the *in silico* classifications obtained were used to map the evolutionary history of the classes retrieved in our analyses using the maximum likelihood criteria in the Malpighiaceae DNA-based molecular phylogeny (**Figure 6**). The classifications obtained in the positive and negative ionization modes were combined to provide a general overview of the total number of chemical classes obtained for the family. Classes retrieved as homoplasies or synapomorphies for the major Malpighiaceae clades (A–J) are described in **Table 1**, while the ones retrieved for all Malpighiaceae clades and genera are extensively presented in **Supplementary Table 2**. From the 113 *in silico* classes retrieved, 35 were present in all genera, such as fatty acyls, flavonoids, glycerolipids, phenols, and purine nucleosides, compounds widely distributed in plants, with a variety of ecological roles in these organisms. Nonetheless, future additional studies sampling the remaining botanical

FIGURE 6 | Summary of the maximum likelihood ancestral state reconstruction for the *in silico* classifications obtained at a class level. Each chemical class was treated as a character (0–77), and character states were binary-coded for each genus (1: present; 0: absent). Black and red circles represent homoplasies and synapomorphies, respectively. Clades highlighted represent the Malpighiaceae major clades recognized by recent molecular phylogenetic studies according to Davis and Anderson (2010).

families included in the order Malpighiales are needed to properly evaluate the relevance of these 35 classes of metabolites for the Malpighiaceae family as a whole.

Regarding the 10 major clades of Malpighiaceae, all of them were recovered with at least one homoplastic or synapomorphic class of metabolites supporting them (**Table 1**; **Supplementary Table 2**). The 10 major clades in Malpighiaceae were characterized by the presence and absence of 22 and 23 classes of metabolites (i.e., homoplasies), respectively, with the absence of glycerophospholipids being recovered as a synapomorphy of Clade D. It is worth mentioning that even though most classes of secondary metabolites circumscribing

all the 10 major clades of Malpighiaceae were recovered as homoplasies, their presence or absence was recorded exclusively for each clade, with reversions (i.e., parallelisms) only being recorded within a few distantly related subclades or genera. The results retrieved from this analysis corroborate some of the conclusions obtained from the molecular networks and from the *in silico* classifications. For instance, glycerophospholipids, fatty acids, fatty esters, furofuranoid lignans, and prenol lipids classes were widely distributed in our molecular networking analyses for all phylogenetic clades, and these classes were also extensively recovered in the ancestral character reconstructions (i.e., appeared in different geological

TABLE 1 | Characters retrieved from the ancestral characters reconstruction (clades) based on the classifications obtained *in silico* for Malpighiaceae samples.

	Classes present[†]	Classes absent
Clade A (Byrsonimoid clade)	Macrolactams (both); Sulfenyl compounds (pos)	–
Clade B (Acridocarpoid clade)	Diazanaphthalenes (pos); Isoflavonoids (both); Oxacyclic compounds (both)	Tetrahydrofurans (both)
Clade C (Mcvaughioid clade)	Linear 1,3-diarylpropanoids (neg)	Dithioles (neg); Indoles and derivatives (both)
Clade D (Barnebyoid clade)	Diarylheptanoids (neg); Keto acids and derivatives (neg); Oxazinanes (pos)	Benzopyrans (both); Furanoid lignans (both); Glycerophospholipids (both)[§], Lignan glycosides (both); Naphthalenes (both); Naphthopyrans (pos); Propargyl-type 1,3-dipolar organic compounds (pos); Pteridines and derivatives (both); Tetrahydrofurans (both)
Clade E (Ptilochaetoid clade)	2-arylbenzofuran flavonoids (neg); Dibenzylbutane lignans (neg); Isoflavonoids (both); Oxacyclic compounds (both); Oxanes (both); Pyrrolidines (neg); Thiocarbonyl compounds (pos)	Organothiophosphorus compounds (pos); Thiophenes (both)
Clade F (Bunchosioid clade)	Azolidines (pos); Benzodioxoles (both); Organochlorides (both); Quinolizines (pos)	Organic carbonic acids and derivatives (both); Organic phosphoric acids and derivatives (both)
Clade G (Hiraeoid clade)	Piperidines (pos)	Benzofurans (neg); Benzopyrans (both); Dithioles (neg); Furanoid lignans (both); Hydroxy acids and derivatives (neg); Naphthopyrans (pos); Pteridines and derivatives (both); Pyrimidine nucleosides (both)
Clade H (Tetrapteroid clade)	–	Organic phosphonic acids and derivatives (pos)
Clade I (Malpighioid clade)	(3′➔5′)-dinucleotides and analogues (neg); Imidolactams (both); Piperidines (pos)	Benzofurans (neg); Furanoid lignans (both); Lignan glycosides (both)
Clade J (Stigmaphylloid clade)	Macrolactams (both)	Biotin and derivatives (both); Sulfenyl compounds (pos)

The ionization mode in which each classification was obtained is described (pos, positive ionization mode; neg, negative ionization mode; and both, both ionization modes).
[†]Classes retrieved as present in all Malpighiaceae clades: Alkyl halides (both); Allyl-type 1,3-dipolar organic compounds (both); Aryl halides (both); Azacyclic compounds (both); Azoles (both); Benzene and substituted derivatives (both); Boronic acid derivatives (both); Carboxylic acids and derivatives (both); Cinnamic acids and derivatives (both); Diazinanes (pos); Diazines (both); Fatty Acyls (both); Flavonoids (both); Glycerolipids (both); Heteroaromatic compounds (both); Imidazopyrimidines (both); Macrolides and analogues (pos); Organic metal salts (pos); Organic sulfonic acids and derivatives (both); Organonitrogen compounds (both); Organooxygen compounds (both); Phenol ethers (both); Phenols (both); Prenol lipids (both); Purine nucleosides (both); Purine nucleotides (both); Pyridines and derivatives (both); Sphingolipids (both); Steroids and steroid derivatives (both); Sulfonyls (pos); and Thioethers (both).
[§]Synapomorphy.

times). In addition, it also clearly shows which classes were important to circumscribe specific clades/genera.

Our chemotaxonomic approach based on MS/MS analyses of Malpighiaceae plant samples allowed us to obtain a comprehensive overview of the classes of secondary metabolites produced by this taxon. Plant secondary metabolites are known to show patterns of occurrence in certain taxa (Wink, 2003). However, it is important to emphasize that the chemical diversity of a sample is highly influenced by many factors, such as genetic variation and environmental influences (soil nutrients, humidity, herbivory, and ecological interactions, among others; Isah, 2019). Studies have shown that even the same species collected in different biomes (Bueno et al., 2021) or different seasons (Zanatta et al., 2021) can produce different relative amounts of specific metabolites. Therefore, even for a single species, a range of factors can be explored to understand how the metabolites are affected to infer their possible ecological roles. However, it is important to emphasize that the ancestral character reconstructions obtained only focus on qualitative characters (presence and absence) rather than quantitative ones.

Our study provides a starting point for follow-up and systematic evaluation of such factors, and in-depth studies must be conducted to confirm and expand these chemotaxonomic conclusions for both Malpighiaceae and Malpighiales. In fact, there is a big jump from a large (but limited) sampling, as the one we present here, toward a chemotaxonomic investigation of the entire plant family. Nonetheless, as shown by several

recent studies mentioned above, using phylogenetic methods with chemistry of natural products data is a promising and revolutionary new line of research that aims to elucidate the evolution of specialized metabolites in living organisms. Therefore, future investigations must be conducted to confirm the hypotheses raised in our study, especially for species that do not present any previous phytochemical study. In addition, our efforts were directed to obtain chemical information at a major clades and generic levels. There might be discrepancies if the ancestral character state reconstructions are obtained at different taxonomic levels, such as intrageneric and species levels.

Finally, the relevance of our evolutionary approach to the study of secondary metabolites can be evidenced if we consider as examples three relevant classes of plant secondary metabolites (**Table 1**: furanoid lignans, isoflavonoids, and piperidines). The absence of the phenylpropanoids furanoid lignans in clades D, G, and I is regarded as an informative homoplasy in our analyzes since this information can be used to point which clades in the family one must focus on in future studies to search for this class of metabolites. This information can also be used in chemosystematic studies of Malpighiaceae to chemically characterize these clades and help establish a new classification system based on chemical compounds and morphology, following de Almeida et al. (2017). On the other hand, the presence of the phenolic isoflavonoids in clades B and E, and the alkaloid class of piperidines in clades G and I are also informative homoplasies that can be interpreted in the same light as the

furanoid lignans. In addition, it is important to emphasize that the *in silico* classifications obtained rely on structural databases populated with compounds from diverse sources, including plants and microorganisms metabolites. Therefore, some *in silico* classifications of metabolites more usually found in microorganisms may occur in plant datasets (for instance, the features classified as lactams). A deeper investigation in each taxonomic group should be performed in the future to confirm their presence. In fact, the population of such databases with more compounds derived from plants is necessary to have more accurate information regarding plant species in the future.

The chemical characterization of all Malpighiaceae clades is the first step toward enabling a new research line on the evolution of secondary metabolites in this plant family since this family already has a dated and calibrated molecular phylogeny available in the literature (Wink, 2003; Davis et al., 2014). Merging both analyses would allow us to infer the geological time in which all classes of secondary metabolites have arisen in Malpighiaceae and correlate these dates with past biogeographic events, such as the colonization of different biomes by the most recent common ancestor of all analyzed clades (de Almeida et al., 2018). Additionally, it would also be possible to evaluate which Malpighiaceae lineages experienced a higher diversification throughout the geological time and identify, and which classes of secondary metabolites are correlated with these diversification events (Xi et al., 2012).

CONCLUSION

Metabolomics analyses based on tandem mass spectrometry and bioinformatics tools have enabled a more comprehensive investigation of the metabolites produced by organisms, and have been increasingly used for this purpose. In fact, due to the low amount of material necessary, it is possible to investigate entire families for chemoevolutionary studies based on the remaining samples from molecular phylogenies used for total DNA extraction. Even though there was a significant advance in this field, many caveats must be considered for proper use. The ionization mode and extraction protocols must be carefully evaluated since these factors influence the results, especially when aiming at chemotaxonomic investigations. Our results showed that positive and negative ionization modes lead to complementary results both in library searches and *in silico* classification tools. However, as the public libraries are more populated with data acquired in the positive ionization mode, less extensive information can be retrieved from analyses performed in the negative ionization mode. More complete and precise results will certainly be obtained for chemotaxonomic studies as these databases get more populated and new bioinformatic tools are developed. Similarly, conclusions obtained from *in silico* approaches must be confirmed with complementary techniques, and the classical methodologies are of great value for deeper investigations. In addition, the population of structural databases with more plant-derived compounds will be of great value to have more accurate results for these organisms.

Our study explored several Malpighiaceae plant species, genera, and clades for the first time, which greatly improved

the chemical knowledge of this family. There are several challenges in performing chemotaxonomic investigations at a plant family level, and the evolutionary conclusions retrieved must be carefully inspected; however, they can be of great value to underpin interesting features in the chemodiversity of a certain taxon. We hope that our findings guide future studies in Malpighiaceae as we reported evidence of specific classes of compounds that most likely occur in specific clades or genera. Therefore, if a particular chemical class is of interest for presenting specific biological activities, one can focus their search on specific groups pointed out in this study. All the software and libraries used in this study are publicly available, making this workflow accessible to be reproduced in other taxa. In addition, we expect that the workflow followed in this study will be used in future studies in several fields, such as chemotaxonomy, metabolomics, chemical ecology, and for the discovery of new natural products.

DATA AVAILABILITY STATEMENT

The mass spectrometry data can be accessed on the Mass spectrometry Interactive Virtual Environment (MassIVE) as the dataset MSV000085119, which is publicly available. The Feature-Based Molecular Networking jobs on GNPS can be accessed online at: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2c5f11403ac847a298e4d7866a491143 and https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=501c16500476451f978311057266fbdf for positive and negative ionization modes, respectively.

The Cytoscape visualization files, the feature tables (.csv) and ion information (.mgf) files exported from MZmine2, and the files generated in the Qemistree workflow in Qiime2 are also available in MSV000085119. The chemical hierarchy tree for interactive visualization and scripts used in this project are available in https://github.com/helenamrusso/Malpighiaceae_supplementary.

AUTHOR CONTRIBUTIONS

HM-R designed the study, performed the extraction and LC–MS/MS analyses, processed and analyzed the MS/MS data, performed the ancestral character reconstructions, inspected the results, wrote the manuscript, and revised the manuscript. RA designed the study, collected and/or identified the species, performed the ancestral character reconstructions, inspected the results, wrote the manuscript, and revised the manuscript. WDGN assisted in the MS/MS data analysis, inspected the results, and revised the manuscript. PCPB inspected the results and revised the manuscript. AC-R performed the extraction and LC–MS/MS analyses, inspected the results, and revised the manuscript. AB inspected the results and revised the manuscript. PD designed the study, acquired funding, and revised the manuscript. VB designed the study, acquired funding, and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.854842/full#supplementary-material

REFERENCES

Aguiar, R. M., David, J. P., and David, J. M. (2005). Unusual naphthoquinones, catechin and triterpene from *Byrsonima microphylla*. *Phytochemistry* 66, 2388–2392. doi: 10.1016/j.phytochem.2005.07.011

Allevato, D. M., Groppo, M., Kiyota, E., Mazzafera, P., and Nixon, K. C. (2019). Evolution of phytochemical diversity in *Pilocarpus* (Rutaceae). *Phytochemistry* 163, 132–146. doi: 10.1016/j.phytochem.2019.03.027

Anderson, W. R. (1990). The origin of the Malpighiaceae-The evidence from morphology. *Mem. N. Y. Bot. Gard.* 64, 210–224.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral. Ecol.* 26, 32–46. doi: 10.1111/j.1442-9993.2001.01070.pp.x

Anderson, W. R. (2006). Eight segregates from the Neotropical genus *Mascagnia* (Malpighiaceae). *Novon J. Bot. Nomenc.* 16, 168–204. doi: 10.3417/1055-3177(2006)16[168:ESFTNG]2.0.CO;2

Anderson, C. (2011). Revision of *Ryssopterys* and transfer to *Stigmaphyllon* (Malpighiaceae). *Blumea* 56, 73–104. doi: 10.3767/000651911X573444

Anderson, W. R., and Davis, C. C. (2006). Expansion of *Diplopterys* at the expense of *Banisteriopsis* (Malpighiaceae). *Harv. Pap. Bot.* 11, 1–16. doi: 10.3100/1043-4534(2006)95[1:EODATE]2.0.CO;2

Angiosperm Phylogeny Group (1998). An ordinal classification for the families of flowering plants. *Ann. Mo. Bot. Gard.* 85, 531–553. doi: 10.2307/2992015

Angiosperm Phylogeny Group (2003). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.* 141, 399–436. doi: 10.1046/j.1095-8339.2003.t01-1-00158.x

Angiosperm Phylogeny Group (2009). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot. J. Linn. Soc.* 161, 105–121. doi: 10.1111/j.1095-8339.2009.00996.x

Angiosperm Phylogeny GroupChase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., et al. (2016). An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* 181, 1–20. doi: 10.1111/boj.12385

Aron, A. T., Gentry, E. C., McPhail, K. L., Nothias, L.-F., Nothias-Esposito, M., Bouslimani, A., et al. (2020). Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* 15, 1954–1991. doi: 10.1038/s41596-020-0317-5

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M.International Natural Product Sciences Taskforce, and Supuran, C. T. (2021). Natural products in drug discovery: advances and opportunities. *Nat. Rev. Drug Discov.* 20, 200–216. doi: 10.1038/s41573-020-00114-z

Barrère, C., Hubert-Roux, M., Afonso, C., Rejaibi, M., Kebir, N., Désilles, N., et al. (2014). Tandem mass spectrometry of low solubility polyamides. *Anal. Chim. Acta* 808, 3–9. doi: 10.1016/j.aca.2013.03.064

Bauermeister, A., Mannochio-Russo, H., Costa-Lotufo, L. V., Jarmusch, A. K., and Dorrestein, P. C. (2022). Mass spectrometry-based metabolomics in microbiome investigations. *Nat. Rev. Microbiol.* 20, 143–160. doi: 10.1038/s41579-021-00621-9

Beaulieu, W. T., Panaccione, D. G., Quach, Q. N., Smoot, K. L., and Clay, K. (2021). Diversification of ergot alkaloids and heritable fungal symbionts in morning glories. *Commun. Biol.* 4:1362. doi: 10.1038/s42003-021-02870-z

Belwal, T., Devkota, H. P., Hassan, H. A., Ahluwalia, S., Ramadan, M. F., Mocan, A., et al. (2018). Phytopharmacology of Acerola (*Malpighia* spp.) and its potential as functional food. *Trends Food Sci. Technol.* 74, 99–106. doi: 10.1016/j.tifs.2018.01.014

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Bondoc, K. G. V., Lee, H., Cruz, L. J., Lebrilla, C. B., and Juinio-Meñez, M. A. (2013). Chemical fingerprinting and phylogenetic mapping of saponin congeners from three tropical holothurian sea cucumbers. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* 166, 182–193. doi: 10.1016/j.cbpb.2013.09.002

Bueno, P. C. P., Abarca, L. F. S., Anhesine, N. B., Giffoni, M. S., Pereira, F. M. V., Torres, R. B., et al. (2021). Infraspecific chemical variability and biological activity of *Casearia sylvestris* from different Brazilian biomes. *Planta Med.* 87, 148–159. doi: 10.1055/a-1301-0183

Buer, C. S., Imin, N., and Djordjevic, M. A. (2010). Flavonoids: new roles for old molecules. *J. Integr. Plant Biol.* 52, 98–111. doi: 10.1111/j.1744-7909.2010.00905.x

Cai, L., Xi, Z., Amorim, A. M., Sugumaran, M., Rest, J. S., Liu, L., et al. (2019). Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. *New Phytol.* 221, 565–576. doi: 10.1111/nph.15357

Cameron, K. M., Chase, M. W., Anderson, W. R., and Hills, H. G. (2001). Molecular systematics of Malpighiaceae: evidence from plastid rbcL and matK sequences. *Am. J. Bot.* 88, 1847–1862. doi: 10.2307/3558361

Cantrell, K., Fedarko, M. W., Rahman, G., McDonald, D., Yang, Y., Zaw, T., et al. (2021). EMPress enables tree-guided, interactive, and exploratory analyses of multi-omic data sets. *mSystems* 6, e01216–e01220. doi: 10.1128/mSystems.01216-20

Cao, S., Guza, R. C., Miller, J. S., Andriantsiferana, R., Rasamison, V. E., and Kingston, D. G. I. (2004). Cytotoxic triterpenoids from *Acridocarpus vivy* from the Madagascar rain forest. *J. Nat. Prod.* 67, 986–989. doi: 10.1021/np040058h

Cardoso Taketa, A. T., Lozada-Lechuga, J., Fragoso-Serrano, M., Villarreal, M. L., and Pereda-Miranda, R. (2004). Isolation of nor-secofriedelanes from the sedative extracts of *Galphimia glauca*. *J. Nat. Prod.* 67, 644–649. doi: 10.1021/np0304666

Chen, Y., Mulder, P. P. J., Schaap, O., Memelink, J., Klinkhamer, P. G. L., and Vrieling, K. (2020). The evolution of pyrrolizidine alkaloid diversity among and within Jacobaea species. *J. Syst. Evol.* 60, 361–376. doi: 10.1111/jse.12671

Coley, P. D., Endara, M.-J., Ghabash, G., Kidner, C. A., Nicholls, J. A., Pennington, R. T., et al. (2019). Macroevolutionary patterns in overexpression of tyrosine: an anti-herbivore defence in a speciose tropical tree genus, *Inga* (Fabaceae). *J. Ecol.* 107, 1620–1632. doi: 10.1111/1365-2745.13208

Creydt, M., and Fischer, M. (2017). Plant metabolomics: maximizing metabolome coverage by optimizing mobile phase additives for nontargeted mass spectrometry in positive and negative electrospray ionization mode. *Anal. Chem.* 89, 10474–10486. doi: 10.1021/acs.analchem.7b02592

Crüsemann, M., O'Neill, E. C., Larson, C. B., Melnik, A. V., Floros, D. J., da Silva, R. R., et al. (2017). Prioritizing natural product diversity in a collection of 146 bacterial strains based on growth and extraction protocols. *J. Nat. Prod.* 80, 588–597. doi: 10.1021/acs.jnatprod.6b00722

Davioud, E., Bailleul, F., Delaveau, P., and Jacquemin, H. (1985). Iridoids of guyanese species of *Stigmaphyllon*. *Planta Med.* 51:78. doi: 10.1055/s-2007-969406

Davis, C. C., and Anderson, W. R. (2010). A complete generic phylogeny of Malpighiaceae inferred from nucleotide sequence data and morphology. *Am. J. Bot.* 97, 2031–2048. doi: 10.3732/ajb.1000146

Davis, C. C., Anderson, W. R., and Donoghue, M. J. (2001). Phylogeny of Malpighiaceae: evidence from chloroplast ndhF and trnl-F nucleotide sequences. *Am. J. Bot.* 88, 1830–1846. doi: 10.2307/3558360

Davis, C. C., Marinho, L. C., and Amorim, A. M. (2020). *Andersoniodoxa*, a replacement name for *Andersoniella* (Malpighiaceae). *Phytotaxa* 470, 121–122. doi: 10.11646/phytotaxa.470.1.9

Davis, C. C., Schaefer, H., Xi, Z., Baum, D. A., Donoghue, M. J., and Harmon, L. J. (2014). Long-term morphological stasis maintained by a plant-pollinator mutualism. *Proc. Natl. Acad. Sci. U. S. A.* 111, 5914–5919. doi: 10.1073/pnas.1403157111

de Almeida, R. F., Amorim, A. M., da Silva Corrêa, A. M., and Van den Berg, C. (2017). A new infrageneric classification for *Amorimia* (Malpighiaceae) based on morphological, phytochemical and molecular evidence. *Phytotaxa* 313, 231–248. doi: 10.11646/phytotaxa.313.3.1

de Almeida, R. F., Amorim, A. M. A., and van den Berg, C. (2018). Timing the origin and past connections between Andean and Atlantic seasonally dry tropical forests in South America: insights from the biogeographical history of *Amorimia* (Malpighiaceae). *Taxon* 67, 739–751. doi: 10.12705/674.4

de Almeida, R. F., Francener, A., Pessoa, C., Sebastiani, R., Oliveira, Y. R., Amorim, A. M. A., et al. (2021). Malpighiaceae. Flora do Brasil 2020 em construção. Available at: http://floradobrasil.jbrj.gov.br/reflora/floradobrasil/FB155 (Accessed January 13, 2022).

de Almeida, R. F., and van den Berg, C. (2020). Biogeography of stigmaphyllon (Malpighiaceae) and a meta-analysis of vascular plant lineages diversified in the Brazilian Atlantic rainforests point to the late Eocene origins of this megadiverse biome. *Plan. Theory* 9:1569. doi: 10.3390/plants9111569

de Almeida, R. F., and van den Berg, C. (2021). Molecular phylogeny and character mapping support generic adjustments in the Tetrapteroid clade (Malpighiaceae). *Nord. J. Bot.* 39:2876. doi: 10.1111/njb.02876

de Pinna, M. C. C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7, 367–394. doi: 10.1111/j.1096-0031.1991.tb00045.x

De Vos, R. C. H., Moco, S., Lommen, A., Keurentjes, J. J. B., Bino, R. J., and Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 2, 778–791. doi: 10.1038/nprot.2007.95

Demarque, D. P., Crotti, A. E. M., Vessecchi, R., Lopes, J. L. C., and Lopes, N. P. (2016). Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Nat. Prod. Rep.* 33, 432–455. doi: 10.1039/c5np00073d

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., et al. (2019). SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* 16, 299–302. doi: 10.1038/s41592-019-0344-8

Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., et al. (2021). Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* 39, 462–471. doi: 10.1038/s41587-020-0740-8

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* 112, 12580–12585. doi: 10.1073/pnas.1509788112

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Edler, D., Klein, J., Antonelli, A., and Silvestro, D. (2021). raxmlGUI 2.0: a graphical interface and toolkit for phylogenetic analyses using RAxML. *Methods Ecol. Evol.* 12, 373–377. doi: 10.1111/2041-210X.13512

Ernst, M., Nothias, L.-F., van der Hooft, J. J. J., Silva, R. R., Saslis-Lagoudakis, C. H., Grace, O. M., et al. (2019). Assessing specialized metabolite diversity in the cosmopolitan plant genus *Euphorbia* L. *Front. Plant Sci.* 10:846. doi: 10.3389/fpls.2019.00846

Floros, D. J., Jensen, P. R., Dorrestein, P. C., and Koyama, N. (2016). A metabolomics guided exploration of marine natural product chemical space. *Metabolomics* 12:145. doi: 10.1007/s11306-016-1087-5

Fraige, K., Dametto, A. C., Zeraik, M. L., de Freitas, L., Saraiva, A. C., Medeiros, A. I., et al. (2018). Dereplication by HPLC-DAD-ESI-MS/MS and screening for biological activities of *Byrsonima* species (Malpighiaceae). *Phytochem. Anal.* 29, 196–204. doi: 10.1002/pca.2734

Gallon, M. E., Monge, M., Casoti, R., Da Costa, F. B., Semir, J., and Gobbo-Neto, L. (2018). Metabolomic analysis applied to chemosystematics and evolution of megadiverse Brazilian *Vernonieae* (Asteraceae). *Phytochemistry* 150, 93–105. doi: 10.1016/j.phytochem.2018.03.007

Gemperline, E., Keller, C., and Li, L. (2016). Mass spectrometry in plant-omics. *Anal. Chem.* 88, 3422–3434. doi: 10.1021/acs.analchem.5b02938

Gong, Y.-X., Jiang, N., Zhang, Z.-R., Tian, Y.-H., and Yu, W.-B. (2020). Complete plastome sequence of *Aspidopterys obcordata* (Malpighiaceae), a traditional Dai medicinal plant from Xishuangbanna, Yunnan, China. *Mitochondrial DNA B Resour.* 5, 3533–3535. doi: 10.1080/23802359.2020.1827998

Huerta-Reyes, M., Herrera-Ruiz, M., González-Cortazar, M., Zamilpa, A., León, E., Reyes-Chilpa, R., et al. (2013). Neuropharmacological in vivo effects and phytochemical profile of the extract from the aerial parts of *Heteropterys brachiata* (L.) DC. (Malpighiaceae). *J. Ethnopharmacol.* 146, 311–317. doi: 10.1016/j.jep.2012.12.049

Isah, T. (2019). Stress and defense responses in plant secondary metabolites production. *Biol. Res.* 52:39. doi: 10.1186/s40659-019-0246-3

Ivanova, P. T., Milne, S. B., Byrne, M. O., Xiang, Y., and Brown, H. A. (2007). Glycerophospholipid identification and quantitation by electrospray ionization mass spectrometry. *Methods Enzymol.* 432, 21–57. doi: 10.1016/S0076-6879(07)32002-8

Jo, S., Kim, Y.-K., Cheon, S.-H., and Kim, K.-J. (2019). The complete plastome sequence from the family Malpighiaceae, *Bunchosia argentea* (Jacq.) DC. *Mitochondrial DNA B Resour.* 4, 1027–1029. doi: 10.1080/23802359.2019.1584065

Kang, K. B., Ernst, M., Hooft, J. J. J., Silva, R. R., Park, J., Medema, M. H., et al. (2019). Comprehensive mass spectrometry-guided phenotyping of plant specialized metabolites reveals metabolic diversity in the cosmopolitan plant family Rhamnaceae. *Plant J.* 98, 1134–1144. doi: 10.1111/tpj.14292

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–1649. doi: 10.1093/bioinformatics/bts199

Lavrynenko, O., Nedielkov, R., Möller, H. M., and Shevchenko, A. (2013). Girard derivatization for LC-MS/MS profiling of endogenous ecdysteroids in *Drosophila*. *J. Lipid Res.* 54, 2265–2272. doi: 10.1194/jlr.D035949

Lee, S. T., Cook, D., Pfister, J. A., Allen, J. G., Colegate, S. M., Riet-Correa, F., et al. (2014). Monofluoroacetate-containing plants that are potentially toxic to livestock. *J. Agric. Food Chem.* 62, 7345–7354. doi: 10.1021/jf500563h

Li, D., and Gaquerel, E. (2021). Next-generation mass spectrometry metabolomics revives the functional analysis of plant metabolic diversity. *Annu. Rev. Plant Biol.* 72, 867–891. doi: 10.1146/annurev-arplant-071720-114836

Ludwig, M., Nothias, L.-F., Dührkop, K., Koester, I., Fleischauer, M., Hoffmann, M. A., et al. (2020). Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nat. Mach. Intell.* 2, 629–641. doi: 10.1038/s42256-020-00234-6

Lumbsch, H. T., Schmitt, I., Barker, D., and Pagel, M. (2006). Evolution of micromorphological and chemical characters in the lichen-forming fungal family Pertusariaceae. *Biol. J. Linn. Soc. Lond.* 89, 615–626. doi: 10.1111/j.1095-8312.2006.00696.x

Maddison, W. P., and Maddison, D. R. (2006). Mesquite: a modular system for evolutionary analysis. Available at: https://www.mesquiteproject.org/ (Accessed March 30, 2020).

Mamode Cassim, A., Gouguet, P, Gronnier, J., Laurent, N., Germain, V., Grison, M., et al. (2019). Plant lipids: key players of plasma membrane organization and function. *Prog. Lipid Res.* 73, 1–27. doi: 10.1016/j.plipres.2018.11.002

Mannochio Russo, H., Ferreira Queiroz, E., Marcourt, L., Rutz, A., Allard, P.-M., de Almeida, R. F., et al. (2020). Phytochemical analysis of the methanolic leaves extract of *Niedenzuella multiglandulosa* (Malpighiaceae), a plant species toxic to cattle in Brazil. *Phytochem. Lett.* 37, 10–16. doi: 10.1016/j.phytol.2020.02.005

Mannochio-Russo, H., Bueno, P. C. P., Bauermeister, A., de Almeida, R. F., Dorrestein, P. C., Cavalheiro, A. J., et al. (2020). Can statistical evaluation tools for chromatographic method development assist in the natural products workflow? A case study on selected species of the plant family Malpighiaceae. *J. Nat. Prod.* 83, 3239–3249. doi: 10.1021/acs.jnatprod.0c00495

Martucci, M. E. P., Loeuille, B., Pirani, J. R., and Gobbo-Neto, L. (2018). Comprehensive untargeted metabolomics of Lychnnophorinae subtribe (Asteraceae: Vernonieae) in a phylogenetic context. *PLoS One* 13:e0190104. doi: 10.1371/journal.pone.0190104

Medema, M. H. (2021). The year 2020 in natural product bioinformatics: an overview of the latest tools and databases. *Nat. Prod. Rep.* 38, 301–306. doi: 10.1039/d0np00090f

Menezes, A. P. A., Resende-Moreira, L. C., Buzatti, R. S. O., Nazareno, A. G., Carlsen, M., Lobo, F. P., et al. (2018). Chloroplast genomes of *Byrsonima* species (Malpighiaceae): comparative analysis and screening of high divergence sequences. *Sci. Rep.* 8:2210. doi: 10.1038/s41598-018-20189-4

Mohammadi, S. A., and Prasanna, B. M. (2003). Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.2135/cropsci2003.1235

Neri-Numa, I. A., Soriano Sancho, R. A., Pereira, A. P. A., and Pastore, G. M. (2018). Small Brazilian wild fruits: nutrients, bioactive compounds, health-promotion properties and commercial interest. *Food Res. Int.* 103, 345–360. doi: 10.1016/j.foodres.2017.10.053

Newman, D. J., and Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.* 83, 770–803. doi: 10.1021/acs.jnatprod.9b01285

Nixon, K. C. (1999). Winclada 1.0. Available at: http://www.diversityoflife.org/winclada/ (Accessed March 30, 2020).

Nothias, L.-F., Petras, D., Schmid, R., Dührkop, K., Rainer, J., Sarvepalli, A., et al. (2020). Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* 17, 905–908. doi: 10.1038/s41592-020-0933-6

Paula-Freire, L. I. G., Mendes, F. R., Molska, G. R., Duarte-Almeida, J. M., and Carlini, E. A. (2013). Comparison of the chemical composition and biological effects of the roots, branches and leaves of *Heteropterys tomentosa* A Juss. *J. Ethnopharmacol.* 145, 647–652. doi: 10.1016/j.jep.2012.12.004

Pilon, A., Selegato, D., Fernandes, R., Bueno, P., Pinho, D., Carnevale Neto, F., et al. (2020). Metabolômica de plantas: métodos e desafios. *Quim Nova.* 43, 329–354. doi: 10.21577/0100-4042.20170499

Possobom, C. C. F., Guimarães, E., and Machado, S. R. (2010). Leaf glands act as nectaries in *Diplopterys pubipetala* (Malpighiaceae). *Plant Biol.* 12, 863–870. doi: 10.1111/j.1438-8677.2009.00304.x

Qing, Z., Xu, Y., Yu, L., Liu, J., Huang, X., Tang, Z., et al. (2020). Investigation of fragmentation behaviours of isoquinoline alkaloids by mass spectrometry combined with computational chemistry. *Sci. Rep.* 10:733. doi: 10.1038/s41598-019-57406-7

Queiroz, M. M. F., Queiroz, E. F., Zeraik, M. L., Ebrahimi, S. N., Marcourt, L., Cuendet, M., et al. (2014). Chemical composition of the bark of *Tetrapterys* 36ucronate and identification of acetylcholinesterase inhibitory constituents. *J. Nat. Prod.* 77, 650–656. doi: 10.1021/np401003p

Ramachandran, P., Zhang, N., McLaughlin, W. B., Luo, Y., Handy, S., Duke, J. A., et al. (2018). Sequencing the vine of the soul: full chloroplast genome sequence of *Banisteriopsis caapi*. *Genome Announc.* 6, e00203–e00218. doi: 10.1128/genomeA.00203-18

Rambaut, A., Suchard, M. A., Xie, D., and Drummond, A. J. (2014). Tracer v1.6. Available at: http://beast.bio.ed.ac.uk/Tracer (Accessed March 30, 2020).

Renner, S. S., and Schaefer, H. (2010). The evolution and loss of oil-offering flowers: new insights from dated phylogenies for angiosperms and bees. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 365, 423–435. doi: 10.1098/rstb.2009.0229

Riet-Correa, F., Medeiros, R. M. T., and Schild, A. L. (2012). A review of poisonous plants that cause reproductive failure and malformations in the ruminants of Brazil. *J. Appl. Toxicol.* 32, 245–254. doi: 10.1002/jat.1754

Ronquist, F., and Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574. doi: 10.1093/bioinformatics/btg180

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Graham, J., et al. (2021). Open natural products research: Curation and dissemination of biological occurrences of chemical structures through Wikidata. bioRxiv [Preprint]. doi: 10.1101/2021.02.28.433265

Sainty, D., Bailleul, F., Delaveau, P., and Jacquemin, H. (1981). Malpighiacees: Nouvelle Famille a Iridoides Etude du *Stigmaphyllon sagittatum*. *J. Nat. Prod.* 44, 576–578. doi: 10.1021/np50017a012

Samoylenko, V., Rahman, M. M., Tekwani, B. L., Tripathi, L. M., Wang, Y.-H., Khan, S. I., et al. (2010). *Banisteriopsis caapi*, a unique combination of MAO inhibitory and antioxidative constituents for the activities relevant to neurodegenerative disorders and Parkinson's disease. *J. Ethnopharmacol.* 127, 357–367. doi: 10.1016/j.jep.2009.10.030

Schmidt, J., Raith, K., Boettcher, C., and Zenk, M. H. (2005). Analysis of benzylisoquinoline-type alkaloids by electrospray tandem mass spectrometry and atmospheric pressure photoionization. *Eur. J. Mass Spectrom.* 11, 325–333. doi: 10.1255/ejms.745

Schmitt, I., and Barker, F. K. (2009). Phylogenetic methods in natural product research. *Nat. Prod. Rep.* 26, 1585–1602. doi: 10.1039/b910458p

Sereno, P. C. (2007). Logical basis for morphological characters in phylogenetics. *Cladistics* 23, 565–587. doi: 10.1111/j.1096-0031.2007.00161.x

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Steckel, A., and Schlosser, G. (2019). An organic chemist's guide to electrospray mass spectrometric structure elucidation. *Molecules* 24:611. doi: 10.3390/molecules24030611

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., et al. (2007). Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI). *Metabolomics* 3, 211–221. doi: 10.1007/s11306-007-0082-2

Tripathi, A., Vázquez-Baeza, Y., Gauglitz, J. M., Wang, M., Dührkop, K., Nothias-Esposito, M., et al. (2021). Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat. Chem. Biol.* 17, 146–151. doi: 10.1038/s41589-020-00677-3

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. *GigaScience* 2:16. doi: 10.1186/2047-217x-2-16

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* 34, 828–837. doi: 10.1038/nbt.3597

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry* 64, 3–19. doi: 10.1016/S0031-9422(03)00300-5

Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H., and Bertrand, S. (2019). 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. doi: 10.1021/acs.analchem.8b05112

Wu, Q., Yuan, Q., Liu, E.-H., Qi, L.-W., Bi, Z.-M., and Li, P. (2010). Fragmentation study of iridoid glycosides and phenylpropanoid glycosides in radix Scrophulariae by rapid resolution liquid chromatography with diode-array detection and electrospray ionization time-of-flight mass spectrometry. *Biomed. Chromatogr.* 24, 808–819. doi: 10.1002/bmc.1368

Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., et al. (2012). Phylogenomics and a posteriori data partitioning resolve the cretaceous angiosperm radiation Malpighiales. *Proc. Natl. Acad. Sci. U. S. A.* 109, 17519–17524. doi: 10.1073/pnas.1205818109

Yang, J. Y., Sanchez, L. M., Rath, C. M., Liu, X., Boudreau, P. D., Bruns, N., et al. (2013). Molecular networking as a dereplication strategy. *J. Nat. Prod.* 76, 1686–1699. doi: 10.1021/np400413s

Zanatta, A. C., Vilegas, W., and Edrada-Ebel, R. (2021). UHPLC-(ESI)-HRMS and NMR-based metabolomics approach to access the seasonality of *Byrsonima intermedia* and *Serjania marginata* From Brazilian Cerrado Flora diversity. *Front. Chem.* 9:710025. doi: 10.3389/fchem.2021.710025