Journal of Chromatography Open 2 (2022) 100071

Contents lists available at ScienceDirect

Journal of Chromatography Open

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

Identifying foliar volatile organic compounds of *Plectranthus* and *Coleus* (Lamiaceae) as predictive markers of genus using GC×GC-TOFMS and machine learning

Daniel T. Pretorius, Egmont Rohwer, Yvette Naudé*

Department of Chemistry, University of Pretoria, Hatfield 0002, South Africa

ARTICLE INFO

Keywords:
Plectranthus and coleus
Volatile markers
Chemotaxonomy
Metabolomics
Gc×gc-tofms
Machine learning

ABSTRACT

Two prominent plant genera, *Plectranthus*, and *Coleus*, many species of which are indigenous to southern Africa, have been previously classified as a single genus of the name *Plectranthus*. However, phylogenetic analysis of markers of the plastid genome of subtribe *Plectranthinae* (family: Lamiaceae) has led to the recognition of *Coleus* as a sister taxon to *Plectranthus*. The purpose of this study is to analyse the profiles of foliar volatile organic compounds (VOCs), from the leaves of southern African species of *Plectranthus* and *Coleus*, as predictive chemotaxonomic markers of genus, using two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS) and machine learning. Foliar VOCs from fresh crushed leaves of representative species of each genera (nine *Plectranthus*, six *Coleus*) were extracted in triplicate using static headspace solid-phase microextraction (HS-SPME), and analysed using GC×GC-TOFMS. Profiles of the foliar VOCs for each sample were constructed from their total ion chromatograms, and machine learning algorithms were used to model the data, to make predictions on the genus of new samples, and to tentatively identify putative markers of genus for *Plectranthus* and *Coleus*. A high predictive accuracy (up to 90%) was obtained, with a sensitivity (for genus *Coleus*) of up to 100%. Top ranking variables included C_6–C_{15} compounds of various chemical classes, but most notably of the sesquiterpene isomers, which were found to occur more prevalently in genus *Coleus*.

1. Introduction

The genera *Plectranthus* L'Hérit. and *Coleus* Lour. are part of family Lamiaceae, a large and diverse family of angiosperms which includes many commonly known herbs and shrubs of horticultural and phytochemical significance [1–4]. Within Lamiaceae is tribe Ocimeae, and the largest subtribe within Ocimeae is *Plectranthinae* [5,6]. The latter contains eleven genera, including *Plectranthus* and *Coleus*. Most species of *Plectranthus* and *Coleus* are herbaceous perennials indigenous to paleotropical regions of the globe [3,4].

The taxonomical relationship of the genus *Coleus* in relation to *Plectranthus* is a matter that only recently, in light of genomic evidence, has seen overall consensus [1,2]. In the late eighteenth century, the first descriptions of specimens of *Plectranthus* and *Coleus* were published independently [2], and on the basis of observations on stamen morphology, the two genera were classified separately [2]. However, in 1962, a revised analysis of stamen morphology proposed subsuming genus *Coleus* under *Plectranthus* [2,7]. This reclassification was not met with broad consensus [2,8-13]. The recent phylogenetic data from a number of markers of the plastid genome of subtribe *Plectranthinae*, has led to a more definitive phylogeny in which the genus *Coleus* (along with four other genera) is delimited from *Plectranthus* (along with five other genera) as its own clade [1,2,14].

More recently, chemotaxonomical evidence consistent with the new phylogeny has been brought to light in the form of a variety of diterpenoids ($C_{20}H_{32}$), particularly a group of C-14-deoxy abietanes characteristic of genus *Coleus* [15].

Plants produce volatile organic compounds (VOCs) as secondary metabolites which may function as molecular signals mediating ecological interactions of the plant with other organisms in the environment [16–19].They can be obtained *in situ*, from particular organs (leaves, flowers, roots, or fruit), or *in vivo* from a whole plant [20,21]. The primary extraction methods include liquid-liquid extraction (LLE), steam distillation/hydrodistillation [22], supercritical fluid extraction (SFE) [22,23] and solid-phase extraction (SPE) [24]. The most widely utilised method, however, is headspace solid-phase microextraction (HS-SPME) [25–29]. The advantage of using polymeric sorbents and headspace sampling is that they exclude higher molecular weight species, thus limiting potential matrix interferents present in the sample. For this reason, HS-SPME is a more selective technique for the sampling of VOCs, and

* Corresponding author.
E-mail address: yvette.naude@up.ac.za (Y. Naudé).

https://doi.org/10.1016/j.jcoa.2022.100071
Received 30 September 2022; Received in revised form 24 November 2022; Accepted 24 November 2022

obviates the preparation and clean-up stages involved in solvent-based methods. The disadvantage of such sorbents is their lower affinity for polar molecules which may be of importance to the analysis.

Extracts of plant VOCs can be composed of a variety of trace chemical constituents, and analysis of such complex extracts, particularly at trace-level limits of detection, requires an instrumental method of high sensitivity and specificity, which can be achieved using gas chromatography-mass spectrometry (GC–MS) [20–22]. But although GC–MS is well suited to the analysis of complex samples, coelution is still a complicating factor. A more comprehensive analysis, with improved resolution of coeluting components, and thus broader analyte coverage, can be achieved by two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS).

This study is focussed on foliar VOCs, which are produced by leaves, as potential markers of genus in *Plectranthus* and *Coleus*. Though there have been numerous studies on the volatile composition of individual specimens of *Plectranthus* and *Coleus* [30–33], there appears currently to be no genus-wide data on the VOC metabolomes of members of these two genera— a gap which this study intends to address in a preliminary investigation, with the aim of determining whether a chemotaxonomic grouping of *Plectranthus* and *Coleus* is reflected on the level of the volatile metabolome. Beyond interest from a chemotaxonomical perspective, such information could harbour potential for future applications, for example, in the prospecting for compounds of phytochemical or ethnobotanical significance.

2. Methods and materials

2.1. Ethical considerations

This study was approved by the Ethics committee of the Faculty of Natural and Agricultural Sciences (reference: NAS256/2020) of the University of Pretoria.

2.2. Reagents and chemical standards

Methanol, acetone, acetonitrile, *n*-hexane and the solution of *n*-alkanes (C_8_C_{28}), used for the calculation of linear retention indices of selected analytes, were purchased from Merck, Pretoria, South Africa.

2.3. Sample population for foliar VOC sampling

The plants used in this study were indigenous southern African *Plectranthus* and *Coleus*, sourced from local nurseries and private gardens, in Gauteng, South Africa. All plants were potted, 24–48 h prior to sampling, in soil from the same batch, and were watered at the same time on their respective days of sampling.

Fifteen species of both genera were included in the study. Of genus *Plectranthus*, nine species were included in the study: *P. ambiguus* (Bolus) Codd, *P. chimanimanensis* S.Moore, *P. ecklonii* Benth., *P. fruticosus* L'Hér., *P. oertendahlii* T.C.E.Fr., *P. saccatus* Benth., *P. strigosus* Benth. ex E.Mey., *P. verticillatus* (L.f.) Druce, and *P. zuluensis* T.Cooke, Bull. Of genus *Coleus*, the six species included were: *C. hadiensis* (Forssk). A.J.Paton, *C. hereroensis* (Engl.) A.J.Paton, *C. livingstonei* A.J.Paton, *C. longipetiolatus* Gürke, *C. madagascariensis* (Pers.) A.Chev., and *C. neochilus* (Schltr.) Codd. Samples of each species were taken in triplicate. Although replicates were taken, each was treated as a discrete sample (*i.e.*, replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning (n = 45). Air blanks and a soil blank were taken to aid in accounting for air and soil VOC contaminants.

Species were identified using species descriptions [34] and herbarium specimens. Voucher specimens of the plants sampled were submitted for record to the H.G.W.J. Schweickerdt Herbarium (PRU) of the University of Pretoria (c.f.: Supplementary information).

2.4. HS-SPME of foliar VOCs

Fresh leaves were picked and removed of petioles, weighed to a mass of two grams, and crushed with a mortar and pestle. The crushed leaf matter was enclosed in a glass vial (40 mL) with a screw cap with a central hole (3.2 mm radius) lined with Teflon® septa (Separations, South Africa). The vial was left to stand in a water bath at 40˚C, for 15 min, to allow for equilibration of the system to occur. Headspace (HS) extraction of the foliar VOCs was performed using a SPME device with a fused-silica fibre coated with 100-micron-diameter polydimethylsiloxane (PDMS) (Supelco, Sigma-Aldrich®, Kempton Park, South Africa). The fibre was chosen on the basis of its selectivity for nonpolar compounds, particularly of the terpene variety, which were assumed to be the major foliar components. However, it should be noted that this choice will have discriminated against polar species of potential importance. The extraction time was 15 min, and the temperature of extraction 40˚C. Thermal desorption of sorbed analytes from the SPME fibre for instrumental analysis by GC×GC-TOFMS was performed in the GC inlet directly after sampling as described in 2.5. The fibre was conditioned at 280˚C in split mode (50:1) for 20 min prior to analysis. Triplicate air blanks and a soil blank (from the same soil in which all specimens were potted) were taken to aid in accounting for air and soil contaminants.

2.5. Instrumental and analytical methods: comprehensive gc×gc-tofms

Comprehensive two-dimensional chromatographic separation and mass spectrometric separation and analysis was performed on a LECO® Pegasus® 4D GC×GC-TOFMS with an Agilent® 7890A chromatograph and a dual quad-jet cryogenic modulator (LECO®, Kempton Park, South Africa), operated by ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®). The hot jets were operated with nitrogen gas produced by a nitrogen gas generator (Peak Scientific, South Africa), and the cold jets were operated with nitrogen gas cooled with liquid nitrogen (Afrox, South Africa). The primary (1D) column was an Rxi-1MS apolar capillary column of length 30 m, 250 μm ID and 0.25 μm film thickness; the secondary (2D) column was an Rxi-17SilMS mid-polar capillary column of length 0.760 m, 250 μm ID and 0.25 μm film thickness (Restek, Bellefonte, PA, USA).The carrier gas (ultra-high purity grade helium [Afrox, Gauteng, South Africa]) flow rate was constant at 1.4 mL/min. Thermal desorption of the sorbed analytes from the SPME fibre was done in a SPME inlet liner (straight design, unpacked, 78.5 mm (L) x 6.5 mm (OD) x 0.75 mm (ID) (Supelco, Merck, Kempton Park, South Africa) of the GC inlet at 230˚C. The splitless time was 30 s, the inlet purge flow rate was 30 mL/min, with an inlet septum purge flow rate of 3 mL/min. The fibre was removed from the inlet after 5 min.

The initial temperature for the primary oven was held at 40˚C for 1.5 min, and ramped to 280˚C at a rate of 6˚C/min, with a hold time at this temperature of 2.83 min. The total run time was 44.5 min. The secondary oven and the modulator temperatures were offset by +5˚C and +15˚C respectively to that of the primary oven. The transfer line to the TOFMS and the inlet of the chromatograph were maintained at a temperature of 280˚C and 230˚C respectively. The modulation period was 2 s, with a hot pulse time of 0.6 s and a cool time between stages of 0.4 s.

The TOFMS was operated at an acquisition rate of 100 spectra/s over a mass range of 35–500 Daltons. The ionisation energy was 70 eV in electron impact ionisation mode (EI+), the voltage of the detector was 1650 V and the temperature of the ion source was 230 ˚C.

2.6. Data acquisition and processing

ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®) was used for data acquisition and chromatographic peak alignment. A S/N threshold of 100 was set, and deviations in retention time were bound within the modulation period (2 s), for 1D peaks, and 0.1 s for 2D peaks. Tentative identification of the analyte compounds was achieved

by comparison of experimental mass spectra with reference spectra of the National Institute of Standards and Technology (NIST) library (version 2.2), with the minimum similarity threshold for a match set at 75%.

Retention indices of reported analyte compounds were calculated using the method of the linear temperature programmed retention index, as developed by Van den Dool and Kratz [35], as well as from a least-squares linear regression equation, using a series of *n*-alkanes (C_8–C_{28}; Merck, Pretoria, South Africa). Compounds having experimental RI values within ±35 units of the literature values are reported, with most falling within ±15 units.

For data processing prior to statistical analysis, contaminant compounds were removed from each individual sample data set, including organosiloxanes, halogens, boronic compounds and metallic complexes. Blank corrections were performed, for the soil-blank with respect to the air-blank (triplicate-averaged), and for each replicate with respect to the air- and soil-blanks. Compounds with resultant negative values were removed.

2.7. Machine learning (regression and classification)

Supervised machine learning was performed using R© computational and statistical software (version 1.3.959) with the Classification and Regression Training (caret) package (version 6.0–86) [36]. The method was based on that of Kuhn, 2008 [37]. Three algorithms were used to construct regression and classification models of the data: an elastic-net regression (using the glmnet algorithm), a random forest (ranger) and a support vector machine (svmPoly). These models are suited to high dimensional datasets, allow for ranked variable selection, and are not very sensitive to outliers and correlated predictors.

The dataset (after contaminant removal and blank correction) was shuffled and randomly split with a 0.5 ratio into a training and testing set. Prior to machine learning, the training dataset was pre-processed, using pre-processing functions of the caret package [36,37], in two steps: 1) The data was normalised, i.e.: centred and scaled; 2) variables that had zero or near-zero variance were removed from the dataset. Although replicates were taken, each was treated as a discrete sample (i.e.: replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning.

The parameters and coefficients of each model were computed and optimised with five-fold, five-times-repeated cross-validation. The function selects those model parameters with the highest AUC/ROC values (area under the curve of the receiver operating characteristic) as determined by cross-validation. Top variables were identified using the inbuilt functions available for the ranking of variable importance [36]. Peak area values of the top variables were normalised (i.e.: centred and scaled) and visualised as a heatmap, using the heatmap.2 function in the gplots package (version 3.1.3).

3. Results and discussion

3.1. Chromatographic data from gc×gc-tofms

Total ion chromatograms (TICs) of foliar compounds from the leaves of southern African *Plectranthus* and *Coleus* were obtained by GC×GC-TOFMS. Fig. 1 presents the contour plots of an extract of *C. neochilus*, which is representative of the chromatographic trend observed across the samples of both genera. There are two regions with prominent peaks— the first lies in the 1D retention time range of ±430–700 s, and the second in the range of ±1040–1300 s. The first region is populated by peaks of the monoterpenes ($C_{10}H_{16}$) and monoterpenoids, and the second region by peaks of the sesquiterpenes ($C_{15}H_{24}$) and sesquiterpenoids. The first region also consists of peaks from C_6–C_{10} species, which may be green leaf volatiles (GLVs) — VOCs which are released upon rupture of the foliar tissue, or with changes in temperature and light — which have been reported to consist of terpenes in addition to C_6 species [20,38,39]. No peaks with retention times greater than 1500 s are observed, indicating a low occurrence of VOCs of greater than fifteen carbon units. Figs. 1 and 2 (annotated contour plots of extracts of *C. neochilus* and *P. oertendahlii*) are presented as representative of their respective genus. In general (but not without exception) *Coleus* samples are observed to show a greater number of peaks.

It should be noted that the chromatographic profiles do not reflect absolute quantities, and that due to differences in compound partition

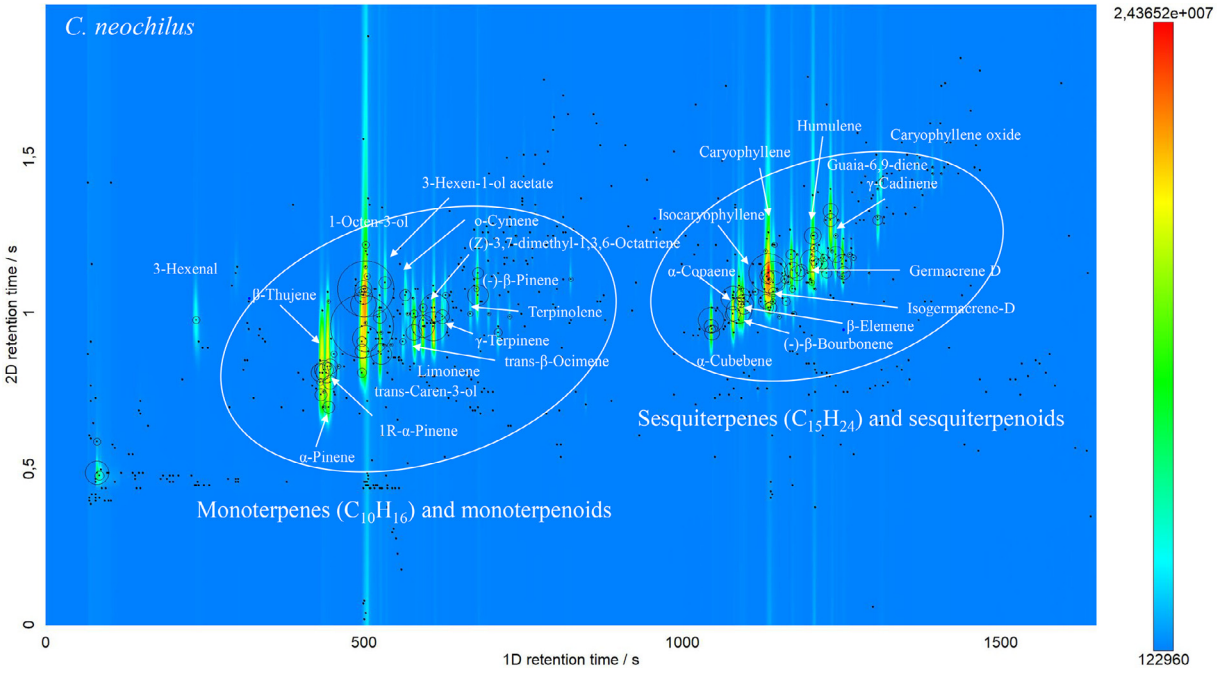

Fig. 1. TIC contour plot (with superimposed peak markers) of an extract of C. neochilus. Peaks of interest fall mostly within the retention time regions of 400–700 s (monoterpenes) and 1040–1300s (sesquiterpenes). Circles indicate peak size.

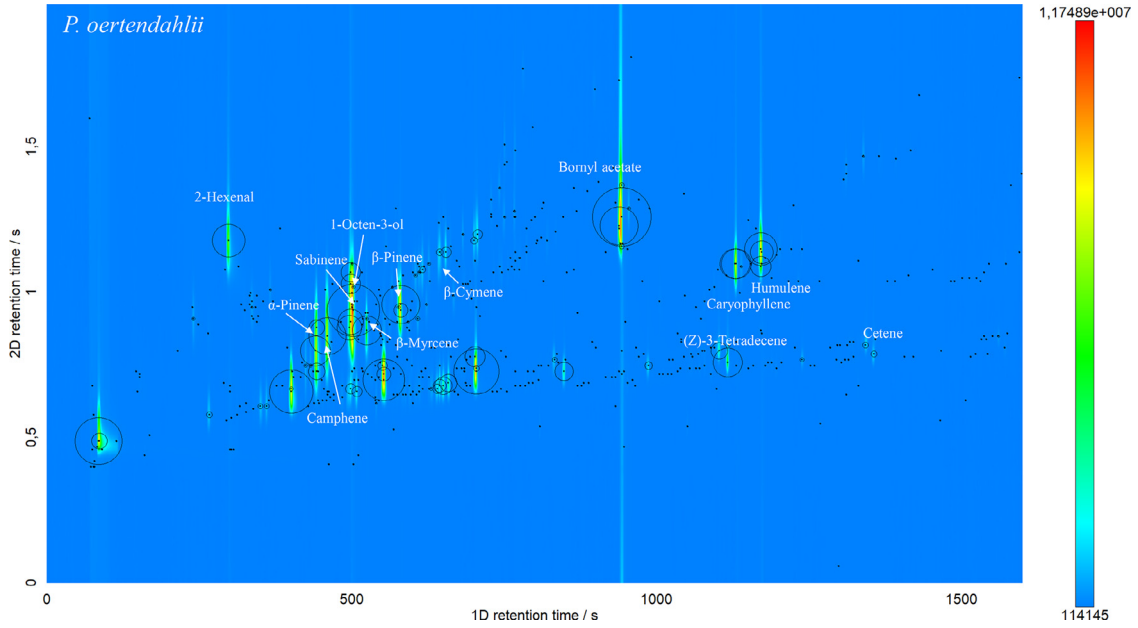

Fig. 2. TIC contour plot (with superimposed peak markers) of an extract of P. oertendahlii. Peaks of interest fall mostly within the retention time regions of 400–700 s (monoterpenes) and 1040–1300s (sesquiterpenes). Circles indicate peak size.

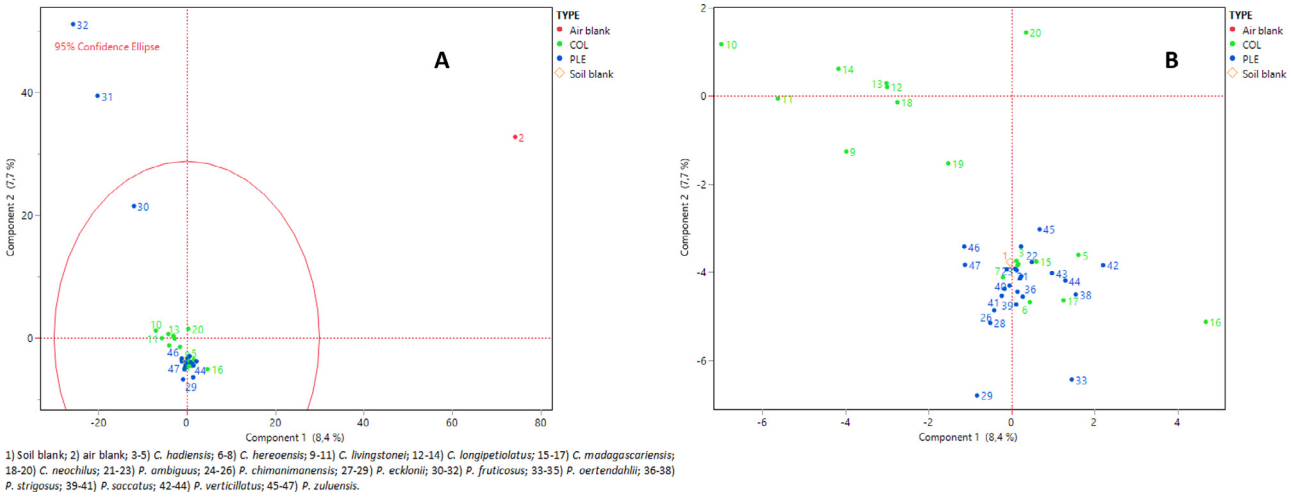

Fig. 3. (A) Score plot for the first two principal components of the blank (air and soil), *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with the 95% confidence ellipse shown. 1) Soil blank; 2) air blank; 3–5) *C. hadiensis*; 6–8) *C. hereoensis*; 9–11) *C. livingstonei*; 12–14) *C. longipetiolatus;* 15–17) *C. madagascariensis*; 18–20) *C. neochilus*; 21–23) *P. ambiguus*; 24–26) *P. chimanimanensis*; 27–29) *P. ecklonii*; 30–32) *P. fruticosus*; 33–35) *P. oertendahlii*; 36–38) *P. strigosus*; 39–41) *P. saccatus*; 42–44) *P. verticillatus*; 45–47) *P. zuluensis*. (B) Amplified view of the region around the origin.

coefficients, relative peak areas for certain compounds may be biased towards high or low values. However, such an effect is assumed to be consistent from sample to sample, permitting inter-sample comparison.

3.2. Preliminary statistical analysis: principal component analysis

As a preliminary assessment of variation in the foliar VOC profiles of members of *Plectranthus* and *Coleus*, principal component analysis (PCA) was performed on the full dataset, described by 1794 variables (compounds), prior to processing for supervised learning.

Fig. 3A is a PCA score plot for the first and second components, of the blank (air and soil), *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with the 95% confidence ellipse shown, and reveals the clustering of the samples in terms of the predictor loading scores. PCA decomposes the full dataset into 45 principal components capturing 99.95% of the variance. The first component, which accounts for the greatest variation, represents only 8.4% of the cumulative total, suggesting that the variation is not strongly influenced by a small number of predictors. The foliar samples, with the exception of the three replicates of *P. fruticosus*, cluster within ±5–7 units of the origin of the first component. The *P. fruticosus* replicates are widely distributed along both components, and appear to be outliers.

The point for the air-blank falls at high values of the first component (Fig. 3A), which indicates that the blank correction procedure for airborne contaminants was effective. The point corresponding to the soil blank, however, falls within the main cluster of samples. This is expected in light of the fact that soil components tend to accumulate inside leaves. It should be noted that the soil blank itself is corrected by the air blank, which leads to negative peak area values for the former. Since a negative peak area is not meaningful, these values were converted to zero during processing, which resulted in a value of zero for all predictors for the soil sample. For this reason, the point for the soil blank on the score plot

Table 1
List of the top predictor compounds for the elastic-net, random forest and support vector machine models.

Elastic-net		Random forest		Support-vector machine	
Predictor compound	Scaled score	Predictor compound	Scaled score	Predictor compound	Scaled score
Cyclohexanone, 2,2,6-trimethyl-	100.00	(E)−4-Oxohex-2-enal	100.00	Hexanal	100.00
β-Ylangene	90.00	Furan-2-ethyl	71.42	(E)−4-Oxohex-2-enal	90.65
Hexanal	72.97	Hexanal	68.29	*β*-Cubebene	86.92
(E)−4-Oxohex-2-enal	54.66	Ylangene	68.26	Ylangene	83.18
Octane, 1,1′-oxybis-	47.26	*α*-Cubebene	67.81	*β*-Copaene	81.31
1,2,4,4-tetramethylcyclopentene	35.56	2-Hexenal	64.77	Furan-2-ethyl	77.57
Furan-2-ethyl	22.41	3-Hexen-1-ol, acetate, (Z)-	62.61	*α*-Cubebene	75.70
1-Hexene	1.97	*α*-Cadinene	61.26	Isogermacrene D	70.09
		Isogermacrene D	58.85	trans-*β*-Ionone	70.09
		trans-*β*-Ionone	57.91	2,4-Hexadienal, (E,E)-	64.49
		γ-Cadinene	56.59	*γ*-Cadinene	64.49
		α-Selinene	54.97	3-Hexen-1-ol, (Z)-	62.62
		β-Cubebene	54.25	1,2,4,4-tetramethylcyclopentene	61.68
		(Z,E)-*α*-Farnesene	53.68	2(5H)-Furanone, 5-ethyl-	61.68
		β-Calacorene	52.43	*β*-Calacorene	60.75
		Decanal	51.24	3-Hexen-1-ol, acetate, (Z)-	60.75
		2(5H)-Furanone, 5-ethyl-	49.18	*α*-Cadinene	58.88
		Octane, 1,1′-oxybis-	48.49	Octane, 1,1′-oxybis-	57.01
		Bornyl acetate	47.42		

Table 2
Retention indices (RI) of selected top predictor compounds of genus *Plectranthus* and *Coleus*.

Tentative identification	CAS number	Molecular formula	Chemical class	MW (g/mol)	RI_{Exp} nonpolar	RI_{Lit} nonpolar (NIST)	MS similarity match
α-Cubebene	17,699–14–8	C15H24	Sesquiterpene	204	1408	1366; 1351	753–916
β-Cubebene	13,744–15–5	C15H24	Sesquiterpene	204	1415	1384; 1381	755–893
β Ylangene	20,479–06–5	C15H24	Sesquiterpene	204	1435	1418; 1425	759–895
β-Copaene	374,189 *NIST	C15H24	Sesquiterpene	204	1633	1598	777–903
α-Cadinene	24,406–05–1	C15H24	Sesquiterpene	204	1539	1522; 1534	794–922
γ-Cadinene	39,029–41–9	C15H254	Sesquiterpene	204	1519	1505, 1507	931–941
α-Selinene	473–13–2	C15H24	Sesquiterpene	204	1496	1523; 1500	754–930
Isogermacrene D	317,819–80–0	C15H24	Sesquiterpene	204	1444	1431; 1442	752–916
(Z,E)-α-Farnesene	26,560–14–5	C15H24	Sesquiterpene	204	1493	1486; 1477	763–929
β-Calacorene	50,277–34–4	C15H20	Bicyclic sesquiterpene	204	1541	1548; 1543	812–893
trans-β-Ionone	79–77–6	C13H20O	Ionone	192	1479	1463; 1462	762–859
Bornyl acetate	76–49–3	C12H20O2	Terpene derivative	196	1292	1269; 1270	785–917
2(5H)-Furanone, 5-ethyl-	2407–43–4	C6H8O2	Unsaturated lactone	112	1019	984; 963	753–877
2,4-Hexadienal, (E,E)-	142–83–6	C6H8O	Unsaturated aldehyde	96	870	877; 877	770–932
3-Hexen-1-ol, (Z)-	928–96–1	C6H12O	Unsaturated alcohol	100	824	872; 838	753–952
3-Hexen-1-ol, acetate, (Z)-	3681–71–8	C8H14O2	Unsaturated ester	142	981	987; 981	816–954
1,2,4,4-tetramethylcyclopentene	65,378–76–9	C9H16	Cyclic alkene	124	834	857,6; 856,5	794–872
Cyclohexanone, 2,2,6-trimethyl-	2408–37–9	C9H16	Cyclic ketone	140	998	1013; 1008	827–857
(E)−4-Oxohex-2-enal	374,042 *NIST	C6H8O2	Unsaturated ketoaldehyde	112	985	958; 950	751–880
Octane, 1,1′-oxybis-	629–82–3	C16H34O	Ether	242	1650	1657; 1660	869–927
Decanal	112–31–2	C10H20O	Aldehyde	156	1202	1183, 1184	824–832

*CAS number not available, NIST entry number.

falls close to the origin, in the region of low eigenvalues for the first two components, indicating that the variation observed in the PCA plot is not weighted strongly towards soil compounds present in the leaves.

An amplified view of the region near the origin of the score plot (Fig. 3B) shows species of both genera to cluster within about two units either side of the origin of the first components, and within about two units of the negative region of the second component. However, the replicates of *C. neochilus* (18–20), *C. livingstonei* (9–11), and *C. longipetiolatus* (12–14) are distributed along a greater distance, and at greater absolute values, of the first component, and are separated from the cluster of *Plectranthus* samples along both components. This suggests that these samples are characterised by high variance predictors (predictors with high loading scores).

3.3. Data pre-processing for machine learning

The full dataset of foliar VOC profiles consists of 45 observations (replicate samples) described by 1794 predictors, corresponding to 1794 compounds identified by mass-spectral similarity scoring (with a minimum similarity threshold of 75%) using reference spectra from the NIST database. Splitting of the data by a 0.5 ratio results in 22 samples in the training set and 23 samples in the testing set. The near-zero variance removal function of the pre-processing step, performed on the 22 samples in the training set, results in the removal of 1160 predictors, and the retention, centring and scaling of 634 predictors. This can be interpreted as a 65% reduction in the dimensionality of the dataset, and emphasises the importance of the pre-processing step in paring the dataset of uninformative variables.

3.4. Testing and prediction

The predictions of each model are summarised in a confusion matrix (c.f.: Supplementary information) with corresponding performance statistics. Genus *Coleus* is defined as the *positive* class, and genus *Plectranthus* the *negative* class. The no-information rate (NIR) is equivalent to the percentage of samples in the majority category (in this case *Plectranthus*), or in other words, the percentage of true negatives and false positives in the testing set. The NIR can be interpreted as the predictive accuracy obtained by a null model that classifies every sample as *negative/Plectranthus*. For each of the three models, the accuracy (the

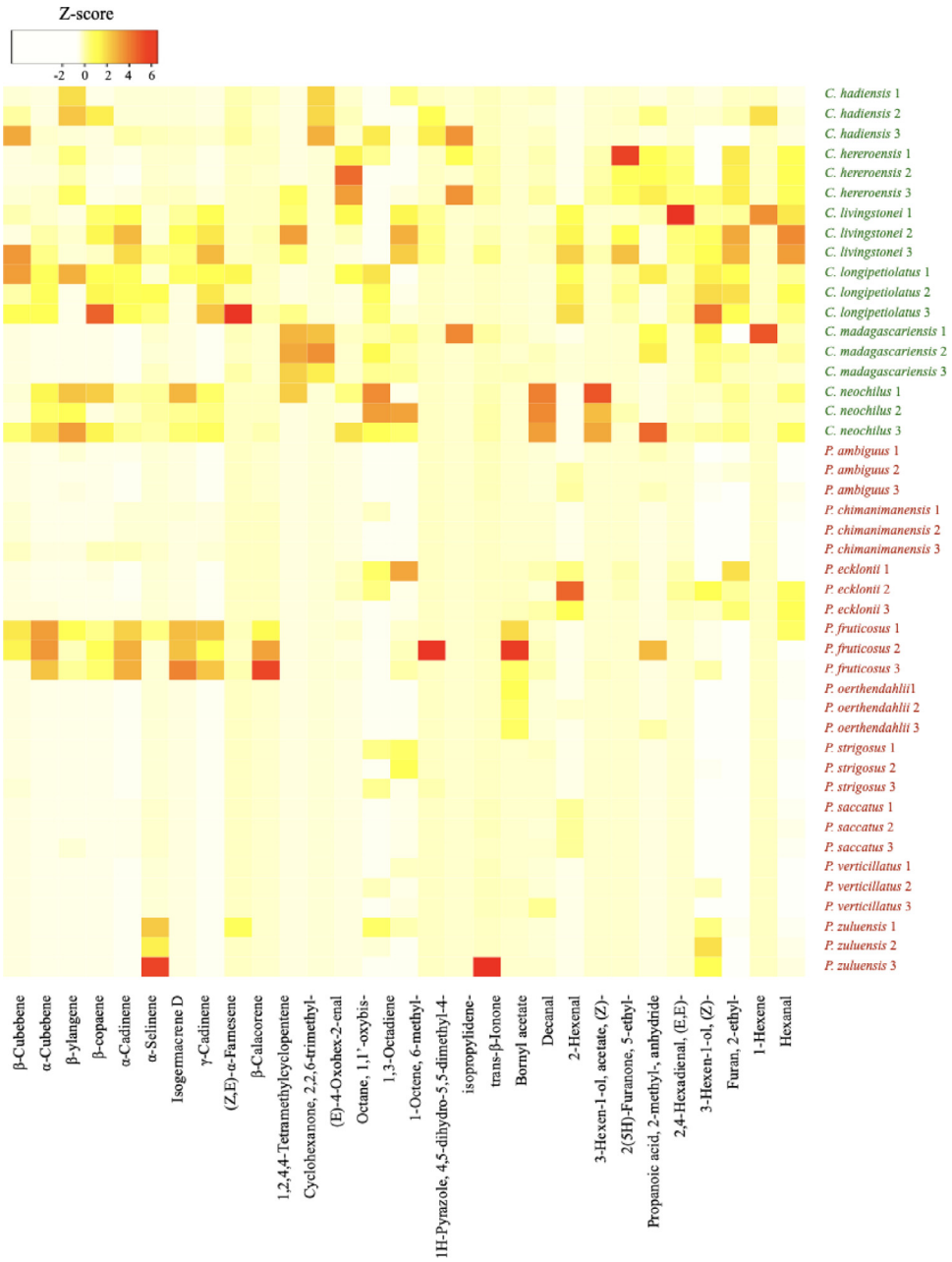

Fig. 4. Heatmap of species-wise relative abundance of the compounds ranked as top variables by machine learning. Green = *Coleus*; red = *Plectranthus*. The colour intensity scale depicts the relative abundance Z-score.

percentage of observations correctly classified) is high (more than 85%) and greater than the NIR (70%), and thus the models can be concluded to be predictive. However, it should be emphasised that testing was performed on a split of the dataset, and may thus be overfit to this set, even if the testing set was not used for model construction. The performance of the model may suffer if applied to a testing set sampled and analysed independently of the training set (for example, using different individual specimens for the training and testing groups, or collecting the samples for the training set one summer, and those for the second set the following summer).

Two key statistics are the sensitivity and the specificity [36]. Since *Coleus* is defined as positive, the sensitivity measures the ability of the model to correctly classify foliar VOC samples of species from genus *Coleus*. The support-vector machine demonstrates the highest sensitivity (100%), followed by the random forest (86%) and the elastic-net (71%). The high sensitivity of the three models tested means that they each also have a high positive prediction value (PPV). The PPV accounts for the sensitivity, the prevalence (the proportion of positive/*Coleus* samples) and the false positive rate [36].

The specificity is the proportion of the true negatives of all the cases predicted to be negative, and in this case measures the ability of the model to correctly classify samples from genus *Plectranthus*. The random forest has the highest specificity (100%). Since the specificity for each model is high, the negative prediction value (NPV) is significant for each model. The NPV accounts for the specificity, the false negative rate, and the proportion of negative/*Plectranthus* samples in the testing set [36].

3.5. Variable importance ranking

The machine learning algorithms employed are able to identify key predictors that distinguish the categorical types in question. The algorithms of the caret package have inbuilt functions available for the ranking of variable importance [36]. Table 1 lists the top-ranking predictors/compounds for the three models and their scaled scores. A few

of the compounds are common to more than one of the output lists in Table 1, although their specific ranks and scores differ. A number of the top variables are molecules belonging to the isomeric class of sesquiterpenes, of molecular formula $C_{15}H_{24}$. These tentatively identified species include β-ylangene, α-cubebene, β-cubebene, β-copaene, α-cadinene, γ-cadinene, isogermacrene D, α-selinene and (Z,E)-α-farnesene. A bicyclic sesquiterpene, β-calacorene ($C_{15}H_{20}$), is also ranked as a top predictor.

The other high-ranking compounds are of diverse organic classes of lower molecular weight than the sesquiterpenes. This includes a group of C_6 and C_8 compounds, including furan, 2-ethyl; 1-hexene; hexanal; 2-hexenal; (E,E)−2,4-Hexadienal; (Z)−3-hexen-1-ol; and (Z)−3-hexen-1-ol, acetate, likely to be GLVs [20,38,39]. Also included are similar compounds such as the unsaturated ketoaldehyde, (E)−4-oxohex-2-enal, and cyclohexanone, 2,2,6-trimethyl-, which are also likely to be GLVs. Notably, there are no monoterpenes or monoterpenoids with high variable importance scores.

The top compound for the elastic-net regression is the cyclic ketone, cyclohexanone, 2,2,6-trimethyl-, for the random forest the unsaturated aldehyde, (E)−4-oxohex-2-enal, and for the support-vector machine the aldehyde hexanal (Table 1). The two latter aldehydes are within the top four ranks for all three models. Overall, the sesquiterpene compounds do not feature on the output list of the elastic-net regression, which moreover, is most different from the other two algorithms in terms of its variable output. However, ylangene has the second highest score for the elastic-net (90), and the fourth highest scores for the random forest and support-vector machine (83 and 68, respectively).

Note that the outputs differ between the elastic-net, on the one hand, and the random forest and support-vector machine, on the other. For the former, the magnitudes of the regression coefficients are used to select the top variables, whereas for the latter, AUR/ROC values are computed for each variable using in-training splits, and those with the highest values are selected [36].

3.6. Retention indices of top-ranking compounds

Retention indices (RIs) for selected top-ranking compounds are reported in Table 2, along with the unique CAS number[1] for the tentative identification and the mass spectral similarity to reference spectra from the NIST database. The retention times for furan-2-ethyl; 1-hexene; 2-hexenal and hexanal, fall outside of the retention time range of the reference *n*-alkane series, and thus have RI<800.

3.7. Relative abundance of top-ranking compounds

The species-wise normalised peak area values for the top compounds are plotted as a heatmap in Fig. 4, representing their relative abundance across the two genera. Most, including the sesquiterpenes and the C_6–C_9 molecules, are seen to be more abundantly distributed within genus *Coleus*, and are thus candidate markers of this clade. Notable exceptions, however, include α-cubebene, γ-cadinene, isogermacrene D, and the bicyclic sesquiterpene, β-calacorene, which are also present in *P. fruticosus*. This is consistent with the outlying dispersion of the points for *P. fruticosus* on the PCA score plot (Fig. 3). In addition, the sesquiterpenes α-selinene and (Z,E)-α-farnesene, (Z)−3-hexen-1-ol as well as trans-β-ionone, are noticeably present in *P. zuluensis*, despite their overall distribution in genus *Coleus*.

4. Conclusion

The complexity of foliar VOC profiles, obtained by GC×GC-TOFMS, poses a substantial challenge to comprehensive analysis. The full set of foliar VOCs from the leaves of the species of southern African *Plectranthus* and *Coleus* consists of a high dimensional set of over a thousand tentative compound identifications. Pre-processing, including the removal of near-zero variance predictors, aids in reducing the dimensionality of the dataset. The chromatograms of species from both genera are characterised by peak clusters in two main regions of the separation space, corresponding to isomers of the monoterpenes and sesquiterpenes.

The machine learning algorithms used to model the data show a high degree of accuracy (up to 90%) in the prediction of genus, with a sensitivity (for genus *Coleus*) of up to 100%. The top-ranking variables listed by each of the models include C_6–C_{15} compounds of a variety of chemical classes. One key group are the sesquiterpenes (including β-ylangene, α-cubebene, β-copaene, α-cadinene and isogermacrene D), which are observed to be more abundantly and widely distributed overall across those *Coleus* specimens included in the sample population, and are thus potential markers of genus *Coleus*. Another group of VOCs with high predictor scores are C_6–C_9 unsaturated compounds (potential GLVs) also observed to have a greater distribution across the *Coleus* samples. A range of retention indices are found for the putative markers, which in the case of the sesquiterpenes suggests a greater isomeric variety than is captured in this analysis, as well as a rich phytochemistry for both genera.

Larger, independently sampled population sizes, representative of a wider species range, and the use of certified standards for compound identification, would aid in testing and expanding these findings for future studies.

Declaration of competing interests

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

CRediT authorship contribution statement

Daniel T. Pretorius: Formal analysis, Writing – original draft. **Egmont Rohwer:** Resources. **Yvette Naudé:** Supervision, Resources, Writing – review & editing.

Data Availability

Data will be made available on request.

Acknowledgements

We would like to acknowledge Damian Vaz de Sousa (Department of Plant Sciences, Faculty of Natural and Agricultural Sciences, University of Pretoria, Pretoria, Gauteng, South Africa) for his assistance in obtaining and identifying the specimens included in this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcoa.2022.100071.

References

[1] A. Paton, M. Mwanyambo, Culham A, Phylogenetic study of *Plectranthus, Coleus* and allies (Lamiaceae): taxonomy, distribution and medicinal use, Bot. J. Lin. Soc. 188 (2018) 355–376. https://doi.org/10.1093/botlinnean/boy064.

[2] A.J. Paton, M. Mwanyambo, R.H.A. Govaerts, K. Smitha, S. Suddee, P.B. Phillipson, T.C. Wilson, P.I. Forster, A. Culham, Nomenclatural changes in Coleus and Plectranthus (Lamiaceae): a tale of more than two genera, PhytoKeys 129 (2019) 1–158 10.3897/phytokeys.129.34988. https://doi.org/10.17660/ActaHortic.2001.552.18.

[3] C.W. Lukhoba, M.S.J. Simmonds, A.J. Paton, *Plectranthus*: a review of ethnobotanical uses, J. Ethnopharmacol. 103 (1) (2006) 1–24. https://doi.org/10.1016/j.jep.2005.09.011.

[4] L.J. Rice, G.J. Brits, C.J. Potgieter, J. Van Staden, *Plectranthus*: a plant for the future? S. Afr. J. Bot. 77 (4) (2011) 947–959. https://doi.org/10.1016/j.sajb.2011.07.001.

[1] In cases where the CAS is not available, the NIST entry number for the compound is reported.

[5] J.F.B. Pastore, R.M. Harley, F. Forest, A. Paton, C. van den Berg, Phylogeny of the subtribe Hyptidinae (Lamiaceae tribe Ocimeae) as inferred from nuclear and plastid DNA, Taxon 60 (5) (2018) 1317–1329. https://doi.org/10.1002/tax.605008.
[6] J.-.S. Zhong, J. Li, L. Li, J.G. Conran, H.-.W. Li, Phylogeny of *Isodon* (Schrad. ex. Benth) Spach (Lamiaceae) and related genera inferred from nuclear ribosomal ITS, *trnL-trnF* region, and *rps*16 intron sequences and morphology, Syst. Bot. 35 (1) (2010) 207–219. https://doi.org/10.1600/036364410790862614.
[7] J.K. Morton, Cytotaxonomic studies on the West African Labiatae, Bot. J. Linn. Soc. 58 (372) (1962) 231–283. https://doi.org/10.1111/j.1095-8339.1962.tb00896.x.
[8] S.T. Blake, A revision of *Plectranthus* (Labiatae) in Australasia, Contr. Queensland Herb. 9 (1971) 1–120. https://doi.org/10.5479/si.00810282.75.
[9] L.E. Codd, *Plectranthus*, in: O.A. Leistner (Ed.), Flora of southern africa, Lamiaceae, Bot. Res. Inst., 28(4), Pretoria, 1985, pp. 137–72.
[10] P.I. Forster, Five new species of *Plectranthus* L. Hér.(Lamiaceae) from New South Wales and Queensland, Austrobaileya 8 (3) (2011) 387–404 http://www.jstor.org/stable/41965592.
[11] A.J. Paton, G. Bramley, O. Ryding, R.M. Polhill, Y.B. Harvey, M. Iwarsson, F. Willis, P.B. Phillipson, K. Balkwill, D. Oteino, R.M. Harley, Lamiaceae, in: J. Timberlake (Ed.), Flora Zambesiaca, R. Bot. Gard., Kew, 8(8), London, 2013, pp. 346.
[12] C.Y. Wu, Y.C. Huang, Coleus, Flora Reipublicae Popularis Sinicae 66 (1977) 536–544.
[13] L.H. Cramer, A revision of *Coleus* (Labiatae) in Sri Lanka (Ceylon), Kew Bull 33 (3) (1978) 551–561. https://doi.org/10.2307/4109658.
[14] A.J. Paton, D. Springate, S. Suddee, D. Otieno, R.J. Grayer, Harley M.M, F. Willis, M.S.J. Simmonds, M.P. Powell, V. Savolainen, Phylogeny and evolution of basils and allies (Ocimeae, Labiatae) based on three plastid DNA regions, Mol. Phylogenet. Evol. 31 (1) (2004) 277–299. https://doi.org/10.1016/j.ympev.2003.08.002.
[15] R.J. Grayer, A.J. Paton, M.S.J. Simmonds, M-J.R. Howes, Differences in diterpenoid diversity reveal new evidence for separating the genus *Coleus* from *Plectranthus*, Nat. Prod. Rep. 38 (10) (2019) 1720–1728. https://doi.org/10.1039/D0NP00081G.
[16] M.E. Maffei, Sites of synthesis, biochemistry and functional role of plant volatiles, S. Afr. J. Bot. 76 (4) (2010) 612–631. https://doi.org/10.1016/j.sajb.2010.03.003.
[17] G. Arimura, R. Ozawa, T. Shimoda, T. Nishioka, W. Boland, J. Takabayashi, Herbivory-induced volatiles elicit defence genes in lima bean leaves, Nature 406 (6795) (2000) 512–515. https://doi.org/10.1038/35020072.
[18] R. Sasso, L. Iodice, M.C. Digilio, A. Carretta, L. Ariati, E. Guerrieri, Host-locating response by the aphid parasitoid *Aphidius ervi* to tomato plant volatiles, J. Plant Interact. 2 (3) (2008) 175–183. https://doi.org/10.1080/17429140701591951.
[19] C. Kost, M. Heil, Herbivore-induced plant volatiles induce an indirect defence in neighbouring plants, J. Ecol. 94 (3) (2006) 619–628. https://doi.org/10.1111/j.1365-2745.2006.01120.x.
[20] D. Tholl, W. Boland, A. Hansel, F. Loreto, U.S.R. Röse, J.-.P. Schnitzler, Practical approaches to plant volatile analysis, Plant J 45 (4) (2006) 540–560. https://doi.org/10.1111/j.1365-313X.2005.02612.x.
[21] C. Cagliero, G. Mastellone, A. Marengo, C. Bicchi, B. Sgorbini, P. Rubiolo, Analytical strategies for in-vivo evaluation of plant volatile emissions— A review, Anal. Chim. Acta. 1147 (1) (2021) 240–258. https://doi.org/10.1016/j.aca.2020.11.029.
[22] Z. Zhang, G. Li, A review of advances and new developments in the analysis of biological volatile organic compounds, Microchem. J. 95 (2) (2010) 127–139. https://doi.org/10.1016/j.microc.2009.12.017.
[23] S.M. Pourmortazavi, S.S. Hajimirsadeghi, Supercritical fluid extraction in plant essential and volatile oil analysis, J. Chromatogr. A. 1163 (1–2) (2007) 2–24. https://doi.org/10.1016/j.chroma.2007.06.021.
[24] J.S. Fritz, Analytical solid-phase extraction, Wiley-VCH, New York, 1999.
[25] Z. Zhang, J. Pawliszyn, Headspace solid-phase microextraction, Anal. Chem. 65 (14) (1993) 1843–1852. https://doi.org/10.1021/ac00062a008.
[26] C. Bicchi, C. Cordero, C. Iori, P. Rubiolo, Headspace sorptive extraction (HSSE) in the headspace analysis of aromatic and medicinal plants, J. High Resolut. Chromatogr. 23 (9) (2000) 539–546. https://doi.org/10.1002/1521-4168(20000901)23:9⟨539::AID-JHRC539⟩3.0.CO;2-3.
[27] F. Belliardo, C. Bicchi, C. Cordero, E. Liberto, P. Rubiolo, B. Sgorbini, Headspace-solid-phase microextraction in the analysis of the volatile fraction of aromatic and medicinal plants, J. Chromatogr. Sci. 44 (7) (2006) 416–429. https://doi.org/10.1093/chromsci/44.7.416.
[28] F. Zhu, J. Xu, Y. Ke, S. Huang, F. Zeng, T. Luan, G. Ouyang, Applications of *in vivo* and *in vitro* solid-phase microextraction techniques in plant analysis: a review, Anal. Chim. Acta. 794 (2013) 1–14. https://doi.org/10.1016/j.aca.2013.05.016.
[29] Y. Naudé, R. Makuwa, V. Maharaj, Investigating volatile compounds in the vapour phase of (1) a hot water infusion of rhizomes, and of (2) rhizomes of Siphonochilus aethiopicus using head space solid phase microextraction and gas chromatography with time of flight mass spectrometry, S. Afr. J. Bot. 106 (2016) 144–148. https://doi.org/10.1016/j.sajb.2016.07.006.
[30] M.B. Ngassoum, L. Jirovetz, G. Buchbauer, W. Fleischhacker, Investigation of essential oils of *Plectranthus glandulosus* Hook f. (Lamiaceae) from Cameroon, J. Essent. Oil Res. 13 (2) (2000) 73–75. https://doi.org/10.1080/10412905.2001.9699615.
[31] R.H. Alasbahi, M.F. Melzig, *Plectranthus Barbatus*: a review of phytochemistry, ethnobotanical uses and pharmacology – part 1, Planta Med 76 (7) (2010) 653–661. https://doi.org/10.1055/s-0029-1240898.
[32] L. Mota, A.C. Figueiredo, L.G. Pedro, J.G. Barroso, M.C. Miguel, M.L. Faleiro, L. Ascensão, Volatile-oils composition, and bioactivity of the essential oils of *Plectranthus barbatus, P. neochilus*, and *P. ornatus* grown in Portugal, Chem. Biodivers. 11 (5) (2014) 719–732. https://doi.org/10.1002/cbdv.201300161.
[33] P. Aziz, N. Muhammed, A. Intisar, M.A. Abid, M.I. Din, M. Yaseen, R. Kousar, A. Aamir, Quratulain, Ejaz, R, Constituents and antibacterial activity of leaf essential oil of *Plectranthus scutellarioides*, Plant Biosyst 155 (6) (2021) 1247–1252. https://doi.org/10.1080/11263504.2020.1837279.
[34] E.J. Van Jaarsveld, V. Thomas, The southern african *plectranthus* and the art of turning shade into glade, Fernwood Press, Simon's Town, 2006.
[35] H. van Den Dool, P.D. Kratz, A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography, J. Chromatogr. 11 (1963) 463–471. https://doi.org/10.1016/S0021-9673(01)80947-X.
[36] M. Kuhn, The caret package. http://topepo.github.io/caret/index.html, 2019 (accessed 23 November 2021).
[37] M. Kuhn, Building predictive models in R using the caret package, J. Stat. Softw. 28 (5) (2008) 1–26. https://doi.org/10.18637/jss.v028.i05.
[38] F. Loreto, C. Barta, F. Brilli, I. Nogues, On the induction of volatile organic compound emissions by plants as consequence of wounding or fluctuations of light and temperature, Plant Cell Environ 29 (9) (2006) 1820–1828. https://doi.org/10.1111/j.1365-3040.2006.01561.x.
[39] F. Brilli, T.M. Ruuskanen, R. Schnitzhofer, M. Müller, M. Breitenlechner, V. Bittner, G. Wohlfahrt, F. Loreto, A. Hansel, Detection of plant volatiles after leaf wounding and darkening by proton transfer reaction "time-of-flight" mass spectrometry (PTR-TOF), PLoS ONE 6 (5) (2011) e20419. https://doi.org/10.1371/journal.pone.0020419.