



Chemical diversity and species differentiation in Brazilian *Vanilla*: insights from LC-HRMS/MS metabolomics

Gesiane S. Lima^{1,2} · Giovanni B. Bevilaqua¹ · Hugo G. Machado¹ · Rosa B. N. Alves³ · Luciano B. Bianchetti³ · Lanaia I. L. Maciel¹ · Bianca M. M. G. Acioli¹ · Nerilson M. Lima⁴ · Roberto F. Vieira³ · Boniek Gontijo¹ · Gabriel F. dos Santos¹

Received: 8 January 2026 / Accepted: 9 March 2026 / Published online: 28 March 2026
© The Author(s) 2026

Abstract

Introduction *Vanilla* species represent a taxonomically complex and economically important group of orchids, yet species discrimination and chemotaxonomic characterization remain challenging due to high intrageneric metabolic variability and limited molecular insights.

Objectives We aimed to apply an untargeted metabolomic approach to characterize the metabolic signatures of three Brazilian *Vanilla* species (*V. pompona*, *V. phaeantha*, and *V. calyculata*) and to evaluate the relative contributions of species identity and biome origin to metabolic diversification.

Methods The leaf metabolome of 102 *Vanilla* were profiled using high-resolution LC-HRMS/MS. Multivariate analyses including PCA and PLS-DA served to explore variance structures and build discriminant models. Discriminant features were selected using integrated criteria (VIP scores and volcano-plot significance). Annotation leveraged GNPS molecular networking and in silico databases to classify key metabolite classes.

Results Unsupervised PCA revealed pronounced intraspecific metabolic heterogeneity, preventing spontaneous grouping by species or biome. Supervised PLS-DA models, however, provided robust species classification ($Q^2 = 0.74\text{--}0.90$), while biome-based models lacked predictive power, indicating limited environmental influence on leaf metabolomes. Consensus selection between models identified 17 core biomarkers, including phenolic acids, cinnamic acid derivatives, C-glycosylated flavonoids, lipids, terpenoids, and N-containing compounds, with ferulic acid (a key precursor of vanillin) emerging as a prominent discriminating metabolite.

Conclusions Untargeted LC-HRMS metabolomics coupled with chemometric modelling delineates species-specific metabolic fingerprints within *Vanilla*, offering a generalizable strategy for chemotaxonomy and species authentication. The conserved metabolic features also spotlight biologically meaningful pathways for biodiversity assessment and valorization of native *Vanilla* resources.

Keywords *Vanilla* · Metabolomics · LC-HRMS/MS · *Vanilla pompona* Schiede · *Vanilla phaeantha* Rchb.f. · *Vanilla calyculata* Schltr.

Gesiane S. Lima, Giovanni B. Bevilaqua and Hugo G. Machado have contributed equally to this work.

✉ Roberto F. Vieira
roberto.vieira@embrapa.br

✉ Gabriel F. dos Santos
franco.gabriel@ufg.br

¹ Institute of Chemistry, Federal University of Goiás, Goiânia, GO 59078-970, Brazil

² Department of Biological Sciences, State University of Santa Cruz, Ilhéus, BA 45662-900, Brazil

³ Embrapa Recursos Genéticos e Biotecnologia, Parque Estação Biológica, Distrito Federal, Brasília 70770-901, Brazil

⁴ Institute of Chemistry, Federal University of Alfenas, Alfenas, MG 37130-001, Brazil

1 Introduction

The genus *Vanilla* Mill. (Orchidaceae) comprises over 100 species distributed across the tropical regions of Africa, Asia, and the Americas (Pansarin & Menezes, 2023). In Brazil, this genus is represented by more than 40 species (Pansarin, 2025). Despite this diversity, the delimitation of species within *Vanilla* remains poorly understood, and several Brazilian taxa are currently synonymized. This taxonomic uncertainty may result in a misestimation of species' geographic distributions (Koch et al., 2013; Pansarin, 2024), limiting our understanding of their true diversity and hindering efforts to conserve and sustainably exploit *Vanilla* genetic resources.

Among the species within this genus, *Vanilla planifolia* is the most extensively cultivated and commercially traded, accounting for the majority of global vanilla extract production. *Vanilla tahitensis* J. W. Moore and *Vanilla pompona* Schiede are known to synthesize distinct aromatic compounds that confer unique flavor profiles to commercial products (de Oliveira et al., 2022). Several other *Vanilla* species also produce aromatic molecules and other secondary metabolites of economic interest (Chen et al., 2025; dos Santos et al., 2024), and the increasing global demand for natural vanilla extract has stimulated the exploration of alternative species as potential sources of valuable compounds. Accurate differentiation among these species is therefore essential for preserving biodiversity, ensuring quality control, and preventing commercial fraud (Chen et al., 2025; Krueger & Krueger, 1985). Beyond these aspects, reliable species identification provides critical insights for conservation strategies, supports the development of sustainable cultivation systems, and promotes the biotechnological exploitation of *Vanilla* species as promising reservoirs of aromatic and bioactive compounds (Flanagan et al., 2022; Hu et al., 2019; Nadal et al., 2025).

A key component of *Vanilla*'s aromatic profile is vanillin, which represents 1–8% of the dry weight of the pods and accounts for approximately 80% of their total aromatic compounds (Perera & Owen, 2010; Zhang et al., 2014). Beyond its well-known role as a flavoring and fragrance agent, vanillin also exhibits antimicrobial and antioxidant properties (Wang et al., 2024), making it a valuable natural preservative in food and cosmetic applications (Brunschwig et al., 2009; Ciriminna et al., 2019; Tomadoni et al., 2016). This reinforces the importance of characterizing the metabolic diversity among *Vanilla* species, since variations in vanillin content and the presence of other bioactive metabolites can greatly influence the sensory, commercial, and functional properties of *Vanilla* products (de Oliveira et al., 2024). Therefore, a comprehensive and accurate taxonomic framework is fundamental to fully explore the

economic, ecological, and biotechnological potential of this diverse genus.

In this context, metabolomics has emerged as a powerful approach for the comprehensive chemical characterization of plants, enabling the differentiation of species based on their metabolic profiles (Salem et al., 2020). Liquid chromatography coupled with mass spectrometry (LC–MS) is one of the most widely employed techniques in metabolomics, offering high-resolution separation and accurate identification of both primary and secondary metabolites (Phapale et al., 2020; Xiao et al., 2012; Zhou et al., 2012). This approach is particularly valuable for analyzing the complex chemical composition of plant species, as it enables the detection of bioactive compounds such as alkaloids, flavonoids, and phenolic derivatives (Álvarez-Fernández et al., 2015; Lima et al., 2025; Niculau et al., 2016). Additionally, LC–MS could facilitate the differentiation of taxa by generating comprehensive metabolic fingerprints, which can be further explored using multivariate statistical analyses and structural annotation tools to identify species-specific biomarkers (Kharazian et al., 2025; Worley & Powers, 2012).

In this study, we apply metabolomics-based approaches using LC–MS to explore the metabolic diversity of Brazilian *Vanilla* species, aiming to differentiate taxa through their distinctive chemical signatures. By combining LC–MS information with chemometric data analysis, we seek to identify key metabolic markers that support accurate species classification, quality control, and the authentication of vanilla-derived products. This integrative approach advances the understanding of *Vanilla* phytochemistry and highlights the potential of metabolomics as a powerful tool for biodiversity assessment, conservation planning, and the sustainable exploitation of *Vanilla* genetic resources.

2 Materials and methods

2.1 Chemicals and materials

HPLC-grade methanol was obtained from Tedia Company (Fairfield, USA). Formic acid and caffeine-(trimethyl- $^{13}\text{C}_3$) were acquired from Sigma-Aldrich (St. Louis, USA), while progesterone- d_9 was supplied by CDN Isotopes (Quebec, Canada). Ultrapure water, with a resistivity of 18.2 M Ω ·cm, was produced using a Master System MS2000 water purification system (Gehaka, São Paulo, Brazil).

2.2 Sample preparation

A total of 102 different *Vanilla* plants were analyzed, comprising fifty from *V. pompona* Schiede, thirty-five from *V. phaeantha* Rchb.f., seventeen from *V. calyculata* Schltr.

The plants were collected from different locations in Brazil (representing different biomes: Cerrado, Amazon, Atlantic Forest, and Caatinga) and cultivated in a greenhouse at the Embrapa Genetic Resources and Biotechnology in Brasília, Federal District, Brazil. All information regarding species, biomes, and the number of samples are presented in the supporting information, Table S1.

The Amazon occupies 49% of Brazil, is predominantly located in the Northern Region, and has an equatorial climate with temperatures of 25–28 °C, high humidity, and abundant, well-distributed rainfall; soils are nutrient-poor and intensely leached, with fertility concentrated in the biomass. The Cerrado covers ~23.3% of the national territory and is the largest tropical savanna in South America, occurring mainly in central Brazil, with a seasonal tropical climate (dry season from May to September and rainy season from October to April), temperatures of 22–27 °C, and old, acidic, highly weathered, nutrient-poor soils. The Caatinga, exclusively Brazilian and the most biodiverse semi-arid region worldwide, presents a semi-arid climate with temperatures above 25 °C, scarce and irregular rainfall, and prolonged droughts; soils are generally shallow and stony, sometimes relatively fertile, but prone to desertification and erosion. The Atlantic Forest extends across 17 states along the Brazilian coast and has a tropical humid climate, high rainfall and humidity, variable temperatures, and ocean-influenced precipitation; soils are typically deep and weathered, potentially fertile, yet prone to erosion (de Queiroz et al., 2023; IBGE. Instituto Brasileiro de Geografia e Estatística, 2019).

The plant material is registered in the Sistema Nacional de Gestão do Patrimônio Genético (SisGen) under code A301AFD. Fresh leaves from each plant were collected, cleaned, and freeze-dried to complete dryness using a Lyotech Freeze Dry System LS 3000 (São Carlos, Brazil) for 48 h. The dried leaves were then ground into a fine powder using an electric grinder and stored at –20 °C until further analysis.

The extraction procedure followed the method described by Vos et al. (De Vos et al., 2007). Briefly, 50 mg of the ground plant material was extracted with 1 mL of a cold methanol: water solution (75:25, v/v) containing 0.1% formic acid. The extraction process involved 10 s of vortexing, followed by 15 min in an ultrasonic bath, and concluded with centrifugation at 15,000 rpm for 10 min. The resulting supernatant was filtered through a 0.22 µm membrane filter and transferred to an HPLC vial (280 µL). Before injection, the samples were spiked with 20 µL of a solution containing caffeine-¹³C₃ and progesterone-*d*₉ at an individual concentration of 50 µg mL^{–1}, resulting in a final concentration of 3.57 µg mL^{–1} for each internal standard. For quality control (QC) samples, 10 µL aliquots from each of the 102 samples

were pooled to generate a representative mixture, which was used to monitor instrument performance, ensure data consistency, and facilitate normalization during metabolomic analysis.

2.3 LC-HRMS/MS analysis

LC-HRMS/MS analyses were performed using an HPLC 1220 Infinity II (Agilent Technologies) coupled to a Q-Exactive hybrid Quadrupole-Orbitrap high-resolution mass spectrometer (Thermo Scientific) equipped with an electrospray ionization (ESI) source. Separation was achieved using a Luna Omega Phenomenex C18 column (4.6 × 150 mm × 5 µm; Agilent) with a gradient elution program. The mobile phases consisted of (A) water with 0.1% formic acid and (B) methanol with 0.1% formic acid. The gradient started at 15% B, held for 4 min, increased linearly to 100% B over 15 min, maintained for 15 min, returned to initial conditions in 1 min, and re-equilibrated for 4 min. The flow rate was set to 0.5 mL min^{–1}, the injection volume 20 µL, and the column temperature was maintained at 40 °C.

The ESI source parameters used in positive mode were optimized as follows: spray voltage 3.5 kV; the capillary temperature was 250 °C; S-lens RF level 60 V; sheath gas flow rate at 47 L min^{–1}; and aux gas flow rate at 11 L min^{–1}. High-resolution mass spectra were acquired in the Full MS/data dependent -MS² (dd-MS²) mode. The mass range in the full MS scanning experiments was *m/z* 100–1200. The top 5 (TopN, 5, loop count 5) most abundant precursors were sequentially transferred for collision-induced fragmentation acquisition. The collision energy for target analytes was 20, 30, and 35 eV. Resolving power was 140,000 and 70,000 for full MS and dd-MS² acquisitions, respectively.

2.4 Metabolite's annotation

Metabolite's annotation was performed using a combination of spectral matching and in silico tools available through the Global Natural Products Social Molecular Networking (GNPS) platform. The spectra data obtained from the LC-MS/MS analysis were first converted to the mzML format using MSConvert (ProteoWizard) and then processed with MZmine (v2.53) for feature detection and alignment. Annotation was primarily based on spectral matching against the GNPS MS/MS spectral libraries, which categorize matches according to spectral quality into gold (well-characterized standards), silver (tentative identification from crude extracts), and bronze (partial annotations) (Wang et al., 2016). Multiple GNPS tools were applied to enhance annotation coverage and confidence, including Classical Molecular Networking (MN) and Feature-Based Molecular Networking (FBMN) for visualization of structurally

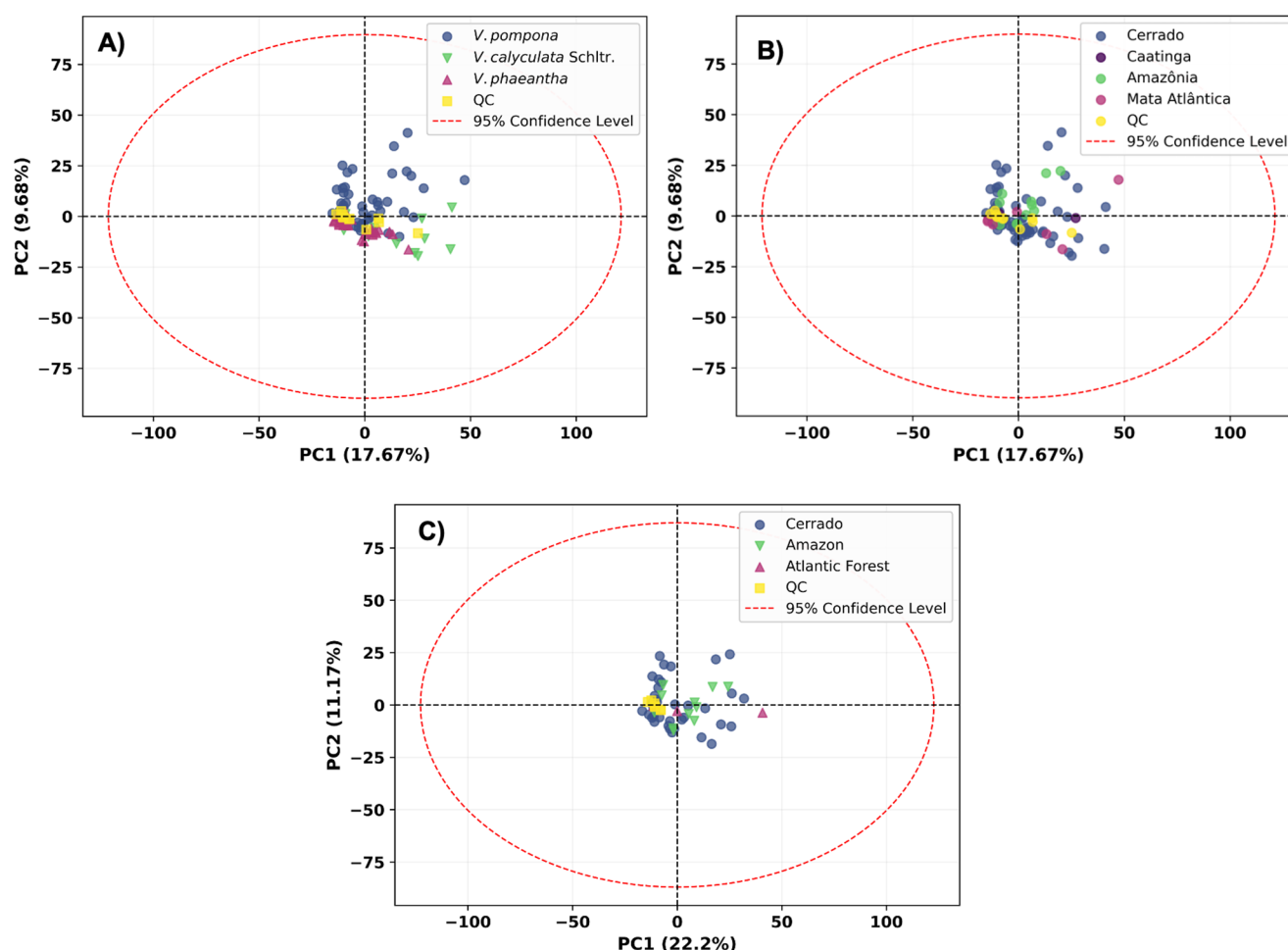


Fig. 1 PCA scores plot of *Vanilla* samples. Samples are colored by **A** species, **B** biomes for all species, and **C** biomes for only *V. pompona*, showing a lack of natural clustering

related metabolite clusters (Nothias et al., 2020; Wang et al., 2016), DEREPLICATOR+ for in silico identification of peptide natural products (Mohimani et al., 2018), and Network Annotation Propagation (NAP) for structure propagation within molecular networks (da Silva et al. 2018). Additional tools such as MOLDISCOVERY (Cao et al., 2021), MS2LDA for substructure discovery (van der Hooft et al., 2016), and MolNetEnhancer for chemical class annotation were also employed (Ernst et al., 2019).

2.5 Chemometric data analysis

The spectral data were extracted directly from the resulting RAW files using the RawFileReader library (github.com/thermofisherlms/RawFileReader), implemented in a custom Python3 script developed in-house. Following raw data extraction, the spectra were processed by discarding signal intensities below 1% of the maximum intensity within each scan, followed by spectral summation and base peak normalization, resulting in relative intensities from 0 to 100%.

A mass alignment was performed across the samples, resulting in a final data matrix containing 113 rows (samples) and 1078 columns, each representing an m/z variable with its respective relative intensity. All statistical processing and data normalization were carried out using custom Python routines. This type of untargeted mass alignment has been shown to be effective for metabolomics analysis in complex plant systems (Gowda & Djukovic, 2014; Smith et al., 2006).

3 Results and discussion

An untargeted metabolomics strategy was employed to investigate the molecular diversity of three Brazilian *Vanilla* species (*V. pompona*, *V. phaeantha*, and *V. calyculata* Schltr.) by analyzing ion abundance and metabolic composition using LC-HRMS/MS. To support metabolite annotation, dereplication procedures were employed by matching candidate structures within comprehensive spectral libraries,

particularly those available through the GNPS platform. These analyses were complemented by *in silico* tools for MS/MS fragmentation prediction and structural elucidation, enhancing the reliability and depth of metabolite identification.

Leaf samples representing the four main Brazilian biomes (Cerrado, Amazon, Atlantic Forest, and Caatinga) were collected from individuals of the *Vanilla* Germplasm Bank at Embrapa Genetic Resources and Biotechnology, in Brasília, Federal District, for sample preparation. All information regarding species, biomes, and the number of samples are presented in the supporting information, Table S1.

The LC-HRMS/MS workflow included regular QC injections across all batches to assess reproducibility and monitor signal drift. All samples were spiked with two internal standards, and their retention times and ion intensities were monitored throughout the analysis. The consistency of these parameters confirmed the robustness and stability of the analytical platform throughout the measurements.

An initial investigation of the LC–MS dataset, which comprises 102 unique samples from three species (*V. pompona*, *V. phaeantha*, *V. calyculata* Schltr.) and four biomes, was performed using Principal Component Analysis (PCA). This unsupervised technique is designed to capture the maximum variance in the data, providing an unbiased view of its inherent structure (Bro & Smilde, 2014).

The resulting PCA scores plot, shown in Fig. 1, reveals no distinct clustering of samples, either when colored by species (Fig. 1A), biomes for all species (Fig. 1B), or biomes for only *V. pompona* (the most abundant distribution) (Fig. 1C). The samples from different groups show extensive overlap, and the QC samples are tightly clustered near the origin, confirming the analytical stability of the LC–MS runs. The absence of spontaneous clustering among samples in the PCA highlights the predominance of intra-specific chemical variability in *Vanilla* leaves. This pattern is consistent with previous metabolomic investigations of *V. planifolia* leaves, which demonstrated that leaf chemistry is strongly influenced by ontogeny, photoperiod, and seasonal

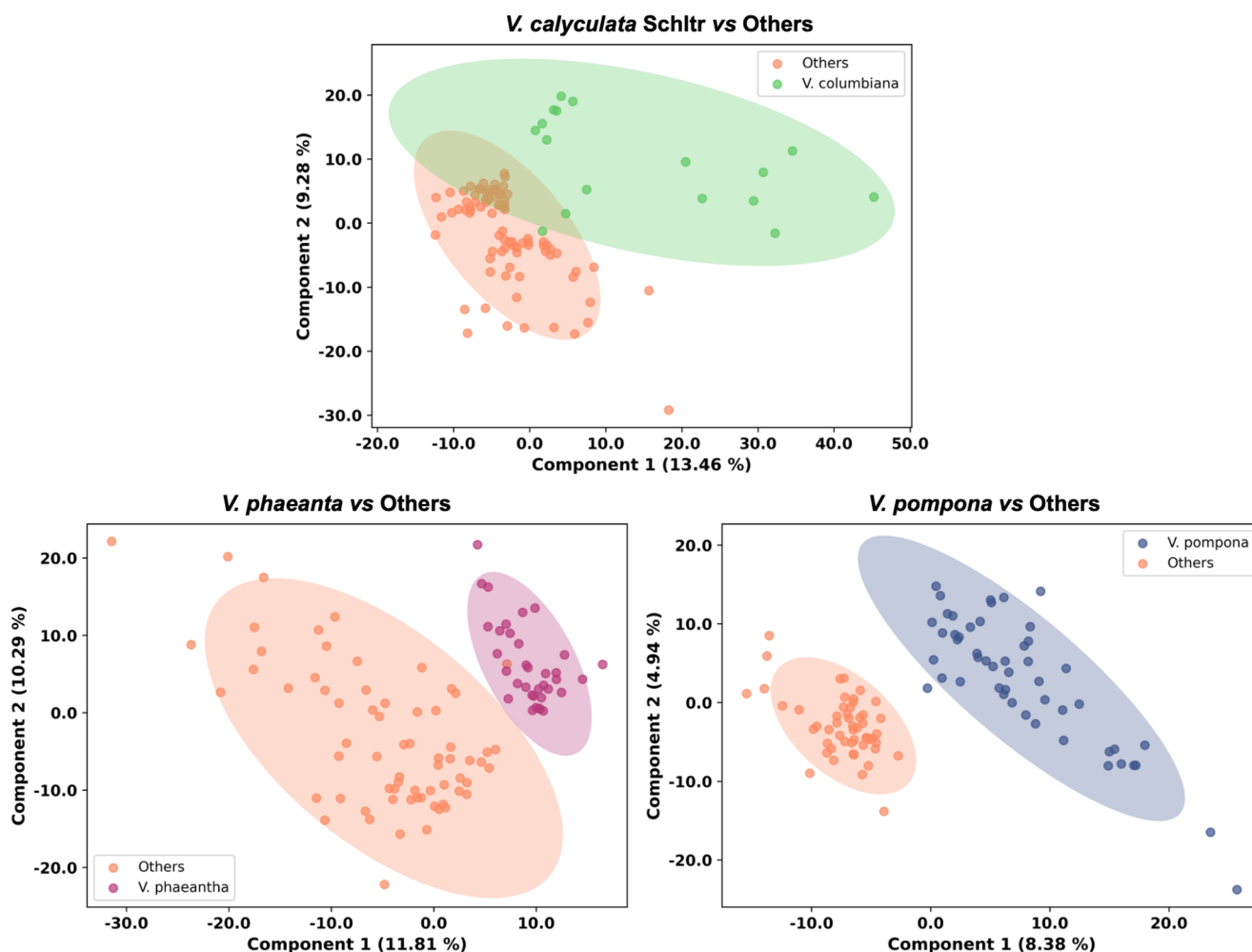


Fig. 2 PLS-DA scores plot for species-based classification models, showing clear separations between the classes

fluctuations, resulting in substantial intraindividual and intraspecific variation that can obscure taxonomic patterns (Palama et al., 2009). Such physiological plasticity inherent to leaf metabolism likely contributed to the broad chemical dispersion observed in our dataset, explaining why unsupervised multivariate methods were insufficient to resolve species-level groupings. Additionally, it is also important to note that the lack of clear species- or biome-specific grouping may be influenced by the fact that only leaf tissue was analyzed in this study. In *Vanilla* species, the characteristic aroma and many of the key secondary metabolites (e.g., vanillin and related phenolics) are predominantly synthesized and accumulated in the pods rather than in leaves, and metabolomic profiling of pods and mature beans has shown clear and statistically significant differences in metabolic composition compared to vegetative tissues (Beer et al., 2025; Gu et al., 2017). Previous studies have shown that developing and curing pods undergo pronounced biochemical transitions, accumulating glucovanillin, vanillin, and other phenylpropanoid derivatives far more intensely than leaves (Palama et al., 2009). Moreover, integrative

metabolomic and transcriptomic work has demonstrated that key steps in vanillin biosynthesis, including those derived from ferulic acid, are strongly upregulated during pod maturation (Hernández-Peña et al., 2025). Thus, future analyses incorporating both pod and mature beans are likely to reveal more diagnostically powerful markers for both chemotaxonomy and flavor-related metabolic pathways that remain undetectable when only leaves are examined.

To deepen the analysis and focus on inter-species discrimination, a supervised approach using Partial Least Squares Discriminant Analysis (PLS-DA) was employed. Unlike PCA, PLS-DA is specifically designed to maximize the covariance between the spectral data (X-matrix) and a predefined class membership variable (Y-matrix), making it ideal for finding latent structures related to class differences (Barker & Rayens, 2003; Brereton & Lloyd, 2014; Lê Cao et al., 2011; Wold et al., 2001). Given the inherently binary nature of the PLS-DA algorithm, a *one-vs-all* strategy was employed to build separate models for species and biome classification. This approach mitigates the method's multi-class limitations by decomposing the problem into multiple

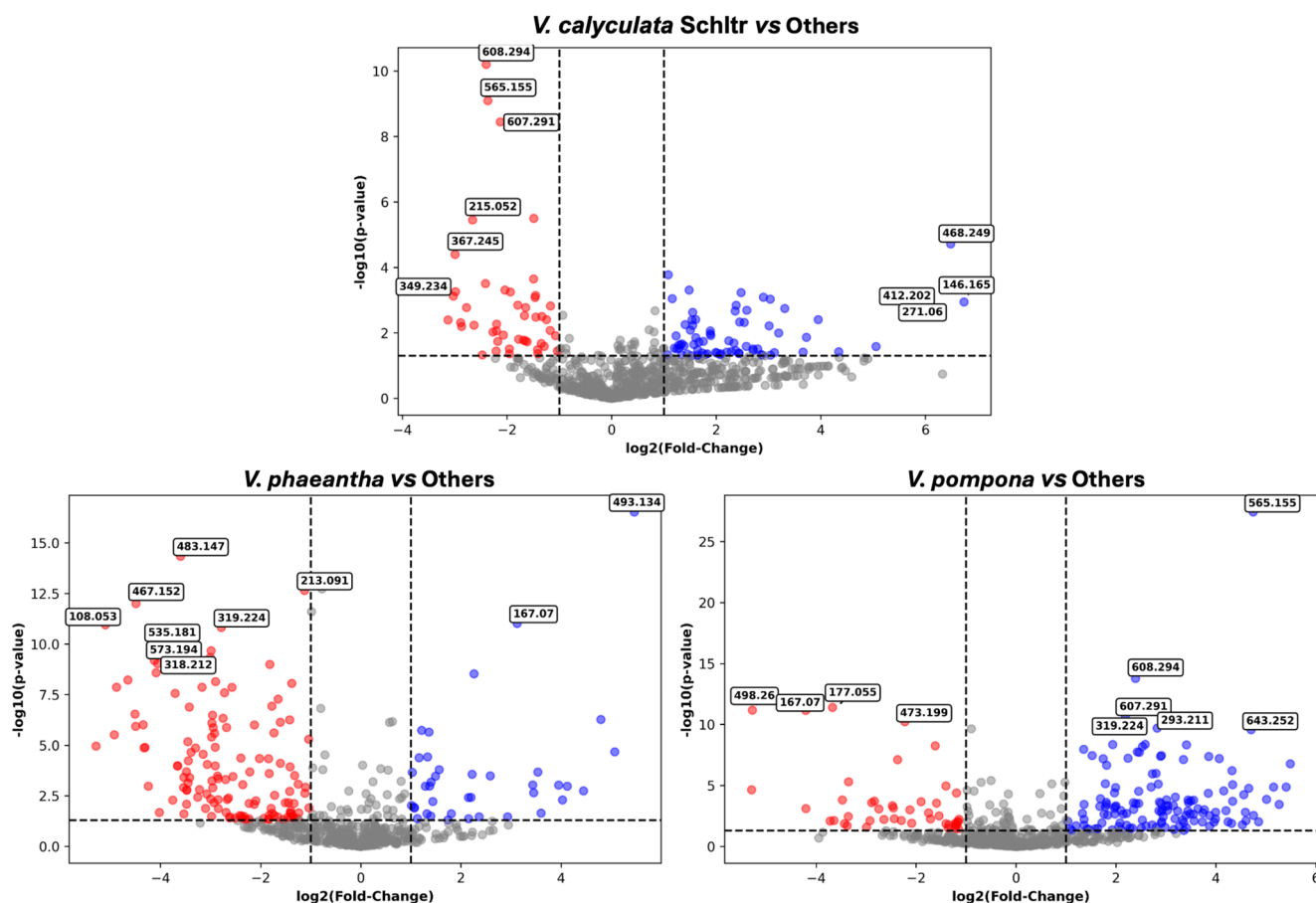


Fig. 3 Volcano plots showing discriminant *m/z* features for each species comparison: *V. pompona* vs. “Others”, *V. phaeantha* vs. “Others”, and *V. calyculata* Schltr. vs. “Others”. Each point represents an *m/z* feature. Features in the upper-right (more abundant in the target

species) and upper-left (more abundant in the “Others”) quadrants are considered significantly different based on statistical significance (*p*-value) and magnitude of change (fold-change)

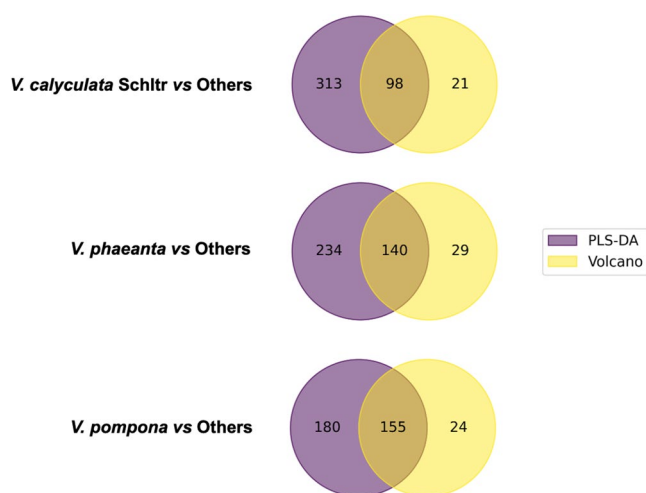


Fig. 4 Venn diagrams show the intersection of discriminating m/z features for species-based datasets, identified by VIP score > 1.0 and Volcano plot significance criteria

distinct two-class models (Brereton & Lloyd, 2014). Model optimization revealed that performance for all models was optimal using five latent components, as metrics plateaued thereafter (Figs. S1 and S2). The models designed to separate each species from the “Others” group yielded excellent results. Specifically, the *V. pompona* model showed outstanding predictive power ($Q^2=0.90$), followed closely by the *V. phaeantha* model ($Q^2=0.82$). The model for *V. calyculata* Schltr. was also robust and predictive ($Q^2=0.74$). More details are shown in Fig. S1 and Table S2. These high Q^2 values, confirmed through leave-one-out cross-validation (LOOCV), indicate that the models identified true, generalizable chemical patterns for each species and were not merely overfitting the data (Neto et al., 2021; Triba et al., 2015). In the context of PLS-DA, Q^2 is a key metric for evaluating predictive ability because it directly reflects the cross-validated prediction error of the model. To further complement the performance assessment, Receiver Operating Characteristic (ROC) curves were also calculated (Xia et al., 2013). The resulting Area Under the Curve (AUC) values confirmed the strong classification capability of the models. The ROC curves are provided in the Supporting Information (Fig. S3). Despite this variability, supervised PLS-DA models achieved strong predictive performance for species discrimination, indicating that species-level chemical fingerprints do exist. Similar findings were reported by Leyva et al., (2021), who demonstrated that leaf metabolomes of *V. planifolia*, *V. pompona*, *V. palmarum*, and *V. ribeiroi* could be reliably discriminated through multivariate analyses despite their high intraspecific variance (Leyva et al., 2021).

In contrast, all models built to classify samples by biome failed to produce predictive results. Despite some high

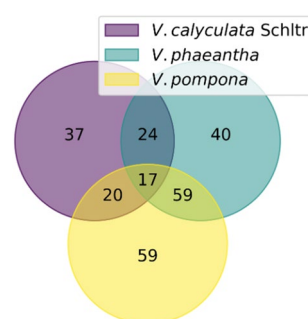


Fig. 5 Three-way Venn diagram comparing the key biomarkers (features present in both VIP > 1.0 and Volcano Plot lists) for *V. calyculata* Schltr., *V. phaeantha*, and *V. Pompona*

accuracy and R^2 values, their Q^2 values were consistently poor, ranging from -0.19 to a maximum of only 0.22 for all biomes across species and from -0.17 to 0.16 in the biome-based models restricted to *V. pompona*. Low or negative Q^2 values indicate poor predictive ability and suggest that the models do not capture a robust relationship between the data matrix and the biome classes. This finding strongly suggests that, while a distinct chemical fingerprint exists for each species, there is no consistent metabolic signature capable of reliably differentiating the biomes represented in this dataset (details in Fig. S2 and Table S2). These findings are also aligned with previous work showing that environmental origin explains far less variance in *Vanilla* leaves than endogenous physiological factors, which dominate the metabolomic landscape (Palama et al., 2009).

The clear separation achieved by the species-specific models is visualized in the PLS-DA scores plots presented in Fig. 2. While all three models effectively discriminated the target species from the “Others” group, they did so with varying degree of distinction, a pattern consistent with their performance metrics. The model for *V. pompona* exhibited the most pronounced separation, corresponding to an accuracy of 100% using five components. Similarly, the *V. phaeantha* model showed a clear separation and achieved 99% accuracy. Although the model for *V. calyculata* Schltr. displayed the least distinct visual separation among the three, it remained highly effective, successfully differentiating the groups with 97% accuracy.

Having established that each species exhibits a robust chemical fingerprint, the next logical step was to identify the specific m/z features driving this discrimination. Discriminant variable selection was performed using a consensus-based strategy to ensure both biological relevance and statistical robustness. Although PLS-DA is well suited for addressing the multicollinearity characteristic of metabolomics datasets, Brereton and Lloyd (Brereton & Lloyd, 2014) caution that the method remains vulnerable to overfitting and the identification of spurious correlations in high-dimensional data. To mitigate these risks in

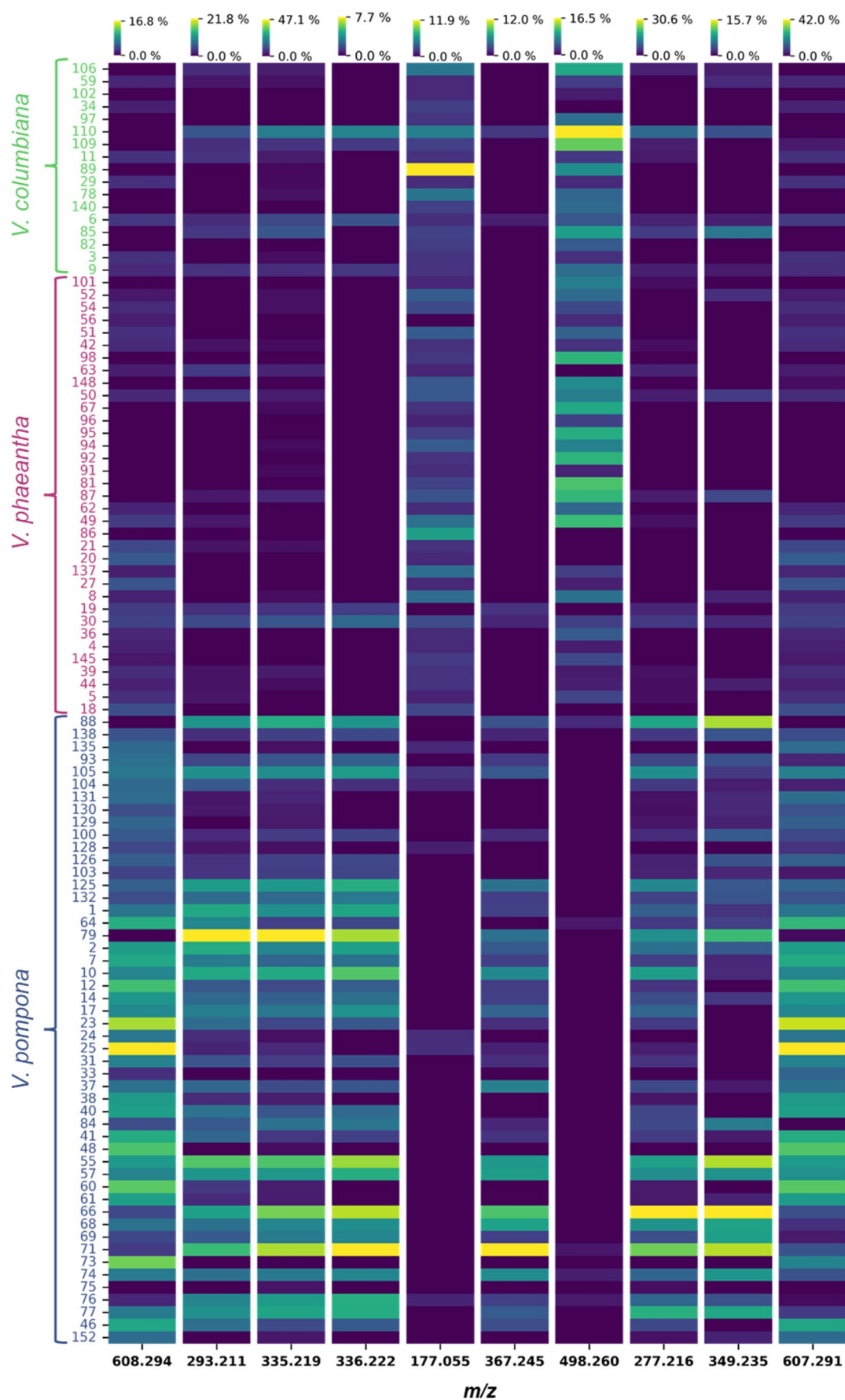


Fig. 6 Heatmap of the top ten shared biomarkers with the highest VIP scores. Samples (rows) are grouped by species, and features (columns) are the selected m/z values. Colors represent normalized relative intensities, where each feature was independently scaled to highlight relative abundance patterns across species

feature selection, variable relevance was defined by the intersection of two complementary criteria: (i) multivariate importance, assessed through VIP scores > 1 in the PLS-DA models, as advocated by Wold et al. (Wold et al., 2001) for summarizing variable contribution in latent variable projections; and (ii) univariate significance and effect magnitude (fold-change), evaluated through Volcano Plot analysis, in accordance with established metabolomics protocols for identifying statistically significant features and distinguishing biological groups (Pan et al., 2021; Xu et al., 2024). This dual strategy ensures that selected metabolites contribute to the model's global covariance structure (PLS) while also exhibiting statistically significant mean differences between groups, thereby aligning with recommendations that complex models be validated against the fundamental statistical properties of the dataset. Model robustness itself was assessed independently through leave-one-out cross-validation (LOOCV), used to calculate Q^2 values, and further supported by permutation testing to verify that the observed discrimination did not arise from random class assignments (Fig. S4). For each species-versus-others comparison, Volcano plots were constructed by plotting statistical significance ($-\log_{10}(p\text{-value})$) against the magnitude of change ($\log_2(\text{fold change})$). As shown in Fig. 3, this approach enables the rapid identification of features that are both statistically significant (high on the y-axis) and exhibit large relative differences between groups (far on the x-axis). To the further control for multiple hypothesis testing inherent to untargeted metabolomics datasets, the univariate p-values were subsequently corrected using the Benjamini-Hochberg false discovery rate (FDR) procedure (Benjamini & Hochberg, 1995). The complete table containing both raw p-values and FDR-adjusted q-values for all detected features is provided in the Supporting Information (Supplementary Spreadsheet), together with updated Volcano plots and Venn diagrams based on the q-value threshold (Fig. S5).

To consolidate the findings from the univariate and multivariate analyses and identify the most reliable biomarkers, the list of significant features obtained from the Volcano Plots ($p < 0.05$ and $|\log_2(\text{FC})| > 1$) was then intersected with the list of variables deemed important by the PLS-DA model ($\text{VIP} > 1.0$). This combined filtering strategy ensures that selected metabolites not only contribute meaningfully to the model's covariance structure but also display robust group-level differences. The overlap between these two feature sets was visualized using a

Venn diagram (Fig. 4), highlighting the metabolites that consistently emerged as discriminant across both analytical frameworks. A similar combination of multivariate and univariate criteria has been shown to improve marker reliability in other metabolomic studies of *Vanilla*, supporting the robustness of the dual-filtering strategy employed here (Leyva et al., 2021).

The interpretation of this diagram is threefold.

- (i) Intersection (VIP & Volcano): Features in this region represent the key biomarkers. They are validated as both statistically significant on their own and highly influential in the multivariate model, making them the highest-confidence candidates for differentiating the species.
- (ii) VIP Only: Features in this region are important for the model's predictive performance but lack strong individual statistical significance. Their contribution is driven by synergistic or combinatorial effects with other variables, underscoring the value of a multivariate perspective.
- (iii) Volcano Only: Features in this region are statistically significant but were not prioritized by the PLS-DA model, possibly due to redundancy with other, more informative variables already captured by the model's latent structure.

Building on these results, we next examined the uniqueness and overlap of the chemical fingerprints across the three species. To do so, the key biomarkers (i.e., the intersection lists) identified for each species were compared using a three-way Venn diagram Fig. 5. This analysis reveals which markers are exclusive to a single species, serving as potential unique identifiers, and which are shared between two or all three species, potentially reflecting common biosynthetic pathways or structurally related metabolites. These patterns of shared and unique metabolites indicate that species differentiation in *Vanilla* arises from a combination of conserved metabolic scaffolds and species-specific modulations. Leyva et al., (2021) reported analogous trends, showing that Peruvian *Vanilla* species share core phenylpropanoid metabolites while differing in their relative abundances and in complementary pathways, supporting the idea of modular metabolic evolution within the genus (Leyva et al., 2021). This final step provides deeper insight into the chemical relationships among the species and yields a refined set of high-priority targets for structural elucidation.

The analysis of the three-way Venn diagram (Fig. 5) revealed a core set of 17 m/z features consistently identified as key molecular markers across all three species-specific models, indicating that these metabolites

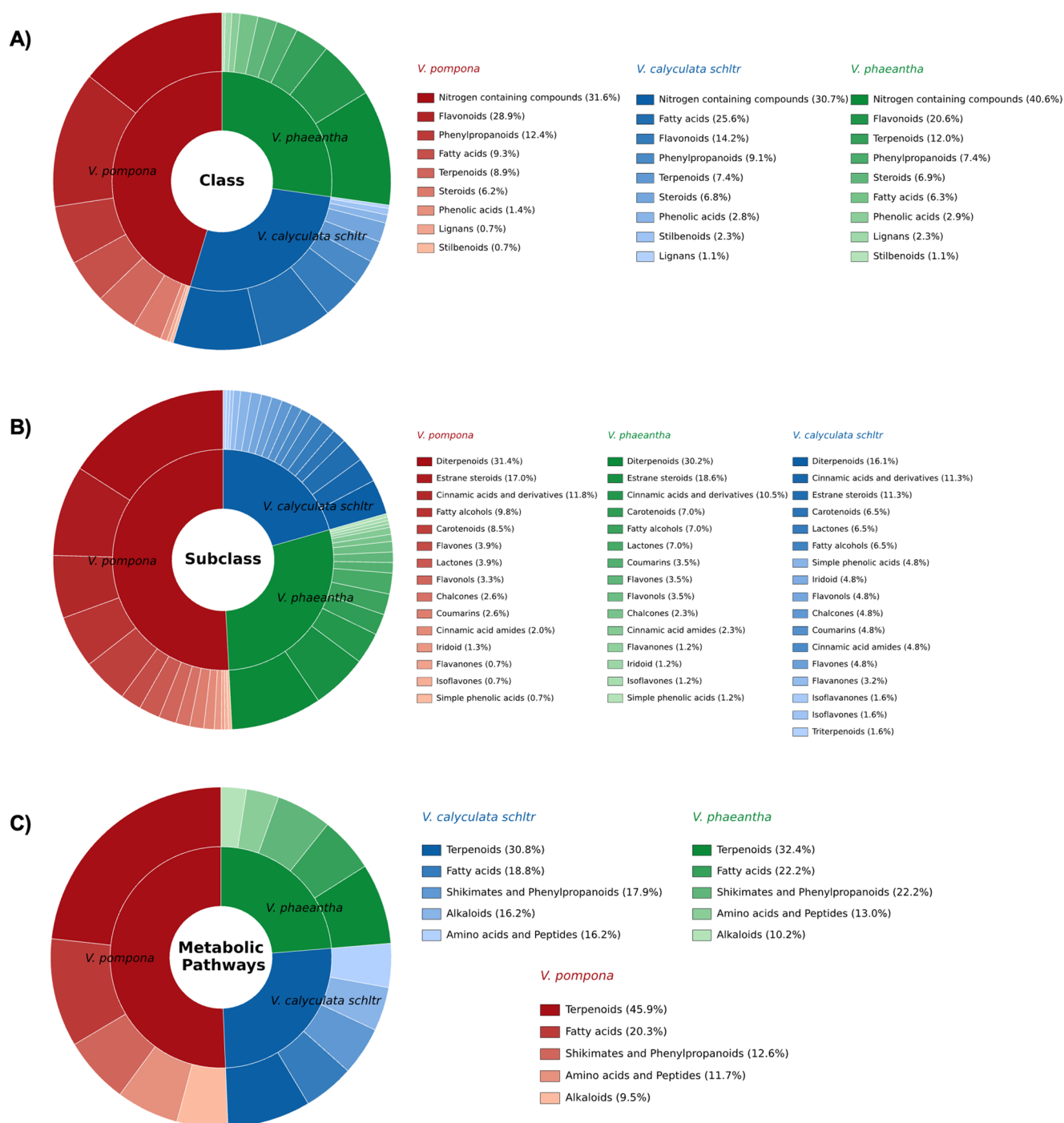


Fig. 7 Classification of annotated metabolites in *V. calyculata* Schltr., *V. pompona*, and *V. phaeantha* based on MolNetEnhancer analysis. **A** Distribution across major chemical classes (flavonoids, terpenoids, lignans, steroids, phenylpropanoids, nitrogen-containing compounds, phenolic acids, fatty acids, stilbenoids), **B** subclass-level annotation

contribute fundamentally to the overall chemical differentiation among the studied *Vanilla* species leaves. Notably, ferulic acid (m/z 177.055 $[M+H]^+$) and two esterified cinnamic acid derivatives—one conjugated to a terpenoid moiety (m/z 607.291 $[M+H]^+$) and another linked to a

(e.g., diterpenoids, sterane steroids, cinnamic acids and derivatives, flavones, flavonols, chalcones, carotenoids, coumarins, isoflavones), and **C** mapping to biosynthetic pathways (terpenoids, fatty acids, shikimates and phenylpropanoids, amino acids and peptides, alkaloids)

nitrogen-containing chain (m/z 498.260 $[M+Na]^+$)—emerged as central phenolic constituents. Additional shared discriminant features included the *C*-glycosylated flavonoid 6,8-di-*C*-glucopyranosylnaringenin (m/z 597.449 $[M+H]^+$); several hydroxy fatty acid derivatives,

represented by ions at m/z 277.216 $[M+H]^+$ and 351.214 $[M+Na]^+$; and a lipid species at m/z 318.248 $[M+H]^+$, putatively assigned to a trihydroxylated sphingoid base. Further discriminant metabolites comprised an oxygenated polyunsaturated fatty acid derivative (m/z 293.211 $[M+H]^+$), a lactone (m/z 367.245 $[M+H]^+$), and a phenolic acid derivative (m/z 349.235 $[M+H]^+$). In addition, three trihydroxylated kaurane-type diterpenoids (m/z 331.224, 333.240, and 335.219 $[M+H]^+$), along with nitrogen-containing metabolites detected at m/z 336.222 and 608.294 $[M+H]^+$ were identified as shared markers, together reflecting the interplay of multiple biosynthetic pathways in shaping species-specific chemical profiles.

Collectively, this metabolite set reinforces the central role of phenolic compounds, particularly phenolic acids and flavonoids, as fundamental secondary metabolites driving interspecific differentiation within the *Vanilla* genus. Among these shared features, ferulic acid emerged as a prominent marker. This finding has important biochemical implications, as ferulic acid is a confirmed precursor in the vanillin biosynthetic pathway, undergoing oxidative and reductive transformations that give rise to vanillin during pod development (Gallage & Möller, 2015; Kundu, 2017; Vogt, 2010). Its consistent detection across species suggests that vanillin-related phenylpropanoid metabolism constitutes a conserved biochemical axis within the genus. Similar observations were made by Leyva et al., (2021), who identified vanillin precursors (including glucosides A and B) as discriminant metabolites in the leaves of four *Vanilla* species, reinforcing the central chemotaxonomic role of phenolics. The prominence of flavonoids, terpenoids, lipid derivatives, and nitrogen-containing metabolites in our shared marker set further highlights the participation of multiple biosynthetic pathways in shaping interspecific differentiation. These compound classes have also been associated with adaptation, stress responses, and defense in *V. planifolia* and related taxa (Leyva et al., 2021; Palama et al., 2009). To further illustrate the abundance patterns of these shared discriminant ions, a heatmap was generated using the ten features with the highest VIP scores from this core set (Fig. 6). To further evaluate the clustering structure of the shared biomarkers, a hierarchical clustering heatmap using the full set of 17 discriminant ions is provided in the Supporting Information (Fig. S6). The complete list of detected features together with their corresponding statistical parameters (VIP scores, p-values, and fold changes) for each species vs. others comparison is provided in the Supporting Information as a supplementary spreadsheet.

The class-level metabolite distribution highlighted clear chemical distinctions among the three *Vanilla* species (Fig. 7). Nitrogen-containing compounds emerged as

the most abundant category across all taxa, particularly in *V. pompona*, suggesting a strong metabolic allocation to alkaloid- and peptide-related pathways in *Vanilla* genus (Dibwe et al., 2023). These compounds are often associated with ecological interactions such as defense and signaling, which may reflect adaptive strategies of *Vanilla* species to their natural habitats. Flavonoids represented another dominant class, with *V. pompona* (28.9%) again showing a markedly higher number of features than the other two species, followed by *V. phaeantha* (20.6%). This pattern underscores the importance of flavonoid metabolism for *Vanilla*, not only as pigments but also as antioxidants and modulators of plant–environment interactions (Beer et al., 2025). Metabolites classified as terpenoids displayed consistently low abundance across all samples (<12%), a trend that may be attributed to their limited ionization efficiency under electrospray conditions, with detection being favored mainly for glycosylated derivatives. Within this context, *V. phaeantha* showed comparatively higher levels of this class. In contrast, phenylpropanoids, predominantly phenolic acids, were more abundant in *V. pompona* (12.4%), highlighting species-specific differences in the distribution of major secondary metabolite classes. Fatty acids, conversely, were particularly enriched in *V. calyculata* Schltr., setting this species apart from the others and indicating a distinct metabolic orientation towards primary-derived lipid pathways. Additional classes, such as terpenoids, steroids, phenylpropanoids, and lignans, were broadly shared but varied in their relative abundance, providing a baseline of conserved chemical traits with species-specific quantitative differences.

Subclass and pathway-level analyses further contextualized these observations (Fig. 7A and C). The enrichment of flavones, flavonols, and cinnamic acid derivatives in *V. pompona* points to a strong investment in the shikimate/phenylpropanoid pathway, consistent with its high number of annotated features in this metabolic route. *V. phaeantha* was distinguished by chalcones and carotenoids, reflecting specialized divisions of flavonoid and terpenoid biosynthesis that could reinforce its intermediate yet distinctive chemical profile. In contrast, *V. calyculata* Schltr. displayed a more balanced yet diverse chemical signature, with contributions from diterpenoids, flavones, and fatty acids, aligning with its dual enrichment in fatty acid and terpenoid pathways. Taken together, these results demonstrate that while the three *Vanilla* species share a common chemical composition, they diverge significantly in the intensity and diversification of specific metabolite classes. Such differences reinforce the discriminant statistical markers identified earlier and reveal the broader biosynthetic frameworks that shape inter-species metabolic variation within the genus.

4 Conclusion

This study demonstrates the power of untargeted LC-HRMS/MS metabolomics combined with chemometric modeling to resolve the chemical diversity of Brazilian *Vanilla* species and to uncover robust species-specific metabolic fingerprints. Despite the pronounced intraspecific variability observed in leaf tissues, our integrative analytical strategy successfully discriminated *V. pompona*, *V. phaeantha*, and *V. calyculata* Schltr., revealing consistent multivariate patterns supported by high predictive power in supervised PLS-DA models. In contrast, the absence of biome-specific clustering underscores that environmental origin was not a major determinant of chemical variation under the conditions studied. By integrating multivariate (VIP scores) and univariate (Volcano Plot) criteria, we identified a high-confidence set of discriminant biomarkers, including phenolic acids, cinnamic acid derivatives, C-glycosylated flavonoids, lipids, terpenoids, and nitrogen-containing metabolites. The shared core of 17 markers, particularly ferulic acid (a biosynthetic precursor of vanillin) highlights central metabolic pathways conserved across these species and reinforces the relevance of phenylpropanoid metabolism in shaping *Vanilla* chemotaxonomy. Class-level and pathway-level annotations further revealed distinct metabolic orientations among the studied taxa, advancing our understanding of intra- and interspecific chemical differentiation within the genus. Together, these findings demonstrate that metabolomics offers a robust and a reliable tool for species authentication, chemotaxonomic studies, and the bioprospecting of aromatic and bioactive compounds in *Vanilla*. The molecular insights generated here provide a foundation for biodiversity conservation, germplasm management, and future valorization strategies involving Brazil's native *Vanilla* species. Expanding future analyses to reproductive organs (green and cured pods) will reveal additional, diagnostically powerful metabolites directly linked to vanilla aroma biosynthesis and commercial quality.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11306-026-02422-8>.

Acknowledgements The authors acknowledge financial support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG (APQ-08297-25), and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for institutional and financial support.

Author contributions **GSL: ** Writing – original draft, Methodology, Formal analysis, Conceptualization. **GBB: ** Writing – review & editing, Methodology, Data curation. **HGM: ** Writing – original draft, Methodology, Data curation. **RBNA: ** Writing – review & editing, Conceptualization, Resources, Funding acquisition. **LBB:

** Writing – review & editing, Conceptualization, Resources, Funding acquisition. **LILM: ** Writing – review & editing, Methodology. **BMMGA: ** Writing – review & editing, Methodology. **NML: ** Writing – original draft, Methodology. **RFV: ** Writing – review & editing, Conceptualization, Funding acquisition, Project administration. **BG: ** Writing – review & editing, Resources, Funding acquisition. **GFS: ** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Conceptualization.

Funding The Article Processing Charge (APC) for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614).

Data availability The mass spectrometric raw data supporting the findings will be made available by the authors upon reasonable request, without undue reservation.

Declarations

Competing interests The authors declare no competing interests.

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

References

- Álvarez-Fernández, M. A., Cerezo, A. B., Cañete-Rodríguez, A. M., Troncoso, A. M., & García-Parrilla, M. C. (2015). Composition of nonanthocyanin polyphenols in alcoholic-fermented strawberry products using LC–MS (QTRAP), high-resolution MS (UHPLC–Orbitrap–MS), LC–DAD, and antioxidant activity. *Journal of Agricultural and Food Chemistry*, 63(7), 2041–2051. <https://doi.org/10.1021/jf506076n>
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3), 166–173. <https://doi.org/10.1002/cem.785>
- Beer, F., Weinert, C. H., Wellmann, J., Hillebrand, S., Ley, J. P., Soukup, S. T., & Kulling, S. E. (2025). Comprehensive metabolome characterization of leaves, internodes, and aerial roots of *Vanilla planifolia* by untargeted LC–MS and GC × GC–MS. *Phytochemical Analysis*, 36(1), 30–51. <https://doi.org/10.1002/pca.3414>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Brereton, R. G., & Lloyd, G. R. (2014). Partial least squares discriminant analysis: Taking the magic away. *Journal of Chemometrics*, 28(4), 213–225. <https://doi.org/10.1002/cem.2609>

- Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Analytical Methods*, 6(9), 2812–2831. <https://doi.org/10.1039/C3AY41907J>
- Brunschwig, C., Collard, F. X., Bianchini, J.-P., & Raharivelomanana, P. (2009). Evaluation of chemical variability of cured vanilla beans (*Vanilla tahitensis* and *Vanilla planifolia*). *Natural Product Communications*. <https://doi.org/10.1177/1934578X0900401016>
- Cao, L., Guler, M., Tagirdzhanov, A., Lee, Y.-Y., Gurevich, A., & Mohimani, H. (2021). MolDiscovery: Learning mass spectrometry fragmentation of small molecules. *Nature Communications*, 12(1), Article 3718. <https://doi.org/10.1038/s41467-021-23986-0>
- Chen, L., Khatri, P. K., Paolini, M., Nardin, T., Roncone, A., Larcher, R., et al. (2025). Authentic aroma and compound-specific isotope ratios ($\delta^{13}\text{C}$, $\delta^2\text{H}$) profiles of vanilla pods (*V. planifolia* and *V. tahitensis*). *Molecules*, 30(4), Article 825. <https://doi.org/10.3390/molecules30040825>
- Ciriminna, R., Fidalgo, A., Meneguzzo, F., Parrino, F., Ilharco, L. M., & Pagliaro, M. (2019). Vanillin: The case for greener production driven by sustainability megatrend. *ChemistryOpen*, 8(6), 660–667. <https://doi.org/10.1002/open.201900083>
- da Silva, R. R., Wang, M., Nothias, L.-F., van der Hooft, J. J. J., Caraballo-Rodríguez, A. M., Fox, E., et al. (2018). Propagating annotations of molecular networks using in silico fragmentation. *PLoS Computational Biology*, 14(4), Article e1006089. <https://doi.org/10.1371/journal.pcbi.1006089>
- de Oliveira, R. T., da Silva Oliveira, J. P., & Macedo, A. F. (2022). Vanilla beyond *Vanilla planifolia* and *Vanilla × tahitensis*: Taxonomy and historical notes, reproductive biology, and metabolites. *Plants*, 11(23), Article 3311. <https://doi.org/10.3390/plants11233311>
- de Oliveira, R. T., da Silva Oliveira, J. P., da Silva, A. L. M., Carrao Dantas, E. K., Koblit, M. G. B., Bello, M. L., et al. (2024). Vanilla from Brazilian Atlantic Forest: In vitro and in silico toxicity assessment and high-resolution metabolomic analysis of *Vanilla* spp. ethanolic extracts. *Food Chemistry*, 456, Article 139948. <https://doi.org/10.1016/j.foodchem.2024.139948>
- de Queiroz, J. C. E., Leite, J. R. S. A., & Vasconcelos, A. G. (2023). Prospecting plant extracts and bioactive molecules with antimicrobial activity in Brazilian biomes: A review. *Antibiotics*, 12(3), Article 427. <https://doi.org/10.3390/antibiotics12030427>
- De Vos, R. C., Moco, S., Lommen, A., Keurentjes, J. J., Bino, R. J., & Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 2(4), 778–791. <https://doi.org/10.1038/nprot.2007.95>
- Dibwe, D. F., Takeishi, N., Oba, S., Sakurai, A., Sakurai, T., Tsukui, T., et al. (2023). Identification of a β -carboline alkaloid from chemoselectively derived vanilla bean extract and its prevention of lipid droplet accumulation in human hepatocytes (HepG2). *Molecules*, 28(24), Article 8024. <https://doi.org/10.3390/molecules28248024>
- dos Santos, E. S. B., de Oliveira, T. T. B., Leal, I. L., de Castro, A., Alves, E. C., Nery, T. B. R., & de Souza, A. L. B. (2024). Study and characterization of vanilla pods *Tahitensis* species for characterization for future application in products derived from vanilla in natura. *Journal of Bioengineering, Technologies and Health*, 6(4), 348–353. <https://doi.org/10.34178/jbth.v6i4.333>
- Ernst, M., Kang, K. B., Caraballo-Rodríguez, A. M., Nothias, L.-F., Wandy, J., Chen, C., et al. (2019). MolNetEnhancer: Enhanced Molecular Networks by Integrating Metabolome Mining and Annotation Tools. *Metabolites*, 9(7), Article 144. <https://doi.org/10.3390/metabo9070144>
- Flanagan, N. S., Navia-Samboni, A., González-Pérez, E. N., & Mendieta-Matallana, H. (2022). Distribution and conservation of vanilla crop wild relatives: The value of local community engagement for biodiversity research. *Neotropical Biology and Conservation*, 17(3), 205–227. <https://doi.org/10.3897/neotropic.al.17.e86792>
- Gallage, N. J., & Möller, B. L. (2015). Vanillin–bioconversion and bioengineering of the most popular plant flavor and its de novo biosynthesis in the *Vanilla* orchid. *Molecular Plant*, 8(1), 40–57. <https://doi.org/10.1016/j.molp.2014.11.008>
- Gowda, G. A. N., & Djukovic, D. (2014). Overview of Mass Spectrometry-Based Metabolomics: Opportunities and Challenges. In D. Raftery (Ed.), *Mass Spectrometry in Metabolomics. Methods in Molecular Biology* (Vol. 1198, pp. 3–12). Humana Press. https://doi.org/10.1007/978-1-4939-1258-2_1
- Gu, F., Chen, Y., Hong, Y., Fang, Y., & Tan, L. (2017). Comparative metabolomics in vanilla pod and vanilla bean revealing the biosynthesis of vanillin during the curing process of vanilla. *AMB Express*, 7(1), Article 116. <https://doi.org/10.1186/s13568-017-0413-2>
- Hernández-Peña, R., Lorenzo-Manzanarez, J. L., Cruz-Ramírez, L. A., Reyes-López, D., Hernández-Domínguez, C., Pascual-Ramírez, F., & Ordaz-Ortiz, J. J. (2025). Unravelling Vanillin Biosynthesis: Integrative Transcriptomic and Metabolomic Insights into *Vanilla planifolia* Pod Development. *Journal of Agricultural and Food Chemistry*, 73(30), 19094–19106. <https://doi.org/10.1021/acs.jafc.5c05293>
- Hu, Y., Resende, M. F. R., Bombarely, A., Brym, M., Bassil, E., & Chambers, A. H. (2019). Genomics-based diversity analysis of *Vanilla* species using a *Vanilla planifolia* draft genome and Genotyping-By-Sequencing. *Scientific Reports*, 9(1), Article 3416. <https://doi.org/10.1038/s41598-019-40144-1>
- IBGE. (2019). Instituto Brasileiro de Geografia e Estatística. *Biomassa e sistema costeiro-marinho do Brasil: compatível com a escala 1:250.000*. Rio de Janeiro.
- Kharazian, N., Dehkordi, F. J., & Xiang, C. (2025). Metabolomics-based profiling of five *Salvia* L. (Lamiaceae) species using untargeted data analysis workflow. *Phytochemical Analysis*, 36(1), 113–143. <https://doi.org/10.1002/pca.3423>
- Koch, A. K., de Fraga, C. N., dos Santos, J. U. M., & Ilkiu-Borges, A. L. (2013). Taxonomic notes on *Vanilla* (Orchidaceae) in the Brazilian Amazon, and the description of a new species. *Systematic Botany*, 38(4), 975–981. <https://doi.org/10.1600/036364413X674706>
- Krueger, D. A., & Krueger, H. W. (1985). Krueger, Dana A., and Harold W. Krueger. “Detection of fraudulent vanillin labeled with carbon-13 in the carbonyl carbon.” 33.3 (1985): 323–325. *Journal of Agricultural and Food Chemistry*, 33(3), 323–325.
- Kundu, A. (2017). Vanillin biosynthetic pathways in plants. *Planta*, 245(6), 1069–1078. <https://doi.org/10.1007/s00425-017-2684-x>
- Lê Cao, K.-A., Boitard, S., & Besse, P. (2011). Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12(1), Article 253. <https://doi.org/10.1186/1471-2105-12-253>
- Leyva, V. E., Lopez, J. M., Zevallos-Ventura, A., Cabrera, R., Cañari-Chumpitaz, C., Toubiana, D., & Maruenda, H. (2021). NMR-based leaf metabolic profiling of *V. planifolia* and three endemic *Vanilla* species from the Peruvian Amazon. *Food Chemistry*, 358, Article 129365. <https://doi.org/10.1016/j.foodchem.2021.129365>
- Lima, N. M., Santos, G. F., de Jesus A. S. Andra, T., Dias, L. S., Silva, P. A., Castro, S. B. R., et al. (2025). Metabolic signatures by LC-HRMS/MS of jabuticaba (*Plinia cauliflora*) juice, liqueur, and wines reveal the wealthiest sources of bioactive metabolites. *Talanta*, 287, Article 127602. <https://doi.org/10.1016/j.talanta.2025.127602>
- Mohimani, H., Gurevich, A., Shlemov, A., Mikheenko, A., Korobeynikov, A., Cao, L., et al. (2018). Dereplication of microbial metabolites through database search of mass spectra. *Nature Communications*, 9(1), Article 4035. <https://doi.org/10.1038/s41467-018-06082-8>

- Nadal, M. C., Teixeira, G., Coronado, J. F. O., Ochiutto, A. A., Cunha Neto, A., & Reis, M. (2025). Potential of Brazil's native vanilla species: A study on distribution and research gaps. *Ornamental Horticulture*. <https://doi.org/10.1590/2447-536x.v31.e312874>
- Neto, J. C. R., Vieira, L. R., de Aquino Ribeiro, J. A., de Sousa, C. A. F., Júnior, M. T. S., & Abdelnur, P. V. (2021). Metabolic effect of drought stress on the leaves of young oil palm (*Elaeis guineensis*) plants using UHPLC–MS and multivariate analysis. *Scientific Reports*, 11(1), Article 18271. <https://doi.org/10.1038/s41598-021-97835-x>
- Niculau, E. S., de Freitas, S. D. L., de Sá, I. C. G., Fernandes, J. B., & Silva, M. F. G. F. (2016). da. Quantitative Analysis of Natural Products in Plants by LC–MS. *Revista Virtual de Química*, 8(1). <https://doi.org/10.5935/1984-6835.20160014>
- 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. *Nature Methods*, 17(9), 905–908. <https://doi.org/10.1038/s41592-020-0933-6>
- Palama, T. L., Khatib, A., Choi, Y. H., Payet, B., Fock, I., Verpoorte, R., & Kodja, H. (2009). Metabolic changes in different developmental stages of *Vanilla planifolia* pods. *Journal of Agricultural and Food Chemistry*, 57(17), 7651–7658. <https://doi.org/10.1021/jf901508f>
- Pan, Y.-Y., Chen, Y.-C., Chang, W.-W., Ma, M.-C., & Liao, P.-C. (2021). Visualization of statistically processed LC–MS-based metabolomics data for identifying significant features in a multiple-group comparison. *Chemometrics and Intelligent Laboratory Systems*, 210, Article 104271. <https://doi.org/10.1016/j.chemolab.2021.104271>
- Pansarin, E. R. (2024). Rediscovery and revalidation of the Brazilian endemic *Vanilla schwackeana* Hoehne (Orchidaceae): Its distribution and phylogenetic position. *Plant Ecology and Evolution*, 157(1), 32–41. <https://doi.org/10.5091/plecevo.110331>
- Pansarin, E. R. (2025). *Vanilla lindmaniana* and *V. palmarum* (Orchidaceae) are distinct allopatric species. *Plant Ecology and Evolution*, 158(1), 53–62. <https://doi.org/10.5091/plecevo.134103>
- Pansarin, E. R., & Menezes, E. (2023). A new remarkable *Vanilla* Mill. (Orchidaceae) species endemic to the Espinhaço Range, Brazil: Its phylogenetic position and evolutionary relationships among Neotropical congeners. *PhytoKeys*, 227, 151–165. <https://doi.org/10.3897/phytokeys.227.101963>
- Perera, C. O., & Owen, E. (2010). Effect of tissue disruption by different methods followed by incubation with hydrolyzing enzymes on the production of vanillin from Tongan Vanilla beans. *Food and Bioprocess Technology*, 3(1), 49–54. <https://doi.org/10.1007/s11947-007-0048-4>
- Phapale, P., Palmer, A., Gathungu, R. M., Kale, D., Brügger, B., & Alexandrov, T. (2020, November 22). Public LC-Orbitrap-MS/MS Spectral Library for Metabolite Identification. <https://doi.org/10.1101/2020.11.21.392266>
- Salem, M. A., de Perez Souza, L., Serag, A., Fernie, A. R., Farag, M. A., Ezzat, S. M., & Alseekh, S. (2020). Metabolomics in the context of plant natural products research: From sample preparation to metabolite analysis. *Metabolites*, 10(1), Article 37. <https://doi.org/10.3390/metabo10010037>
- Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3), 779–787. <https://doi.org/10.1021/ac051437y>
- Tomadoni, B., Viacava, G., Cassani, L., Moreira, M. R., & Ponce, A. (2016). Novel biopreservatives to enhance the safety and quality of strawberry juice. *Journal of Food Science and Technology*, 53(1), 281–292. <https://doi.org/10.1007/s13197-015-2068-9>
- Triba, M. N., Le Moyec, L., Amathieu, R., Goossens, C., Bouchemal, N., Nahon, P., et al. (2015). PLS/OPLS models in metabolomics: The impact of permutation of dataset rows on the K-fold cross-validation quality parameters. *Molecular BioSystems*, 11(1), 13–19. <https://doi.org/10.1039/C4MB00414K>
- van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V., & Rogers, S. (2016). Topic modeling for untargeted substructure exploration in metabolomics. *Proceedings of the National Academy of Sciences*, 113(48), 13738–13743. <https://doi.org/10.1073/pnas.1608041113>
- Vogt, T. (2010). Phenylpropanoid biosynthesis. *Molecular Plant*, 3(1), 2–20. <https://doi.org/10.1093/mp/ssp106>
- Wang, J., An, W., Wang, Z., Zhao, Y., Han, B., Tao, H., et al. (2024). Vanillin has potent antibacterial, antioxidant, and anti-inflammatory activities in vitro and in mouse colitis induced by multidrug-resistant *Escherichia coli*. *Antioxidants*, 13(12), Article 1544. <https://doi.org/10.3390/antiox13121544>
- 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. *Nature Biotechnology*, 34(8), 828–837. <https://doi.org/10.1038/nbt.3597>
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1)
- Worley, B., & Powers, R. (2012). Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1), 92–107. <https://doi.org/10.2174/2213235X11301010092>
- Xia, J., Broadhurst, D. I., Wilson, M., & Wishart, D. S. (2013). Translational biomarker discovery in clinical metabolomics: An introductory tutorial. *Metabolomics*, 9(2), 280–299. <https://doi.org/10.1007/s11306-012-0482-9>
- Xiao, J. F., Zhou, B., & Ressom, H. W. (2012). Metabolite identification and quantitation in LC–MS/MS-based metabolomics. *TrAC Trends in Analytical Chemistry*, 32, 1–14. <https://doi.org/10.1016/j.trac.2011.08.009>
- Xu, W., Huang, J., Wang, P., Yang, Y., Fu, S., Ying, Z., & Zhou, Z. (2024). Using widely targeted metabolomics profiling to explore differences in constituents of three *Bletilla* species. *Scientific Reports*, 14(1), Article 23873. <https://doi.org/10.1038/s41598-024-74204-y>
- Zhang, Y., Mo, L., Chen, F., Lu, M., Dong, W., Wang, Q., et al. (2014). Optimized production of vanillin from green vanilla pods by enzyme-assisted extraction combined with pre-freezing and thawing. *Molecules*, 19(2), 2181–2198. <https://doi.org/10.3390/molecules19022181>
- Zhou, B., Xiao, J. F., Tuli, L., & Ressom, H. W. (2012). LC–MS-based metabolomics. *Molecular BioSystems*, 8(2), 470–481. <https://doi.org/10.1039/C1MB05350G>

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