Food Research International 233 (2026) 118954

Contents lists available at ScienceDirect

Food Research International

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

Volatile fingerprinting and interpretable machine learning for authenticating New Zealand monofloral honeys

Rushan Lakshitha [a], Kevin Kantono [a], Tony Chen [a], Thao T. Le [a], Swapna Gannabathula [b], Nazimah Hamid [a,*]

[a] AUT Centre for Future Foods, School of Science, Faculty of Health and Environment Sciences, Auckland University of Technology, Auckland, New Zealand
[b] The Experiment Company, Auckland, New Zealand

ARTICLE INFO

Keywords:
New Zealand honey
Volatile biomarkers
HS-SPME/GC–MS
Machine learning
Random forest
SHAP
PLS-DA
Authenticity
Traceability

ABSTRACT

Authenticating monofloral honeys is essential for protecting premium markets and ensuring traceability. This study applied an integrated analytical and explainable machine-learning workflow to identify volatile biomarkers for four New Zealand monofloral honeys: thyme, mānuka, kānuka, and clover. Twenty-two samples were profiled using headspace solid-phase microextraction gas chromatography–mass spectrometry (HS-SPME/GC–MS), yielding 122 tentative volatiles across aldehydes, alcohols, acids, esters, terpenoids, and phenolics. Data were analysed using ANOVA, hierarchical clustering, Partial Least Squares Discriminant Analysis (PLS-DA), and Random Forest classification with SHapley Additive exPlanations (SHAP). ANOVA and heatmap analysis revealed honey-specific volatile modules, while PLS-DA confirmed clear supervised separation of floral types. A rule-based SHAP framework was implemented to select biomarkers that were not only influential in the model but also directionally consistent and chemically distinctive. Thyme honey was characterised by short-chain fatty acids and oxygenated terpenoids; mānuka by methoxyacetophenones and benzofuran/methoxylated benzoates; kānuka by anisole-type aromatics and bicyclic monoterpenes; and clover by phenylpropanoid-related aldehydes, fusel alcohols, and linalool-oxide derivatives. These panels achieved non-overlapping group separation and near-perfect cross-validated performance (micro-average ROC-AUC = 0.995). This combined HS-SPME/GC–MS and RF-SHAP approach provides a transparent, statistically supported route to defining interpretable volatile biomarkers, offering a scalable framework for honey authentication, quality assurance, and traceability, and helping safeguard the premium positioning of New Zealand monofloral honeys in global markets.

1. Introduction

Honey, produced by honeybees (*Apis mellifera*), is a complex natural plant-derived product not only as a food, but also for its medicinal and cosmetic uses (Farooqui & A Farooqui, 2011). Botanically, honey is commonly classified as blossom (floral) honey or honeydew honey (Chessum et al., 2022). New Zealand remains a global leader in premium monofloral honey by export value, driven largely by the antibacterial reputation of mānuka honey and strong certification branding. However, this leadership is under pressure because the sector has faced oversupply, price erosion, and governance challenges since the 2020 peak, with export volumes declining by approximately 26% and production falling by about 56% by 2023. These challenges highlight the urgent need for robust, science-based authentication methods, such as biomarker discovery, to protect market integrity and support the New Zealand Honey Strategy Coriolis, 2024–2030 (Coriolis, 2024). This study focuses on four monofloral honeys, namely mānuka (*Leptospermum scoparium*), kānuka (*Kunzea ericoides*), thyme (*Thymus vulgaris*), and clover (*Trifolium repens*).

Chemically, honey is an intricate matrix containing diverse bioactive compounds, such as flavonoids, phenolic acids, and is a rich source of volatile organic compounds (VOCs) that are responsible for its aroma, taste, and bioactivities. VOCs originate from floral nectar, bee-derived enzymes, and post-harvest transformations, and together form a characteristic chemical "fingerprint" of botanical and regional origin (Kaškonienė et al., 2008; Manyi-Loh et al., 2011). Identifying robust biomarkers within these volatile fingerprints is therefore central to authenticating honey, safeguarding quality, and distinguishing closely related floral sources in premium markets (Kaškonienė et al., 2008; Manyi-Loh et al., 2011).

* Corresponding author at: AUT Centre for Future Foods, Auckland University of Technology, Private Bag 92006, New Zealand.
E-mail address: nazimah.hamid@aut.ac.nz (N. Hamid).

https://doi.org/10.1016/j.foodres.2026.118954
Received 20 December 2025; Received in revised form 12 March 2026; Accepted 15 March 2026
Available online 17 March 2026

Historically, melissopalynology has been used to determine floral origin. However, this approach has critical limitations that directly support volatile biomarker discovery. For example, mānuka and kānuka honeys are very difficult to distinguish by pollen due to morphological similarities and overlapping bloom periods, and pollen abundance does not necessarily reflect nectar contribution to the final product (McDonald et al., 2018; Schmidt et al., 2021). Furthermore, NMR, though non-destructive and reproducible, is costly and may miss trace chemical compounds due to signal overlap and lower sensitivity, limiting its effectiveness in honey discrimination (Gerhardt et al., 2018). In contrast, gas chromatography–mass spectrometry (GC–MS) enables sensitive detection of trace volatiles and has demonstrated superior differentiation between closely related floral sources like mānuka and kānuka (Beitlich et al., 2014).

Analysing the volatile profile of honey provides valuable insights for verifying authenticity, ensuring quality, and distinguishing between different products. Gas chromatography-mass spectrometry (GC–MS) remains the benchmark technique for volatile analysis in honey (Breschi et al., 2024). Headspace solid-phase microextraction (HS-SPME), a solvent-free and highly sensitive technique, is particularly suitable for capturing low-abundance compounds (Liang et al., 2023). Combined HS-SPME/GC–MS facilitates efficient characterisation of major aroma-active volatiles, such as aldehydes, alcohols, esters, acids, terpenes, and phenolics, and has been widely used to differentiate monofloral honeys and to relate volatile patterns to sensory attributes (Karabagias et al., 2019; Nascimento et al., 2024). Despite widespread adoption, many published studies have focused on enumerating compounds or applying univariate statistics (e.g., ANOVA) to highlight mean differences across floral sources (Karabagias, 2022; Soria et al., 2009; Wolski et al., 2006). Univariate approaches reveal single-compound effects but overlook multivariate dependencies and co-occurrence structures intrinsic to botanical origin.

To address these gaps, multivariate and machine-learning methods have increasingly been used to model high-dimensional chemical fingerprints. Partial least squares discriminant analysis (PLS-DA) can summarise group-level separation by maximising covariance between X (volatiles) and Y (class labels); however, PLS-DA may be sensitive to noise and prone to overfitting, and often offers limited interpretability when honeys are botanically distinct yet chemically similar (Chen et al., 2017; Taiti et al., 2023). In contrast, ensemble machine-learning models such as Random Forest (RF) provide improved predictive performance, built-in feature selection, and robustness to noisy data, which are characteristics well-suited to natural product variability and complex volatile matrices (Barragán-Hernández et al., 2024; Cardinal et al., 2020).

A further important factor to consider for honey biomarker discovery is interpretability. Beyond accurate classification, stakeholders require transparent rationale for decisions, traceable chemical panels, and biologically plausible pathways that support authenticity claims. SHapley Additive exPlanations (SHAP) is a model-agnostic interpretability tool that quantifies each feature's contribution to individual predictions and to global model behaviour, enabling consistent identification of the compounds most responsible for class separation (Fu et al., 2024; Rodríguez-Pérez & Bajorath, 2019). SHAP is increasingly used in food authenticity/traceability to interpret tree-based models and rank influential variables, yet it is typically applied as a global importance metric or to generate top-k feature lists. Here, we advance SHAP from interpretation to biomarker discovery using a class-specific, rule-based framework that emphasises directional consistency and within-class prevalence, yielding more robust, interpretable markers for authentication. (Kang et al., 2024; Yu et al., 2025; Zhang & Abdulla, 2023). In New Zealand, exported mānuka honey is authenticated using a five-attribute definition (four nectar chemicals plus mānuka DNA), highlighting the manuka-kānuka ambiguity. Therefore, untargeted HS-SPME-GC–MS and machine learning provide a complementary fingerprinting approach beyond targeted marker tests used in regulatory authentication (Ministry for Primary, 2025).

The present study uses HS-SPME coupled with GC–MS to sensitively resolve headspace volatiles from four New Zealand monofloral honeys (thyme, mānuka, kānuka, and clover) and combines complementary statistical and machine learning frameworks, including univariate ANOVA, multivariate clustering, PLS-DA, and Random Forest with SHAP-based interpretation, to identify explainable volatile biomarkers of botanical origin. To our knowledge, this is the first study to integrate HS-SPME/GC–MS with Random Forest and SHAP to discover volatile biomarkers in New Zealand honeys. This comprehensive approach addresses the limitations of melissopalynology, such as pollen ambiguity and the lack of correlation between pollen and nectar, as well as the constraints of NMR, including limited sensitivity and spectral overlap. At the same time, it exploits the proven sensitivity of GC–MS for detecting trace volatile compounds and the analytical strengths of HS-SPME for solvent-free headspace capture. More importantly, the use of SHAP provides interpretable panels that reflect coherent biochemical modules, such as phenylpropanoid-derived aromatics, monoterpenoid ketones and alcohols, and aldehydes, aligning with established knowledge of honey volatile biogenesis and aroma chemistry. This study establishes scientifically robust biomarker panels that enhance honey authentication, reinforce quality assurance practices, and help maintain the global reputation of New Zealand's premium honey industry.

2. Materials and methods

2.1. Honey samples and experiment design

A total of 22 biological honey samples representing four New Zealand monofloral honey types [thyme ($n = 13$), mānuka ($n = 3$), kānuka (n = 3) and clover (n = 3)] were analysed in this study (Supplementary Table S1). Each sample was measured in technical triplicate, resulting in 66 observations. Raw and commercial samples (Supplementary Table S1) were included to capture natural variability and to ensure that the proposed biomarker-based authentication approach is robust across different processing conditions, making it applicable to both unprocessed and market-ready honeys commonly found in retail and export markets. According to the supplier, the raw honeys were obtained directly from carefully selected and trusted beekeepers, with traceability records and harvest information ensuring accurate botanical origin. The commercial honeys were procured from certified suppliers, and all samples were labelled as monofloral based on the origin region, floral source, and beekeepers' harvesting records. Details for each honey sample, including floral type, processing status (raw or commercial), year of collection and geographical origin, are provided in Supplementary Table S1. The sample set was unbalanced, with a higher representation of thyme honey samples ($n = 13$). This design was motivated by multiple factors. First, our previously published data demonstrated that New Zealand thyme honey possesses the highest antioxidant activity among local monofloral honeys (George et al., 2025), so this type was prioritised for deeper chemical and biomarker analysis. Second, mānuka and kānuka honeys are difficult to differentiate via melissopalynology because of overlapping flowering periods and morphologically similar pollen. Lastly, clover honey was included as a widely available, low-value commercial benchmark to enhance model relevance.

2.2. HS-SPME-GC/MS analysis of volatile compounds and data processing

Benzophenone (internal standard) and the n-alkane series (C9-C25) for retention index calibration were purchased from Sigma-Aldrich (St. Louis, MO, USA). Sodium chloride (NaCl) and methanol (MeOH) used for sample preparation were of analytical grade. Volatiles were extracted using headspace solid-phase microextraction (HS-SPME) following adapted methods from previous studies (Bianchi et al., 2011; Makowicz et al., 2019). Briefly, 1.5 g of honey was mixed with 3 mL of a 30% (*w/v*)

NaCl solution in a 10 mL GC vial, spiked with 20 μL of 1 ppm benzophenone (internal standard). Samples were equilibrated at 40 °C for 10 min with stirring. A DVB/CAR/PDMS SPME fibre (50/30 μm) was exposed to the headspace for 40 min at 40 °C and desorbed at 250 °C for 5 min in the GC inlet. Analyses were performed using an Agilent 7890 GC system coupled with a 5977B MSD and a DB-WAX capillary column (30 m × 0.25 mm × 0.25 μm). The oven was programmed from 40 °C (3 min) to 240 °C at 5 °C/min, held for 5 min. The carrier gas was helium at 1.0 mL/min. Mass spectra were acquired in Electron Ionisation mode (70 eV), scanning from *m/z* 35–400. Linear retention indices were calculated using the n-alkane series (C9-C25) analysed under the same GC conditions. Volatile compound identification was tentative and based on both mass spectral matching (≥80% similarity using the NIST17 library) and retention index (RI) confirmation. Compound intensities were reported as relative abundance using peak area ratios relative to the internal standard (benzophenone, 1 ppm). Data extraction and quantification were carried out using Agilent GC–MS software. The final data matrix comprised 66 rows (technical replicates) × retained volatiles ISTD intensities and was used for downstream univariate, multivariate, and machine-learning analyses.

2.3. Statistical and machine learning analysis in Python

All analyses were conducted in Python (version 3.13) using packages including pandas, NumPy, statsmodels, scikit-learn, SHAP, matplotlib, seaborn, and openpyxl. Data preprocessing, statistical modelling, and figure generation were implemented through scripted workflows to ensure full reproducibility.

2.3.1. One-way fixed-effects ANOVA

Because the design was unbalanced and included repeated technical replicates within biological samples, we averaged the replicates to get sample-level means and analysed one mean per biological sample per group (separate, not nested). For each compound, we fitted a one-way ANOVA with honey type as a fixed effect using OLS and Type-II sums of squares, then assessed the group effect with the F-test (statsmodels). Fisher's LSD post-hoc comparisons ($\alpha = 0.05$) were carried out using the ANOVA residual Mean Square Error (MSE) and residual degrees of freedom (df), and summarised groups by mean ± SD with compact letter displays. Post-hoc Fisher's LSD was applied only when the ANOVA was significant ($\alpha = 0.05$), and each group contributed ≥3 samples; otherwise, no pairwise letters were assigned. Compounds were ranked by ANOVA *p*-value, and the top 50 were selected for visualisation and reporting; complete outputs (F and *p* values, per-group summaries, and letters) were exported to Excel. Complete ANOVA results for all compounds, including descriptive statistics, F and *p*-values, and post-hoc group letters where applicable, were also exported to Excel (Supplementary Table S2).

2.3.2. Data transformation

For all downstream multivariate analyses (hierarchical clustering, PLS-DA, Random Forest classification and SHAP), internal-standard-normalised GC–MS peak area ratios were preprocessed using a consistent workflow. Technical replicates were averaged to obtain one intensity profile per biological honey sample. Zero values (signals below the detection limit) were replaced with half of the smallest non-zero peak area. The resulting data matrix was then cube-root transformed to reduce the influence of highly abundant volatiles and stabilise variance, followed by autoscaling to ensure all variables contributed on a comparable scale to subsequent multivariate modelling and visualisation.

2.3.3. Agglomerative clustering (Heatmap based on Top-50 ANOVA features)

A clustered heatmap was generated using the top 50 compounds ranked by one-way ANOVA *p*-values with Fisher's LSD post-hoc comparisons (replicates collapsed to sample means). Rows represent compounds and columns represent samples, which were hierarchically clustered using Ward's linkage on Euclidean distance. Column colours indicate honey type. All figures were exported as high-resolution PNG and SVG formats.

2.3.4. PLS-DA biplot with top 25 loadings

PLS-DA was applied to evaluate group-level separation among honey types. Components were fitted using sklearn.cross_decomposition. PLSRegression, and the biplot displays PLS1 versus PLS2 scores with 95% confidence ellipses derived from score covariance, along with the top loadings represented as arrows. Axis labels indicate the proportion of variance explained in X and Y for each component. Model performance was assessed using stratified five-fold cross-validation at the sample level (replicates collapsed), and the optimal number of components was determined based on cross-validated classification accuracy using out-of-fold predictions. All plots were exported as high-resolution PNG and SVG formats.

2.4. Random forest classification and SHAP-based biomarker discovery

Random Forest (RF) classification was performed using sklearn. Ensemble.RandomForestClassifier with 500 trees to assess discriminative performance between honey types. Model interpretability was obtained via SHapley Additive exPlanations using the shap library: a probability explainer was fit on the trained RF to compute per-feature SHAP values for each honey class. For each class, the Top-10 SHAP features were ranked by mean positive SHAP across the class samples (global importance). Then a stricter biomarker filter was applied to define top 10 biomarkers: (i) mean SHAP >0.0015, (ii) positive SHAP in all target class samples (support = n of that class), and (iii) the feature's central tendency is highest in the target class (dominant abundance).

For each honey type, a standardised four-panel SHAP figure was generated: (1) Top-10 SHAP features, (2) Top-10 biomarkers after filtering, and (3–4) the same two panels coloured by feature intensity to show value-dependence. Model robustness was assessed via stratified 5-fold cross-validation at the Sample level (replicates collapsed), generated out-of-fold ROC curves and AUC (one-vs-rest plus micro/macro) to verify performance above chance, produced a learning curve (accuracy vs training size) to diagnose bias/variance. Biomarker names were truncated for clarity, and all final figures were exported in high-resolution PNG and SVG formats, and the biomarker summary table was exported as a CSV/XLSX file.

3. Results and discussion

The total ion chromatograms (TICs) obtained from HS-SPME-GC–MS analysis of the 22 New Zealand honey samples revealed a complex mixture of VOCs. After initial spectral analysis and elimination of common impurities and noise peaks, 122 volatile features were retained across all honey types. Total ion chromatograms illustrating the volatile profiles of Thyme, Mānuka, Kānuka, and Clover honey are provided in Supplementary Fig. S3. Tentative identification was achieved by mass spectral matching (≥80% similarity) and retention index confirmation against the NIST17 library.

To ensure both statistical support and interpretability, we used a complementary analysis workflow. ANOVA provides compound-wise evidence for group differences, while hierarchical clustering and the heatmap provide an unsupervised view of intrinsic sample structure and co-occurrence modules independent of class labels. PLS-DA provides a supervised low-dimensional projection to visualise separation and identify covariance-driven drivers. Random Forest complements this by modelling non-linear class boundaries and feature interactions among correlated volatiles under stratified cross-validation, while SHAP explains Random Forest predictions via feature-level contributions to class probability.

3.1. One-way fixed-effects ANOVA

Fixed-effects ANOVA results and their full discussion are provided in the Supplementary Information (Supplementary Table S2). Briefly, this analysis identified the most discriminatory volatiles across honey types and provided statistical support for the floral-origin differences observed in the dataset. The ranked outputs were used to guide downstream visualisation and interpretation. Key ANOVA findings are referenced below to support and contextualise the PLS-DA and Random Forest/SHAP results.

3.2. Multivariate analysis of volatile fingerprints: Clustered heatmap

The clustered heatmap results and accompanying discussion are presented in the Supplementary Information (Supplementary Fig. S1). The heatmap illustrates honey-type clustering consistent with floral origin and highlights co-occurrence blocks of compounds that underpin group separation at the fingerprint level. These unsupervised patterns are referenced alongside the PLS-DA and Random Forest/SHAP findings to demonstrate concordance between exploratory structure and supervised classification, supporting that the observed separation is chemically interpretable rather than model-driven.

3.3. Partial least squares discriminant analysis (PLS-DA) analysis of honey volatiles

The PLS-DA biplot (Fig. 1) shows clear supervised separation of thyme, mānuka, kānuka, and clover honeys, with PLS1 explaining 42.4% of X-variance (40.5% of Y) and PLS2 explaining 9.0% of X-variance (26.5% of Y). Model performance improved from 0.59 with one component to 0.72 with two components, reaching 1.00 at three components and remaining stable thereafter (Supplementary Fig. S4), confirming robust discrimination under stratified 5-fold cross-validation.

Vectors pointing strongly toward the thyme cluster along PLS1 include hexanoic acid (caproic acid), butanoic acid (butyric acid), eugenol (2-methoxy-4-(prop-2-en-1-yl)phenol), the terpenoid diol 2,6-dimethylocta-3,7-diene-2,6-diol, and the aromatic ketone 2-aminoacetophenone. All were significantly higher in thyme ($p < 0.001$; Supplementary Table S2) and co-occur in the thyme group of the clustered heatmap (Supplementary Fig. S1). This explains the pronounced positive displacement of thyme on PLS1. Unlike univariate analysis, which evaluates compounds independently, PLS-DA leverages inter-variable covariance, allowing features with modest individual differences to contribute strongly as part of a discriminant pattern. Short-chain acids and terpenoid alcohols are commonly reported in thyme honeys and impart herbal/spicy notes (Alissandrakis et al., 2009; Castro-Vázquez et al., 2009; Karabagias et al., 2014).

Mānuka samples cluster in the lower-left quadrant, driven primarily by negative PLS1 scores, with PLS2 providing additional separation from kānuka. Although characteristic mānuka volatiles such as o-methoxyacetophenone, 4′-hydroxyacetophenone, and benzofuran derivatives (e.g., benzofuran-2-carbaldehyde) do not dominate the top loading arrows, their shared covariance positions mānuka in a distinct space. This pattern aligns with the ANOVA-selected clustered heatmap (Supplementary Fig. S1), which revealed a compact methoxylated-aromatic group including methyl 3,5-dimethoxybenzoate and benzofuran-2-carbaldehyde, significantly higher in mānuka ($p < 0.001$; Supplementary Table S2). The negative PLS1/PLS2 placement therefore reflects a coherent phenolic group rather than isolated compounds. In supervised space, this group pulls mānuka toward negative PLS2 (away from kānuka's anisole and bicyclic ketone profile) and negative PLS1 (opposite thyme). Together, the heatmap and PLS-DA biplot present a consistent picture showing that mānuka is defined by a suite of correlated phenolic and heterocyclic volatiles, supporting multi-compound panel-based authentication rather than reliance on a single marker (Díaz-Galiano et al., 2023; Hegazi et al., 2022). This distributed discriminatory power explains why no single phenolic dominates the top loading rankings in the PLS-DA model, making the collective profile more robust and reliable than any individual compound.

Kānuka's distinct chemical signature is defined by anisole-type aromatics and bicyclic monoterpenes that drive its separation along positive PLS2 and negative PLS1 scores. Two of the longest loading vectors,

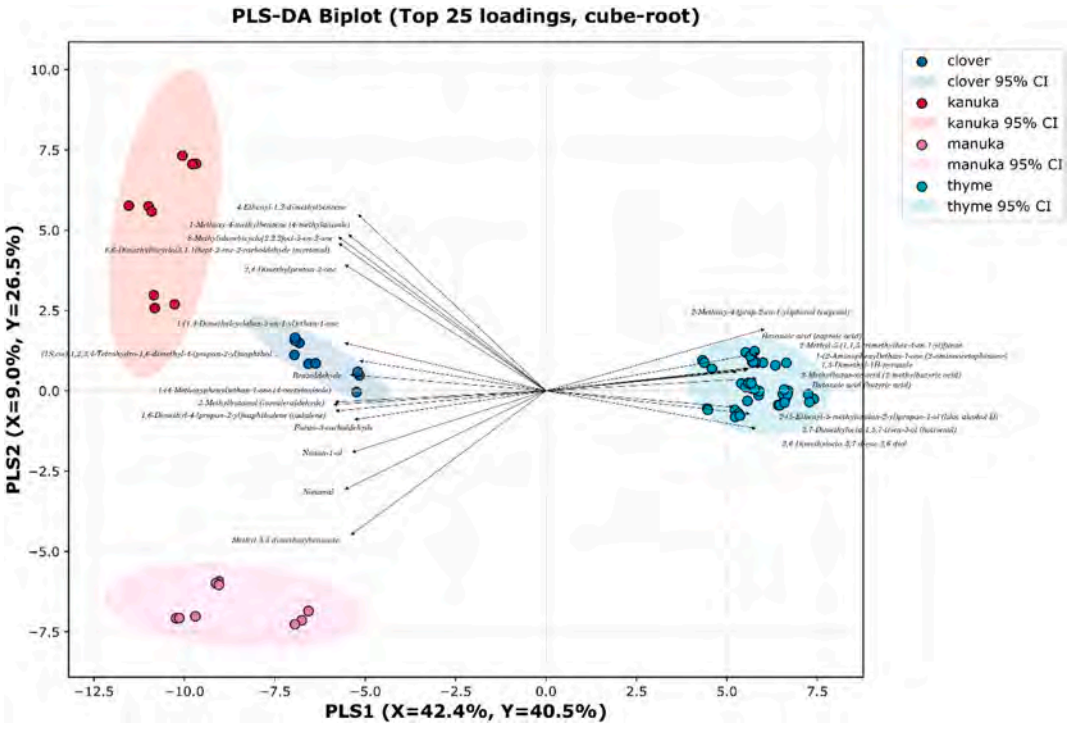

Fig. 1. Partial Least Squares Discriminant Analysis (PLS-DA) biplot of cube-root transformed, auto-scaled volatile profiles for thyme, mānuka, kānuka, and clover honeys.

8-methylidenebicyclo[2.2.2]oct-5-en-2-one and 6,6-dimethylbicyclo [3.1.1]hept-2-ene-2-carbaldehyde (myrtenal), together with 1-methoxy-4-methylbenzene (4-methylanisole) and 1-(4-methoxyphenyl) ethan-1-one (4-acetylanisole), point toward kānuka, capturing the monoterpenoid and anisole/alkylbenzene profile that discriminates this floral type. All four compounds are among the ANOVA-selected Top 50 volatiles (Supplementary Fig. S1) and show significant differences between honey types ($p < 0.001$; Supplementary Table S2), where they co-occur in a kānuka-specific aromatic group. Bicyclic/spiro ketones are characteristic of Myrtaceae secondary metabolism, and anisole-type aromatics (e.g., *p*-methylanisole, acetylanisoles) are well documented in *Kunzea* essential oil (Fuller et al., 2022; Khambay et al., 2003). These compounds co-vary strongly across kānuka samples, collectively driving its distinct separation along PLS2 and PLS1. This agreement between PLS-DA loadings, ANOVA significance, and the clustered heatmap confirms that kānuka's separation reflects a coherent chemical profile rather than numerical artefacts.

Clover samples cluster on the negative side of PLS1 with intermediate PLS2 scores. Among the top loading vectors, 3-methylbutanal (isovaleraldehyde) projects toward clover, while compounds with smaller loading weights, such as 2-(5-ethenyl-5-methyloxolan-2-yl) propan-2-ol (trans-linalool oxide, furanoid) and octanal, contribute in the same direction. All three volatiles were significantly higher in clover ($p < 0.001$; Supplementary Table S2), and the ANOVA-selected clustered heatmap (Supplementary Fig. S1) reveals a clover-specific aldehyde and oxygenated-monoterpene group that includes octanal and structurally related compounds. These compounds are well documented in clover honeys: 3-methylbutanal as a Strecker aldehyde linked to amino acid degradation (Jerković et al., 2016; Karabagias et al., 2019), trans-linalool oxide as a floral aroma contributor (Jerković et al., 2016; Machado et al., 2020), and octanal as a lipid-derived aldehyde common in Trifolium honeys (Karabagias et al., 2019; Machado et al., 2020).

Although PLS-DA (Fig. 1) demonstrates strong supervised separation in a low-dimensional projection, it is primarily a covariance-based visualisation tool and may not capture non-linear class boundaries or higher-order interactions in complex volatile fingerprints. We therefore complemented PLS-DA with Random Forest classification to model non-linear relationships among correlated volatiles and to quantify predictive performance under stratified 5-fold cross-validation at the sample level. SHAP was then applied to explain Random Forest predictions by attributing class-probability contributions to individual compounds, enabling conservative, class-consistent biomarker selection.

3.4. Random forest classification and SHAP interpretation

3.4.1. Model performance

The Random Forest classifier discriminated the four honey types very well under stratified 5-fold cross-validation. The micro-average Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) was 0.995 (95% CI: 0.982–1.000), with class-specific AUCs of 1.000 for clover, thyme, and mānuka, and 0.965 for kānuka (Supplementary Fig. S8). Precision and recall were uniformly high for clover and thyme (1.00/1.00), slightly lower for mānuka (0.75/1.00), and reduced for kānuka (1.00/0.67), reflecting the small sample size and class imbalance (Supplementary Fig. S6). The confusion matrix (Supplementary Fig. S7) shows occasional misclassification of kānuka as clover, while all other classes were perfectly recovered. The learning curve (Supplementary Fig. S5) plateaued near 0.88–0.90 cross-validated accuracy, indicating strong signal but residual variance likely due to limited kānuka representation. Overall, these metrics confirm that Random Forest provides robust predictive performance and clear compositional separation among honey types.

3.4.2. SHAP global feature importance

To interpret the Random Forest model, SHapley Additive exPlanations (SHAP), a model-agnostic approach that quantifies each feature's contribution to predictions was applied. Global SHAP values were computed for all volatiles, ranking features by their mean positive impact on class probability across samples. Fig. 2 (panels 2 A-2D) illustrates the top 10 SHAP features for each honey type, alongside colour-coded plots showing how feature intensity influences SHAP contribution.

The global SHAP rankings reveal that thyme honey is primarily associated with short-chain fatty acids (e.g., butanoic and hexanoic acids), oxygenated terpenoids, and phenolic aromatics. This is consistent with previous reports of thyme honeys being rich in lipid-derived acids and terpene alcohols (Alissandrakis et al., 2009; Karabagias et al., 2014; Wiese et al., 2018). Mānuka shows strong contributions from methoxyacetophenones, benzofuran derivatives, and methoxylated benzoates, aligning with established *Leptospermum* markers (Beitlich et al., 2014; Daher & Gülaçar, 2010; Díaz-Galiano et al., 2023; Hegazi et al., 2022; Oelschlaegel et al., 2012). Kānuka is driven by anisole-type aromatics and bicyclic monoterpenes, in agreement with *Kunzea* essential-oil chemotypes and previous honey studies (Beitlich et al., 2014; Fuller et al., 2022; Khambay et al., 2003; Lewe et al., 2023; Maddocks, 2021). Clover's top SHAP features include branched Strecker aldehydes, fusel alcohols, and linalool-oxide derivatives, consistent with *Trifolium* honey literature (Jerković et al., 2016; Karabagias et al., 2019; Machado et al., 2020). These patterns confirm that the Random Forest model captures chemically coherent volatile blocks rather than isolated markers, aligning with known floral chemistry and supporting robust interpretation.

3.4.3. Biomarker filtering criteria

While global SHAP rankings identify features that strongly influence model predictions, not all high-ranking variables are suitable as authenticity biomarkers. Some compounds may appear important due to statistical interactions or shared occurrence across multiple honey types rather than being uniquely characteristic of a single floral origin. To address this, a rule-based filtering framework was applied to refine SHAP outputs into interpretable, class-specific biomarker panels.

Rather than listing the top SHAP features, a rule-based filter to identify robust authenticity biomarkers. A compound was retained only if it satisfied three criteria:

(i) mean SHAP value greater than 0.0015 for the target class.
(ii) positive SHAP contribution in all samples of that class (full within-class support)
(iii) the highest median abundance in the target class compared to all other classes.

These conditions ensured that selected biomarkers are not only influential in the Random Forest model but also chemically distinctive and consistently present in the honey type they represent. In contrast, raw top-ten SHAP features, while mathematically important, often fail these filters because they include ubiquitous volatiles or compounds whose predictive strength depends on interactions rather than on unique abundance. Applying these rules removes spurious predictors and yields biomarker panels that are statistically robust, biologically plausible, and aligned with known floral chemistry.

Filtering SHAP outputs using abundance and consistency rules bridges the gap between statistical importance and practical authenticity, aligning honey biomarker discovery with best practices established in other complex food matrices (Marcílio & Eler, 2020). Previous honey studies using chemometrics or machine learning have typically reported top-ranked features without applying such interpretability filters, which risks selecting ubiquitous or interaction-driven predictors (Agila & Barringer, 2012; Chen et al., 2017; Gerhardt et al., 2018; Karabagias et al., 2020; Langford et al., 2012; Marcílio & Eler, 2020; Soria et al., 2009).

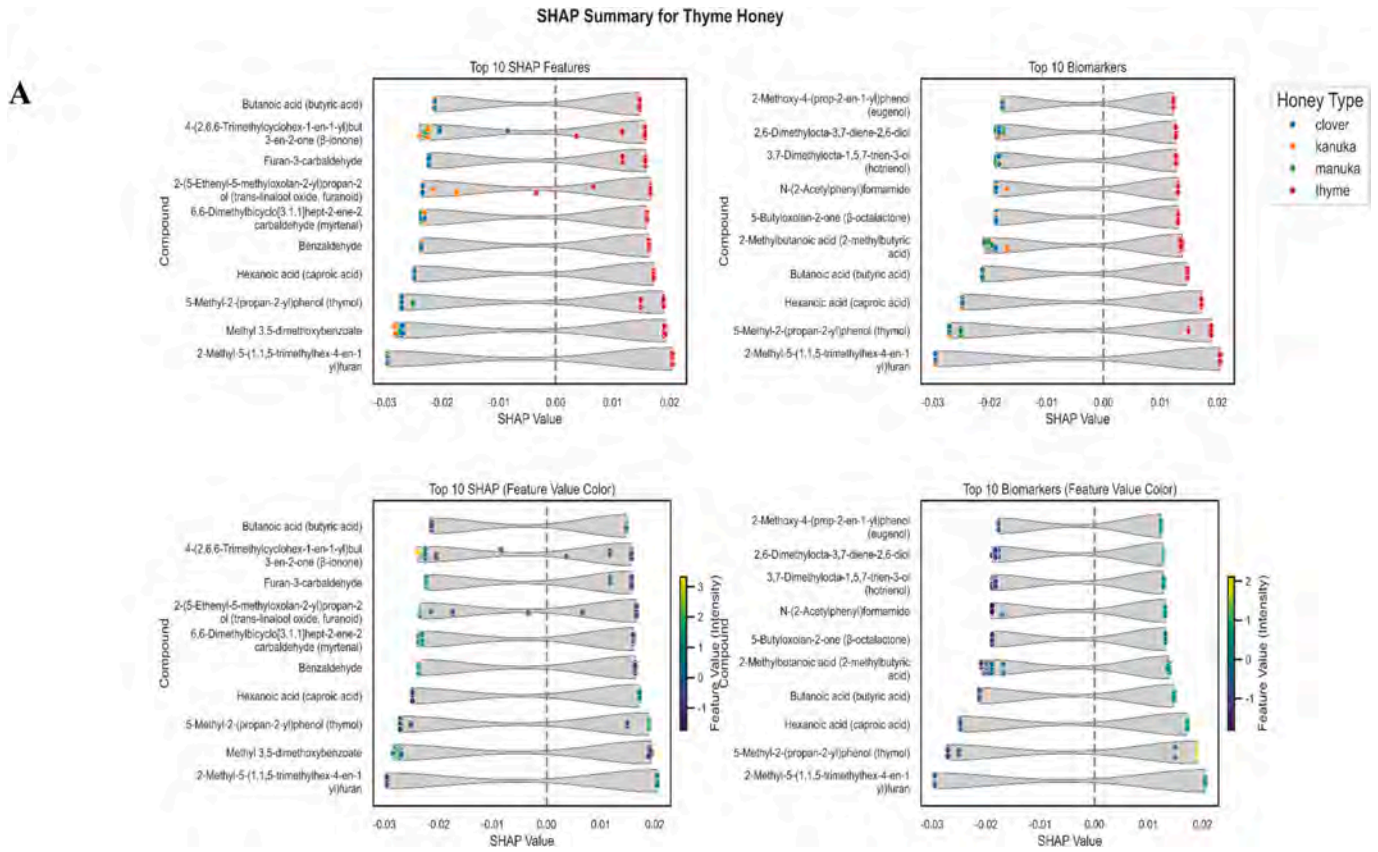

Fig. 2. SHAP-based interpretation of Random Forest classification for New Zealand monofloral honeys. Each panel displays the top 10 SHAP-ranked features (left) and the filtered top 10 biomarker candidates (right), with colour-coded versions below showing feature intensity. Panels: (A) Thyme honey; (B) Mānuka honey; (C) Kānuka honey; (D) Clover honey.

3.4.4. *Honey-specific biomarker panels*

Fig. 2 summarises the SHAP-based interpretation of Random Forest classification for the four honey types. Each subfigure shows the global top ten SHAP features on the left and the filtered biomarker set after applying interpretability criteria on the right, with colour-coded versions below indicating feature intensity. The filtered panels represent compounds meeting all three criteria: high mean SHAP, positive SHAP across all class samples, and highest median abundance in the target class, providing a transparent link between model behaviour and chemically plausible biomarkers.

3.4.4.1. Thyme honey biomarkers. The authentication of thyme honey is based on the Top 10 SHAP biomarkers panel, which applies stricter selection criteria: (i) mean SHAP >0.0015, (ii) positive SHAP for all thyme samples, and (iii) the highest median feature value in thyme. The RF-SHAP analysis (Fig. 2A) presents the top 10 SHAP features and biomarkers for thyme honey, highlighting volatiles supported by literature on thyme nectar and phytochemistry. The highest-ranked biomarker is the terpenoid-derived furan, 2-methyl-5-(1,1,5-trimethylhex-4-en-1-yl) furan, which shows uniformly positive SHAP values for thyme, is highly significant in ANOVA ($p < 0.001$; Supplementary Table S2), and appears in thyme-directed loading vectors in the PLS-DA Top 25 biplot. Furan derivatives have been reported in thyme and other floral honeys as secondary volatiles linked to terpene degradation (Alissandrakis et al., 2009; Karabagias et al., 2019; Wiese et al., 2018), and their formation can involve oxidative cleavage of monoterpenes or Maillard-type reactions during honey storage (Xing & Yaylayan, 2024).

The second-ranked biomarker, thymol (5-methyl-2-(propan-2-yl) phenol), shows positive SHAP values for all thyme honey samples. Despite its borderline ANOVA *p*-value ($p = 0.096$), it appears in thyme-directed loading vectors in the PLS-DA Top 25 biplot. Thymol is a key phenolic marker, reinforcing its role as a signature compound for thyme honey aroma and authenticity (Alissandrakis et al., 2009; Beitlich et al., 2014; Karabagias et al., 2019). The short-chain fatty acid cluster, including butanoic acid, hexanoic acid, heptanoic acid, and 2-methylbutanoic acid, contributes buttery, cheesy, and slightly rancid notes characteristic of thyme honey (Alissandrakis et al., 2009; Karabagias et al., 2014). Two further biomarkers, γ-octalactone (5-butyloxolan-2-one) and N-(2-acetylphenyl)formamide, add chemical diversity. γ-Octalactone, a cyclic ester with sweet, creamy notes, likely arises from lipid oxidation ((Alissandrakis et al., 2009; Karabagias et al., 2019), while N-(2-acetylphenyl)formamide is produced possibly through Maillard-type reactions or phenylpropanoid intermediates (Glagoleva et al., 2022; Xing & Yaylayan, 2024; Yadav et al., 2020). Oxygenated monoterpenoids such as hotrienol (2,6-dimethylocta-3,7-diene-2,6-diol) and 3,7-dimethylocta-1,5,7-trien-3-ol are derived from *Thymus vulgaris* nectar metabolism and impart herbal and floral notes (Alissandrakis et al., 2009; Jerković et al., 2016; Karabagias et al., 2019; Machado et al., 2020; Wiese et al., 2018). Finally, eugenol (2-methoxy-4-(prop-2-en-1-yl)phenol) carries strong SHAP contributions and has been reported alongside thymol as a distinctive aroma constituent in thyme honeys (Escriche et al., 2022).

Comparison of RF-SHAP biomarkers with PLS-DA Top 25 loadings shows strong agreement for core thyme volatiles. The terpenoid-derived furan, oxygenated monoterpenoids (hotrienol, 3,7-dimethylocta-1,5,7-trien-3-ol), and short-chain fatty acids (butanoic acid, hexanoic acid) all load toward the thyme cluster in multivariate space and appear in the ANOVA Top 50, confirming their dual role as variance-driven separators

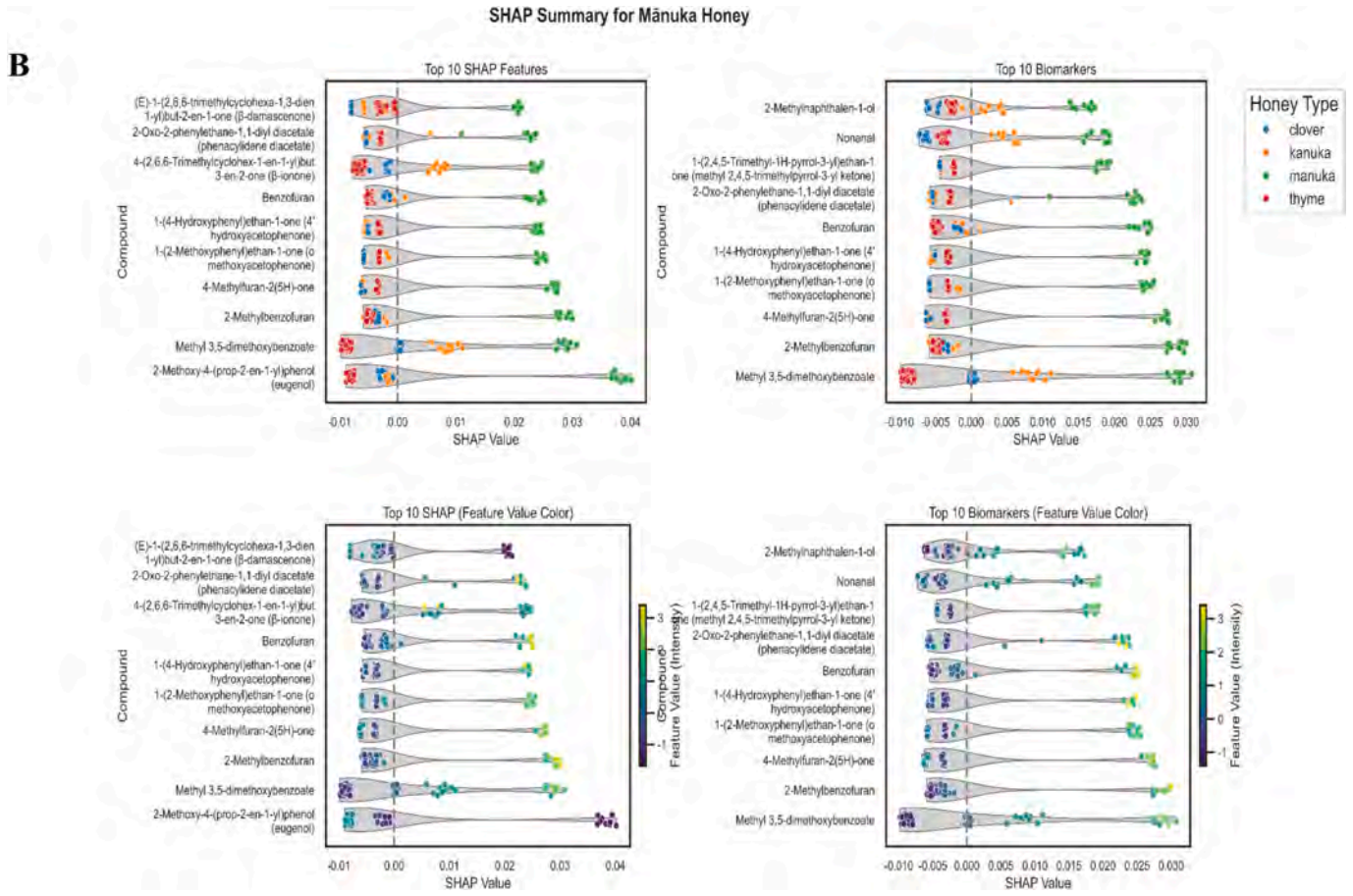

Fig. 2. (*continued*).

and predictive features. Conversely, several SHAP-only biomarkers, such as N-(2-acetylphenyl)formamide and γ-octalactone, were absent from the PLS-DA Top 25, illustrating how machine-learning attribution can reveal botanically specific volatiles overlooked by covariance-based models. This highlights the value of combining SHAP with PLS-DA and ANOVA to construct robust, interpretable biomarker panels for thyme honey authentication.

3.4.4.2. Mānuka honey biomarkers. The RF-SHAP analysis (Fig. 2B) identified a panel of volatiles that consistently distinguish mānuka honey from clover, kānuka, and thyme. These biomarkers are dominated by acetophenones, methoxylated benzoates, benzofuran/naphthalenol derivatives, and selected heterocycles, forming a chemically coherent methoxylated-aromatic module. Comparison with the ANOVA-selected Top 50 and the clustered heatmap (Supplementary Fig. S1) confirms that these compounds co-locate within the mānuka-associated cluster, reflecting their higher abundance in mānuka samples and reinforcing their discriminatory strength.

Acetophenones and methoxylated benzoates represent the characteristic mānuka volatile profile. Compounds such as o-methoxyacetophenone (1-(2-methoxyphenyl)ethan-1-one), 4′-hydroxyacetophenone (1-(4-hydroxyphenyl)ethan-1-one), and methyl 3,5-dimethoxybenzoate consistently show the highest abundance in mānuka, strong SHAP contributions, and prominent PLS-DA loadings. These compounds are widely recognised as signature authenticity markers for Leptospermum honeys (Beitlich et al., 2014; Szafnauer, 2023) and have been confirmed in previous GC–MS studies (Daher & Gülaçar, 2010). Benzofuran derivatives, notably 2-methylbenzofuran, also contribute strongly to the mānuka profile and cluster within the methoxylated-aromatic module, reinforcing their discriminatory strength in multivariate models.

Several biomarkers apart from the known volatile fingerprint of mānuka honey were found. 4-methylfuran-2(5H)-one, a butenolide associated with Maillard chemistry, was ANOVA-significant and retained as a biomarker. While butenolides have been reported in honey (Stephens et al., 2010), this specific compound has not been widely cited as a mānuka marker. Similarly, 2-oxo-2-phenylethane-1,1-diyl diacetate (phenacylidene diacetate), a furanyl diacetate, was ANOVA-significant ($p < 0.001$), retained in the ANOVA Top 50, and showed consistently positive SHAP values across mānuka replicates. To our knowledge, this compound has not been previously reported in mānuka honey or Leptospermum spp., highlighting its novelty. An N-heterocyclic ketone, 1-(2,4,5-trimethyl-1H-pyrrol-3-yl)ethan-1-one (methyl 2,4,5-trimethylpyrrol-3-yl ketone), also met biomarker criteria and represents a novel pyrrole derivative, extending the pyrrole-related aroma chemistry documented for Leptospermum honeys (Chan et al., 2013) and highlighting the contribution of Maillard-type reactions. Additional contributors include nonanal, a C_9 aldehyde with a fresh, fatty-floral aroma, and 2-methylnaphthalen-1-ol, a polycyclic aromatic alcohol aligning with the methoxy/phenolic aromatic profile reported for Leptospermum volatiles (Daher & Gülaçar, 2010). While nonanal has not been previously identified as a distinctive mānuka marker, its higher abundance in mānuka samples and SHAP effect suggest a role in mānuka's characteristic odour.

Several high-ranking SHAP features, such as eugenol (2-methoxy-4-(prop-2-en-1-yl)phenol) and Maillard-type furanone-enone ketones (e.g., 1-(furan-2-yl)but-2-en-1-one), were excluded from the biomarker panel because they occur in other honeys via nectar inputs or heat-related sugar degradation (Manyi-Loh et al., 2011). This filtering step highlights the importance of combining machine-learning attribution with statistical and biological plausibility when defining authenticity markers.

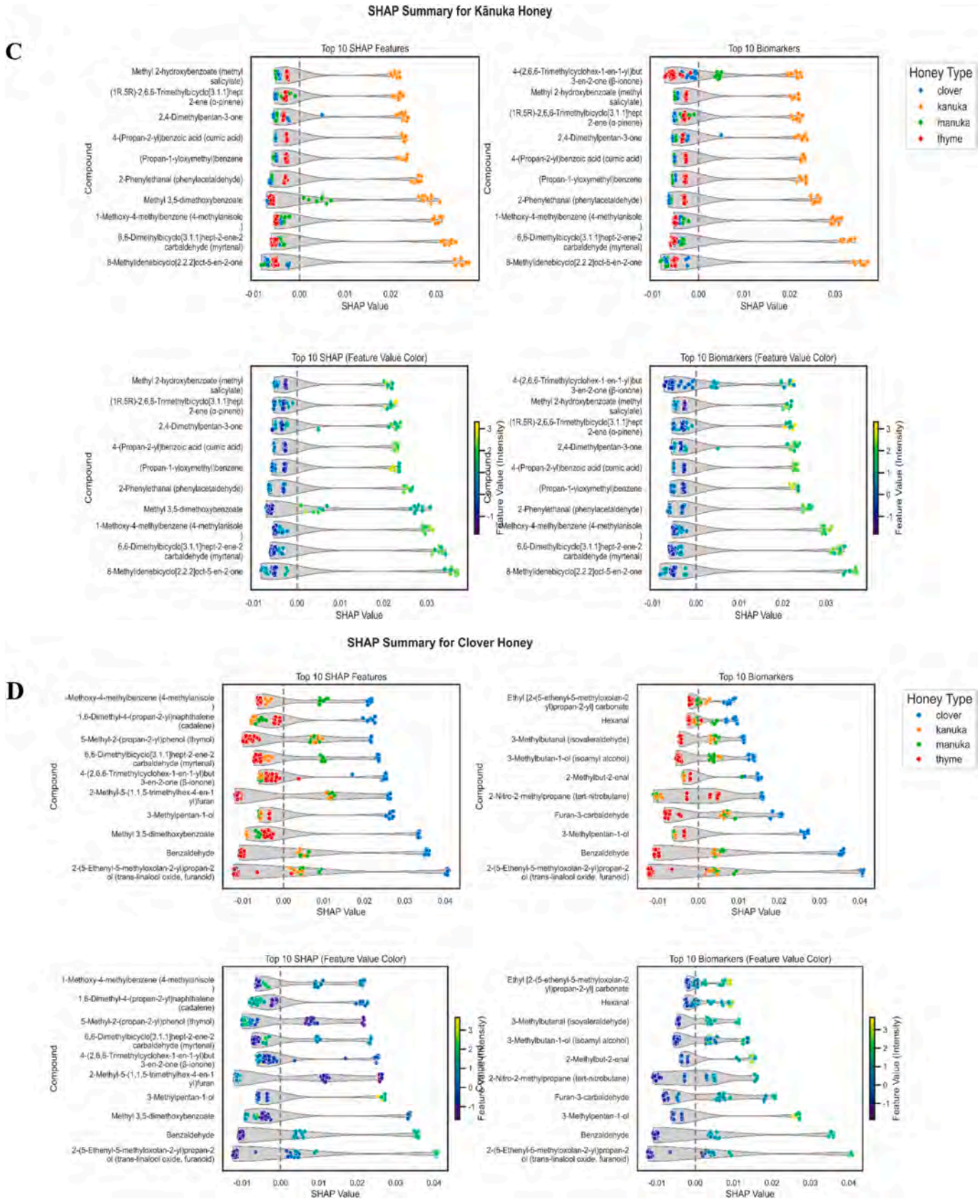

Fig. 2. (*continued*).

Comparison of RF-SHAP biomarkers with PLS-DA Top 25 loadings shows strong agreement for characteristic mānuka volatiles. Acetophenones (o-methoxyacetophenone, 4′-hydroxyacetophenone), methoxylated benzoates (methyl 3,5-dimethoxybenzoate), and benzofuran derivatives (2-methylbenzofuran) all load toward the mānuka cluster in multivariate space and appear in the ANOVA Top 50, confirming their dual role as variance-driven separators and predictive features. Conversely, several SHAP-only biomarkers, such as phenacylidene diacetate, methyl 2,4,5-trimethylpyrrol-3-yl ketone, and nonanal, were absent from the PLS-DA Top 25, illustrating how machine learning can uncover botanically specific volatiles that are overlooked by covariance-based models.

3.4.4.3. Kānuka honey biomarkers. The RF-SHAP analysis (Fig. 2C) identified a panel of ten volatiles that consistently distinguish kānuka honey from clover, thyme, and mānuka. All biomarkers were statistically significant in ANOVA ($p < 0.001$) and retained in the ANOVA-selected Top 50 volatiles used for the clustered heatmap (Supplementary Fig. S1), where they co-locate in a kānuka-enriched cluster. This clustering pattern confirms that these compounds rise and fall together across kānuka replicates, forming a coherent chemical module. Comparison with PLS-DA Top 25 loadings shows strong convergence for core kānuka volatiles: methyl salicylate, cumic acid, and the bicyclic terpenoids load toward the kānuka cluster in multivariate space, validating their dual role as variance-driven separators and predictive features.

Terpenoid-derived compounds dominated the biomarker set. 8-methylidenebicyclo[2.2.2]oct-5-en-2-one and (1*R*,2*R*,5*R*)-2,6,6-trimethylbicyclo[3.1.1]hept-2-ene (a bicyclic monoterpene related to α-pinene) showed consistently positive SHAP values and high mean abundances in kānuka samples. These resinous terpenoids cluster with other essential oil volatiles in the heatmap, reflecting the *Kunzea* chemotype reported in essential oil studies (Fuller et al., 2022). At the top of the SHAP ranking, the norisoprenoid 4-(2,6,6-trimethylcyclohex-1-en-1-yl)but-3-en-2-one (a β-ionone derivative) provided the strongest SHAP effect and was ANOVA-significant ($p < 0.001$), consistent with carotenoid degradation products that impart floral-fruity notes (Machado et al., 2020).

Aromatic derivatives formed a second major group. 4-(propan-2-yl) benzoic acid (cumic acid), 2,4-dimethylbenzaldehyde, (propan-2-yloxymethyl)benzene, and 4-ethenyl-1,2-dimethylbenzene all showed strong SHAP contributions and clustered with benzenoid compounds in the heatmap, consistent with phenylpropanoid metabolism. Methyl 2-hydroxybenzoate (methyl salicylate) ranked near the top of the SHAP list and is widely recognised as a chemotaxonomic marker for *Myrtaceae* honeys, including kānuka (Beitlich et al., 2014). Two additional contributors consolidate the phenylpropanoid module: 2-phenylethanol (phenethyl alcohol), a floral alcohol derived from phenylalanine, showed positive SHAP values and grouped with benzenoids in the heatmap.

Several biomarkers identified, such as methyl salicylate and terpenoid derivatives, are consistent with Kunzea essential oil chemistry and have been previously associated with *Myrtaceae* honeys (Beitlich et al., 2014; Fuller et al., 2022). Others, including certain bicyclic ketones and aromatic ethers, have not been widely documented in kānuka honey, suggesting they represent novel additions to its volatile profile. Taken together, the RF-SHAP biomarker panel includes terpenoids, norisoprenoids, benzenoid acids, and aromatic ethers that provides a robust multivariate basis for distinguishing kānuka from other New Zealand monofloral honeys.

3.4.4.4. Clover honey biomarkers. The global SHAP ranking for clover honey (Fig. 2D, top-left) identified ten volatiles as the most influential predictors in the Random Forest model. However, not all these high-ranking features were retained as authenticity biomarkers. After applying biomarker rules, requiring the highest median abundance in clover, honey consistently positive SHAP values across replicates, and statistical significance in ANOVA, only three of the global Top 10 SHAP features remained: benzaldehyde, 2-(5-ethenyl-5-methyloxolan-2-yl) propan-2-ol (trans-linalool oxide, furanoid), and 3-methylpentan-1-ol. These compounds satisfied all criteria, with benzaldehyde and trans-linalool oxide also appearing in the ANOVA-selected Top 50 panel (Supplementary Table S2; $p < 0.001$) and loading toward clover honey in the PLS-DA biplot (Fig. 2D). Benzaldehyde was consistently dominant in clover honey, aligning with the phenylpropanoid pathway expected for *Trifolium* nectar and widely reported in clover honey literature (Jerković et al., 2016; Machado et al., 2020). *trans*-Linalool oxide, a floral monoterpenoid derivative, is a well-established clover marker (Jerković et al., 2016), while 3-methylpentan-1-ol, a branched fusel alcohol linked to Ehrlich pathway catabolism, represents a rarely reported compound in honey volatiles, suggesting a novel addition to the clover fingerprint (Karabagias et al., 2019; Smit et al., 2009).

The remaining biomarkers in the SHAP biomarker panel (Fig. 2D, top-right), including 3-methylbutanal, 2-methylbut-2-enal, furan-3-carbaldehyde, hexanal, and ethyl [2-(5-ethenyl-5-methyloxolan-2-yl) propan-2-yl] carbonate, were drawn from outside the global SHAP Top 10 but met the biomarker rules. All were ANOVA-significant ($p \leq 0.027$; Supplementary Table S2), with furan-3-carbaldehyde and 2-methylbut-2-enal also appearing in the ANOVA Top 50 panel. Compounds such as furan-3-carbaldehyde and prenyl-type enals are not commonly documented in clover honey profiling studies (Jerković et al., 2016), although they are recognised as Maillard/Strecker products in honey chemistry more broadly (Machado et al., 2020; Xing & Yaylayan, 2024). Their detection here suggests a novel extension of the clover volatile fingerprint. In contrast, some global SHAP Top 10 features (e.g., 4-methyl-1-(methylenecyclopropyl)benzene) were excluded because they lacked clover-highest median or ANOVA support.

A tentative nitroalkane, tert-nitrobutane [2-nitro-2-methylpropane], was observed as a tentatively annotated feature in the SHAP ranking for clover. Given that nitroalkanes are not typical plant-derived nectar volatiles, this assignment should be interpreted cautiously and may reflect background/artefact or contamination (e.g., analytical, environmental, or packaging sources). Therefore, this feature is not proposed as a confirmed botanical biomarker without targeted confirmation (authentic standard).

Hexanal was identified as a clover biomarker primarily through the RF-SHAP approach. Although it was ANOVA-significant ($p = 0.006$; Supplementary Table S2), it did not enter the ANOVA top-50 panel or the PLS-DA top-25 loading set. While hexanal itself has not been widely highlighted in clover honey literature, a structurally related C_8 aldehyde, octanal, has been reported in clover honey volatiles, supporting the plausibility of lipid-derived aldehydes as contributors to *Trifolium* honey aroma (Karabagias et al., 2019; Zamora et al., 2015).

2-(5-ethenyl-5-methyloxolan-2-yl)propan-2-ol (trans-linalool oxide, furanoid) and furan-3-carbaldehyde were both ANOVA-significant ($p < 0.001$; Supplementary Table S2), with clover honey having the highest mean for each; both compounds were retained in the ANOVA-selected Top-50 panel, although neither appeared among the PLS-DA Top-25 loadings. While direct reports of furan-3-carbaldehyde in clover honey are lacking, the detection of trans-linalool oxide in *Trifolium* honeys in previous studies (Jerković et al., 2016; Machado et al., 2020) supports the plausibility of this chemical subgroup occurring together in clover, potentially reflecting common floral-derived metabolic pathways.

The clover honey biomarker panel includes branched-chain aldehydes and alcohols, benzenoids, straight-chain aldehydes, and terpenoid furanic derivatives, forming a chemically coherent profile consistent with *Trifolium* floral chemistry and providing a robust basis for authenticity discrimination.

4. Conclusion

This study demonstrates that combining HS-SPME/GC–MS with

interpretable machine learning provides a powerful and transparent approach for authenticating New Zealand monofloral honeys. By integrating univariate ANOVA, hierarchical clustering, PLS-DA, and Random Forest with SHAP-based interpretation, chemically coherent biomarker panels that deliver robust, non-overlapping separation of thyme, mānuka, kānuka, and clover honeys were identified. These panels reflect biologically plausible modules, short-chain fatty acids and terpenoids for thyme, methoxyacetophenones and benzofuran derivatives for mānuka, anisole-type aromatics and bicyclic monoterpenes for kānuka, and phenylpropanoid-related aldehydes and linalool oxides for clover, rather than isolated markers. The rule-based SHAP framework advances beyond conventional "top-k" feature lists by enforcing directional consistency and within-class prevalence, ensuring biomarkers are both statistically and chemically meaningful. Near-perfect classification performance (micro-average ROC-AUC = 0.995) confirms the reliability of this integrated workflow. Beyond academic significance, these findings offer a scalable solution for honey authentication, quality assurance, and traceability, supporting the integrity and premium positioning of New Zealand honeys in global markets. Importantly, robust authentication frameworks strengthen consumer trust and contribute to sustainable industry practices by reducing fraud, protecting biodiversity-linked products, and supporting transparent supply chains. Because compound identities are primarily based on library matching and retention-index confirmation, and abundances are reported as internal-standard-normalised peak area ratios, targeted confirmation and absolute quantification (authentic standards and appropriate controls) are recommended for regulatory-grade deployment. Moreover, although cross-validation indicates strong internal generalisation, independent external validation across regions, seasons/years, and producers is required before broad commercial or regulatory translation. Future work should expand seasonal and geographic coverage, validate the prioritised driver panels in independent external cohorts, and explore sensory correlations to reinforce regulatory and commercial applications.

CRediT authorship contribution statement

Rushan Lakshitha: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kevin Kantono:** Writing – review & editing, Writing – original draft, Supervision, Software, Methodology, Data curation. **Tony Chen:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis. **Thao T. Le:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Swapna Gannabathula:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Nazimah Hamid:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2026.118954.

Data availability

Data will be made available on request.

References

Agila, A., & Barringer, S. (2012). Application of selected ion flow tube mass spectrometry coupled with chemometrics to study the effect of location and botanical origin on volatile profile of unifloral American honeys. *Journal of Food Science, 77*(10), C1103–C1108.

Alissandrakis, E., Tarantilis, P. A., Pappas, C., Harizanis, P. C., & Polissiou, M. (2009). Ultrasound-assisted extraction gas chromatography–mass spectrometry analysis of volatile compounds in unifloral thyme honey from Greece. *European Food Research and Technology, 229*, 365–373.

Barragán-Hernández, W., López-Campos, Ó., Aalhus, J. L., & Prieto, N. (2024). Using machine-learning approaches to investigate the volatile-compound fingerprint of fishy off-flavour from beef with enhanced healthful fatty acids. *Meat Science, 218*, Article 109643.

Beitlich, N., Koelling-Speer, I., Oelschlaegel, S., & Speer, K. (2014). Differentiation of manuka honey from kanuka honey and from jelly bush honey using HS-SPME-GC/MS and UHPLC-PDA-MS/MS. *Journal of Agricultural and Food Chemistry, 62*(27), 6435–6444.

Bianchi, F., Mangia, A., Mattarozzi, M., & Musci, M. (2011). Characterization of the volatile profile of thistle honey using headspace solid-phase microextraction and gas chromatography–mass spectrometry. *Food Chemistry, 129*(3), 1030–1036.

Breschi, C., Ieri, F., Calamai, L., Miele, A., D'Agostino, S., Melani, F., … Cecchi, L. (2024). HS-SPME-GC-MS analysis of the volatile composition of Italian honey for its characterization and authentication using the genetic algorithm. *Separations, 11*(9), 266.

Cardinal, M., Chaussy, M., Donnay-Moreno, C., Cornet, J., Rannou, C., Fillonneau, C., Prost, C., Baron, R., & Courcoux, P. (2020). Use of random forest methodology to link aroma profiles to volatile compounds: Application to enzymatic hydrolysis of Atlantic salmon (Salmo salar) by-products combined with Maillard reactions. *Food Research International, 134*, Article 109254.

Castro-Vázquez, L., Díaz-Maroto, M. C., González-Viñas, M. A., & Pérez-Coello, M. S. (2009). Differentiation of monofloral citrus, rosemary, eucalyptus, lavender, thyme and heather honeys based on volatile composition and sensory descriptive analysis. *Food Chemistry, 112*(4), 1022–1030.

Chan, C. W., Deadman, B. J., Manley-Harris, M., Wilkins, A. L., Alber, D. G., & Harry, E. (2013). Analysis of the flavonoid component of bioactive New Zealand mānuka (Leptospermum scoparium) honey and the isolation, characterisation and synthesis of an unusual pyrrole. *Food Chemistry, 141*(3), 1772–1781.

Chen, H., Jin, L., Fan, C., & Wang, W. (2017). Non-targeted volatile profiles for the classification of the botanical origin of Chinese honey by solid-phase microextraction and gas chromatography-mass spectrometry combined with chemometrics. *Journal of Separation Science, 40*(22), 4377–4384.

Chessum, K. J., Chen, T., Hamid, N., & Kam, R. (2022). A comprehensive chemical analysis of New Zealand honeydew honey. *Food Research International, 157*, Article 111436.

Coriolis, R. (2024). *nextHoney 2024: After the goldrush*. Coriolis Limited.

Daher, S., & Gülaçar, F. O. (2010). Identification of new aromatic compounds in the New Zealand Manuka honey by gas chromatography-mass spectrometry. *Journal of Chemistry, 7*, S7–S14.

Díaz-Galiano, F. J., Heinzen, H., Gómez-Ramos, M. J., Murcia-Morales, M., & Fernández-Alba, A. R. (2023). Identification of novel unique mānuka honey markers using high-resolution mass spectrometry-based metabolomics. *Talanta, 260*, Article 124647.

Escriche, I., Juan-Borrás, M., Visquert, M., Asensio-Grau, A., & Valiente, J. M. (2022). Volatile profile of Spanish raw citrus honey: The best strategy for its correct labeling. *Journal of Food Processing and Preservation, 46*(6), Article e16200.

Farooqui, T., & A Farooqui, A. (2011). Health benefits of honey: Implications for treating cardiovascular diseases. *Current Nutrition & Food Science, 7*(4), 232–252.

Fu, Q., Wu, Y., Zhu, M., Xia, Y., Yu, Q., Liu, Z., Ma, X., & Yang, R. (2024). Identifying cardiovascular disease risk in the US population using environmental volatile organic compounds exposure: A machine learning predictive model based on the SHAP methodology. *Ecotoxicology and Environmental Safety, 286*, Article 117210.

Fuller, I. D., de Lange, P. J., Burgess, E. J., Sansom, C. E., van Klink, J. W., & Perry, N. B. (2022). Chemical diversity of kānuka: Inter-and intraspecific variation of foliage terpenes and flavanones of Kunzea (Myrtaceae) in Aotearoa/New Zealand. *Phytochemistry, 196*, Article 113098.

George, E. M., Gannabathula, S., Lakshitha, R., Liu, Y., Kantono, K., & Hamid, N. (2025). Antibacterial properties, arabinogalactan proteins, and bioactivities of New Zealand honey. *Antioxidants, 14*(4), 375.

Gerhardt, N., Birkenmeier, M., Schwolow, S., Rohn, S., & Weller, P. (2018). Volatile-compound fingerprinting by headspace-gas-chromatography ion-mobility spectrometry (HS-GC-IMS) as a benchtop alternative to 1H NMR profiling for assessment of the authenticity of honey. *Analytical Chemistry, 90*(3), 1777–1785.

Glagoleva, A. Y., Vikhorev, A. V., Shmakov, N. A., Morozov, S. V., Chernyak, E. I., Vasiliev, G. V., … Shoeva, O. Y. (2022). Features of activity of the phenylpropanoid biosynthesis pathway in melanin-accumulating barley grains. *Frontiers in Plant Science, 13*, Article 923717.

Hegazi, N. M., Elghani, G. E. A., & Farag, M. A. (2022). The super-food Manuka honey, a comprehensive review of its analysis and authenticity approaches. *Journal of Food Science and Technology, 59*(7), 2527–2534.

Jerković, I., Radonić, A., Kranjac, M., Zekić, M., Marijanović, Z., Gudićc, S., & Kliškić, M. (2016). Red clover (*Trifolium pratense* L.) honey: Volatiles chemical-profiling and unlocking antioxidant and anticorrosion capacity. *Chemical Papers, 70*(6), 726–736.

Kang, X., Zhao, Y., Yao, L., & Tan, Z. (2024). Explainable machine learning for predicting the geographical origin of Chinese oysters via mineral elements analysis. *Current Research in Food Science, 8*, Article 100738.

Karabagias, I. K. (2022). Headspace volatile compounds fluctuations in honeydew honey during storage at in-house conditions. *European Food Research and Technology, 248*(3), 715–726.
Karabagias, I. K., Badeka, A., Kontakos, S., Karabournioti, S., & Kontominas, M. G. (2014). Characterization and classification of Thymus capitatus (L.) honey according to geographical origin based on volatile compounds, physicochemical parameters and chemometrics. *Food Research International, 55*, 363–372.
Karabagias, I. K., Badeka, A., & Kontominas, M. G. (2020). A decisive strategy for monofloral honey authentication using analysis of volatile compounds and pattern recognition techniques. *Microchemical Journal, 152*, Article 104263.
Karabagias, I. K., Karabagias, V. K., & Badeka, A. V. (2019). The honey volatile code: A collective study and extended version. *Foods, 8*(10), 508.
Kaškonienė, V., Venskutonis, R., & Čeksteryte, V. (2008). Composition of volatile compounds of honey of various floral origin and beebread collected in Lithuania. *Food Chemistry, 111*, 988–997.
Khambay, B. P. S., Beddie, D. G., Hooper, A. M., & Simmonds, M. S. J. (2003). Isolation, characterisation and synthesis of an insecticidal tetramethyltetrahydrochromenedione-spiro-bicyclo [3.1. 1] cycloheptane from two species of Myrtaceae. *Tetrahedron, 59*(36), 7131–7133.
Langford, V., Gray, J., Foulkes, B., Bray, P., & McEwan, M. J. (2012). Application of selected ion flow tube-mass spectrometry to the characterization of monofloral New Zealand honeys. *Journal of Agricultural and Food Chemistry, 60*(27), 6806–6815.
Lewe, N., Young, M., Vorster, J., Paenga, B., Skinner, D., Harcourt, N., de Lange, P., Haira, T., Blockley-Powell, S., & Munkacsi, A. (2023). Comparison of chemical profiles of Kānuka (Kunzea robusta de Lange & Toelken, Myrtaceae) essential oils. *Phytochemistry Letters, 56*, 50–56.
Liang, D., Wen, H., Zhou, Y., Wang, T., Jia, G., Cui, Z., & Li, A. (2023). Simultaneous qualitative and quantitative analyses of volatile components in Chinese honey of six botanical origins using headspace solid-phase microextraction and gas chromatography–mass spectrometry. *Journal of the Science of Food and Agriculture, 103*(15), 7631–7642.
Machado, A. M., Miguel, M. G., Vilas-Boas, M., & Figueiredo, A. C. (2020). Honey volatiles as a fingerprint for botanical origin—A review on their occurrence on monofloral honeys. *Molecules, 25*(2), 374.
Maddocks, W. A. (2021). Diversity in the essential oil of New Zealand grown Kānuka, *Kunzea ericoides* (A. Rich) joy Thomps. *American Journal of Essential Oils and Natural Products, 9*, 32–38.
Makowicz, E., Jasicka-Misiak, I., Teper, D., & Kafarski, P. (2019). Botanical origin authentication of polish phacelia honey using the combination of volatile fraction profiling by HS-SPME and lipophilic fraction profiling by HPTLC. *Chromatographia, 82*, 1541–1553.
Manyi-Loh, C. E., Ndip, R. N., & Clarke, A. M. (2011). Volatile compounds in honey: A review on their involvement in aroma, botanical origin determination and potential biomedical activities. *International Journal of Molecular Sciences, 12*(12), 9514–9532.
Marcílio, W. E., & Eler, D. M. (2020). *From explanations to feature selection: Assessing SHAP values as feature selection mechanism* (pp. 340–347). Ieee.
McDonald, C. M., Keeling, S. E., Brewer, M. J., & Hathaway, S. C. (2018). Using chemical and DNA marker analysis to authenticate a high-value food, manuka honey. *npj Science of Food, 2*(1), 9.
Ministry for Primary, I. (2025). Ensuring mānuka honey is authentic. In *New Zealand Government*. New Zealand Food Safety.
Nascimento, M. B., Amorim, L. R., Nonato, M. A. S., Roselino, M. N., Santana, L. R. R., Ferreira, A. C. R., … Soares, S. E. (2024). Optimization of HS-SPME/GC-MS method for determining volatile organic compounds and sensory profile in cocoa honey from different cocoa varieties (Theobroma cacao L.). *Molecules, 29*(13), 3194.
Oelschlaegel, S., Gruner, M., Wang, P.-N., Boettcher, A., Koelling-Speer, I., & Speer, K. (2012). Classification and characterization of manuka honeys based on phenolic compounds and methylglyoxal. *Journal of Agricultural and Food Chemistry, 60*(29), 7229–7237.
Rodríguez-Pérez, R., & Bajorath, J. R. (2019). Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry, 63*(16), 8761–8777.
Schmidt, C., Eichelberger, K., & Rohm, H. (2021). New Zealand mānuka honey-A review on specific properties and possibilities to distinguish mānuka from kānuka honey. *Lwt, 136*, Article 110311.
Smit, B. A., Engels, W. J. M., & Smit, G. (2009). Branched chain aldehydes: Production and breakdown pathways and relevance for flavour in foods. *Applied Microbiology and Biotechnology, 81*(6), 987–999.
Soria, A. C., Sanz, J., & Martínez-Castro, I. (2009). SPME followed by GC–MS: A powerful technique for qualitative analysis of honey volatiles. *European Food Research and Technology, 228*(4), 579–590.
Szafnauer, R. (2023). *Automated aroma and flavour profiling of honey using high-capacity Sorptive extraction*.
Taiti, C., Guardigli, G., Babbini, S., Marone, E., Masi, E., Comparini, D., & Mancuso, S. (2023). Characterization of Italian honeys: Integrating volatile and physico-chemical data. *Advances in Horticultural Science, 37*(3), 329.
Wiese, N., Fischer, J., Heidler, J., Lewkowski, O., Degenhardt, J., & Erler, S. (2018). The terpenes of leaves, pollen, and nectar of thyme (Thymus vulgaris) inhibit growth of bee disease-associated microbes. *Scientific Reports, 8*(1), 14634.
Wolski, T., Tambor, K., Rybak-Chmielewska, H., & Kedzia, B. (2006). Identification of honey volatile components by solid phase microextraction (SPME) and gas chromatography/mass spectrometry (GC/MS). *Journal of Apicultural Science, 50*(2), 115–126.
Xing, H., & Yaylayan, V. (2024). Mechanochemistry of Strecker degradation: Interaction of glyoxal with amino acids. *Food Chemistry, 439*, Article 138071.
Yadav, V., Wang, Z., Wei, C., Amo, A., Ahmed, B., Yang, X., & Zhang, X. (2020). Phenylpropanoid pathway engineering: An emerging approach towards plant defense. *Pathogens, 9*(4), 312.
Yu, D., Qu, C., Nie, J., Wen, P., Zhao, Y., Dai, C., … Wu, Q. (2025). Interpretable AI-driven multidimensional chemical fingerprints for geographical authentication of Euryales semen. *npj Science of Food, 9*(1), 133.
Zamora, R., Navarro, J. L., Aguilar, I., & Hidalgo, F. J. (2015). Lipid-derived aldehyde degradation under thermal conditions. *Food Chemistry, 174*, 89–96.
Zhang, G., & Abdulla, W. (2023). Explainable AI-driven wavelength selection for hyperspectral imaging of honey products. *Food Chemistry Advances, 3*, Article 100491.