Machine learning models for terroir classification and blend similarity prediction: A proof-of-concept to enhance cocoa quality evaluation

Machine Learning Models for Terroir Classification and Blend Similarity Prediction: A Proof-of-Concept to Enhance Cocoa Quality Evaluation

Eloisa Bagnulo[1], Felizzato Giorgio[1], Caratti Andrea[1], Cristian Bortolini[2], Chiara Cordero[1], Carlo Bicchi[1], Erica Liberto*

[1] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Turin, Italy

[2] Soremartec Italia S.r.l. (Ferrero S.p.a. group), P.le P. Ferrero 1, 12051 Alba (CN), Italy

ABSTRACT

Flavour is a key quality attribute of cocoa, essential for industry standards and consumer preferences. Automated methods for assessing flavour quality support industrial laboratories in achieving high sample throughput. Targeted and untargeted HS-SPME-GC-MS chromatographic fingerprints of cocoa volatiles from fermented beans and liquors, combined with machine learning (ML), are used for terroir qualification, enabling effective origin classification with both approaches. The targeted method, which aims to identify chemical patterns associated with sensory attributes is used for flavour comparison of origin with a reference. The similarity analysis successfully identified the most suitable origin to create new blends with a similar flavour to the industry standard. The resulting ML, model based on odorants distribution, enabled the prediction of similarity of blends to the industrial reference with an accuracy of 88%, a sensitivity of 90% and a specificity of 84%.

Keywords: cocoa flavour quality, machine learning, origin chemical-sensory blueprint, flavour benchmarking

1 INTRODUCTION

Cocoa is one of the “big” crops under stress, as its production is strongly influenced/limited by climate change and the geopolitical situation in the country of origin, while demand is constantly growing. Around 75% of global production currently comes from West and Central Africa, which accounts for covering more than 70% of the global cocoa market. The most important producing countries are the Ivory Coast, Ghana, Nigeria and Cameroon; the remaining market is covered by Latin America and South East Asia (Figure 1) (Black et al., 2020; ICCO, 2022). The cocoa demand is fuelled by a preference for chocolate with perceived health benefits, exotic flavours, variations such as single-origin and/or dark chocolate, as well as low sugar content. The cocoa market is also powered by the demand for cocoa-based ingredients such as butter and powder used in confectionery, biscuits, and cosmetic formulations (Euromonitor International, 2023; Innovamarketinsights, 2023). Climate change seriously affects cocoa production, in particular in West Africa, since changes in temperatures and precipitations (e.g. drought, heat and water stress, increase in pest and disease pressure) affect cocoa growth, yield and quality of beans (Boeckx, Bauters, & Dewettinck, 2020; Lahive, Hadley, & Daymond, 2019). Extreme weather events, in particular heatwaves or heavy rainfall during harvest, can lead to premature ripening, fermentation irregularities, and mould growth, compromising bean quality. These factors may also change the chemical and sensory profile of cocoa beans (i.e. their flavour), impacting both their market value and the choices of chocolate manufacturers and consumers (Bagnulo et al., 2023; Brazil, 2023; Delgado-Ospina, Molina-Hernández, Chaves-López, Romanazzi, & Paparella, 2021; Somarriba et al., 2021).

The challenge for the cocoa industry is therefore to have enough suitable raw materials to guarantee the consistent flavour quality of cocoa products. There are a number of strategies that helping cocoa farmers to secure or improve their yields and quality (Kolotzek, Helbig, Thorenz, Reller, & Tuma, 2018; Pieter van Donk, D., Akkerman, R. and van der Vaart, 2008). These include i) educational and training programmes on technologies to address climate change and promote more sustainable farming practises, ii) breeding strategies where science could offer a viable solution with drought and disease resistant cocoa varieties, and/or iii) a rethinking of cocoa products (Lahive et al., 2019; Somarriba et al., 2021; Wongnaa & Babu, 2020). The latter aspect has prompted the industry to experiment with different fermentation methods, blending different origins, customising roasting on the specific origin/blend and the sensory characteristics of the final products, as well as the creating new products creation of new products. (Andruszkiewicz, Corno, & Kuhnert, 2021; Balcázar-Zumaeta, Castro-Alayo, Cayo-Colca, Idrogo-Vásquez, & Muñoz-Astecker, 2023; De Vuyst & Leroy, 2020; Gutiérrez-Ríos et al., 2022; Hinneh et al., 2019; John et al., 2020; Kongor et al., 2016; Lefeber, Papalexandratou, Gobert, Camu, & De Vuyst, 2012; Lemarcq et al., 2020; Saunshia, Sandhya, Lingamallu, Padela, & Murthy, 2018; Wen et al., 2023)

Cocoa flavour depends on various factors, including genetic traits, geographical origin, environmental conditions, and processing. In particular, the origin influences the chemical and sensory properties of the cocoa flavour and thus its "terroir". This concept is relatively new for cocoa and it is borrowed from the

wine industry. As with wine, the terroir of the cocoa impacts the flavour and quality of the end product. While country of origin is a broad national designation that is useful for legal or regulatory purposes and indicates that the cocoa was grown in that country, it does not specify the quality attributes. Terroir refers to peculiar sensory characteristics that are due to environmental influences and traditional production and processing methods. This results in a variety of chocolate flavour profiles from different regions, which are specifically used for fine/speciality chocolate or for chocolate from a single origin (Fanning, Eyres, Frew, & Kebede, 2023; Putri, De Steur, Juvinal, Gellynck, & Schouteten, 2024).

Using the concept of terroir for industrial bulk chocolate with a neutral sensory profile might seem counterintuitive, but it is highly valuable for ensuring consistency and predictability in the final product. By selecting regions where cocoa naturally develops mild, uniform flavours (low acidity, fruitiness, or complexity) and controlling post-harvest processes (fermentation, drying), producers achieve a neutral sensory profile essential for blending. Monitoring this "neutral terroir" enables the selection of areas that reliably deliver stable quality year after year, minimizing environmental variability, supports traceability and sustainability, and serves as a strategic advantage for maintaining stable quality or developing new blends in particular in times of cocoa scarcity. Ultimately, terroir was used a tool for control, standardisation, and reliability in large-scale chocolate production traceability of origin based on chemical fingerprints that correlate with sensory profiles can be helpful as a starting point for the development of new products (Beckett, 2009; Fanning et al., 2023; Lucini, Rocchetti, & Trevisan, 2020; Putri et al., 2024). At the same time, to improve traceability and transparency, researchers are integrating instrumental and sensory analyses to establish reliable geographical indicators, which could support cocoa quality certification and maintain market integrity in a complex supply chain (Yu et al., 2025; Nguyen et al., 2023; Sentellas & Saurina, 2023).

Terroir quality identification relies on analytical methods capable of generating detailed diagnostic patterns that correlate with sensory attributes suitable for monitoring and quantification in quality control processes. To define cocoa's sensory characteristics, liquor tasting remains a gold standard, while targeted chemical analyses tracking origin markers are essential when certification is required. This process of terroir quality assessment requires a comprehensive identity card for liquors to align with flavour reference criteria essential for product design. Objective and robust tools to trace the authenticity of origin and quality stability of cocoa products are therefore necessary to support continuity from year-to-year in face of ever-increasing global demand at an industrial level (Bagnulo et al., 2023; Cuadros-Rodríguez, Ortega-Gavilán, Martín-Torres, Arroyo-Cerezo, & Jiménez-Carvelo, 2021).

This study is a proof-of-concept that aims implementing artificial intelligence (AI)-based methods to guide the definition of chemical-sensory blueprints of the terroir. According to literature, AI is a broader concept that encompasses not only the process of learning from a dataset (technically referred to as machine learning, ML) but also the ability to simulate human thought and behaviour in response to specific situations (Ayres, Gomez, Linton, Silva, & Garcia, 2021; Houhou & Bocklitz, 2021). Unsupervised

and supervised ML are gradually transitioning from traditional ML to the broader field of AI, opening up real opportunities to create intelligent protocols capable of responding to the environment in ways that previously required human involvement (Ayres et al., 2021; Bagnulo, Strocchi, Bicchi, & Liberto, 2024; Bressanello et al., 2018, 2021; Houhou & Bocklitz, 2021; Squara et al., 2023).

In this research work the volatile fraction of selected cocoa samples of industrial interest are explored by untargeted and targeted fingerprinting based on headspace solid-phase microextraction (HS-SPME) followed by gas chromatography-mass spectrometry (GC-MS) to discriminate the origins. Targeted odorant patterns are analysed using ML to develop classification models and identify origin diagnostic signatures. A similarity algorithm is used to evaluate their proximity to a standard reference. The closest origin is then used to design blends with different percentages. A ML model was developed and validated to predict the similarity of flavour quality to the reference and cross-checked by the sensory evaluation of the industry panel. The goal is to provide the industry with decision-making tools to better direct the selection of incoming batches and to support the design of blends matching a (sensory)quality benchmark.

2 EXPERIMENTAL

2.1 Cocoa samples and reference compounds

Samples, fermented and dried cocoa beans as well as liquors of industrial interest were sourced by Soremartec Italia srl (Alba, Italy). Beans were of commercial grade and compliant with the industrial quality control. Liquors (n=215) (*Theobroma cacao* L.) cultivar Forastero from Colombia (COL), Cameroon (CAM) and West Africa (Industrial Reference - IREF), as well as Arriba and CCN51 from Ecuador (ECU), and a series of blends (n=66) and beans (n=75) are listed in Table 1. CCN51 samples were from specific controlled fermentation.

Beans were ground in liquid nitrogen, then stored at -18°C before the analysis as well as for liquors.

Bulk cocoa (Forastero) shows strong basic chocolate flavour and accounts for more than 90% of the world production used to produce cocoa mass, powder and dark/milk chocolate. It is mainly cultivated in West Africa. In contrast, fine flavour cocoas (National, Trinitario, Criollo) are highly aromatic characterized by a wealth of sensory notes such as fruity, floral, nutty, woody, spicy etc…(Aprotosoaie, Vlad Luca, & Miron, 2016; Frauendorfer & Schieberle, 2006, 2008; Rottiers et al., 2019; Tuenter et al., 2020). The Nacional variety, unique to Ecuador, yields the characteristic Arriba beans renowned for their discernible floral and spicy flavour profiles. The clone CCN51 is highly adaptable to climatic conditions, disease-resistant and high-yielding; it emerged in Ecuador as a result of various breeding programmes and helped Ecuadorian farmers recover from El Niño in 1997/1998. From there, it then spread throughout South America, i.e. to Ecuador, Colombia, Brazil, Perù (Jaimez et al., 2022; Rottiers et al., 2019).

Pure reference standards for identity confirmation, normal alkanes (*n*-alkanes *n*-C9 to *n*-C25) for Linear Retention Index (I^T_S) determination and α-Thujone as internal standard (ISTD) were from Merk

(Milan, Italy). Alpha-Thujone (ISTD) at a concentration of 1000 mg/L was prepared in degassed sunflower oil and stored in a sealed vial at -18°C, 5.0 µL of this solution were used for internal standardisation.

2.2 Automated Head Space Solid Phase Micro Extraction: sampling device and analysis conditions

Automated Headspace Solid Phase Microextraction (HS-SPME) was performed using a Combi-PAL AOC 5000 (Shimadzu, Milan, Italy) online integrated with a Shimadzu QP2010 GC-MS system equipped with LabSolution® software (Shimadzu, Milan, Italy). A Divinylbenzene/Carboxen/Polydimethylsiloxane (DVB/CAR/PDMS) fiber (50/30 µm df - 2 cm length) from Millipore (Bellefonte, PA, USA) was used for the extraction and pre-concentration of volatiles from the samples headspace (HS). 1.00 g of cocoa powder was weighed in 20 mL HS glass vials and submitted to HS-SPME sampling for 40 min at 50°, agitation was set at 350 rpm speed (Bagnulo et al., 2023; Cordero et al., 2019; Magagna et al., 2017, 2018).

GC-MS analysis- Chromatographic conditions: Injector temperature: 240°C, injection mode: splitless; carrier gas: helium, flow rate: 1 mL/min. Analytes were recovered by thermal desorption into the split/splitless (S/SL) injection port of the GC system at 240°C for 5 min. The GC column was a SolGelwax (100% polyethene glycol) 30 m x 0.25 mm d_c x 0.25 µm d_f from Trajan Analytical Science (Ringwood, Australia). Temperature program was from 40°C (2 min) to 200°C at 3.5°C/min, then to 240°C (5 min) at 10°C/min. The mass spectrometer was set as follows: electron ionization mode (EI) at 70 eV; ion source temperature 200°C; quadrupole temperature 150°C; transfer line at 260°C; scan range: 35-350 m/z and spectrum acquisition speed at 666 spectra/min. Each sample was analysed in triplicate.

2.3 Analytes identification and data analysis

Targeted compounds were identified by matching their EI-MS fragmentation patterns with those of authentic standards and/or spectra stored in commercial (NIST2014, Wiley 7n, FFSNC) and in-house databases. Linear retention indices (I^T_S) were taken as a complementary parameter to support identification, and experimental values were compared to tabulated values in NIST Chemistry WebBook (https://webbook.nist.gov/chemistry/) (see Table 2).

AI-based tools leveraging traditional ML methods were used to decrypt information from the analytical outputs and train learning models for further predictions. Chromatographic responses were normalised to the internal standard, transformed with Log_{10} and autoscaled before data processing. Principal component analysis (PCA) was used as exploratory analysis to assess patterns and outliers. ML by Partial Least Square Discriminant Analysis (PLS-DA) or Orthogonal PLS-DA (OPLS-DA) were used to define classification models and to extrapolate variables through VIP (Variable Importance in Projection) refining the models for origin classification and for flavour quality similarity prediction. For origin's identitation, both bean and liquor samples were divided into a training set (approx. 80%) and an external test set (approx. 20%) using the Kennard-Stone algorithm. For blends benchmarking, the relative odour activity values (ROAV) of the variables important for describing the closest origin with industry standard

and IREF samples were used to train a classification model, which was internally cross-validated and then tested on an external test set of 66 liquors.

PCA and PLS-DA or OPLS-DA analysis were carried out with Pirouette® (Comprehensive Chemometrics Modelling Software, version 4.5-2014) (Infometrix, Inc. Bothell, WA). Similarity indices were calculated using statistical and data analysis solution with XLSTAT version 2021.4.1 (Addinsoft (2022), New York, USA. https://www.xlstat.com/en)

2.4 Sensory evaluation

Liquors were tested by an industrial panel (Soremartec Italia srl), which consisted of 8 trained judges, all with at least 2 years of experience in evaluating cocoa liquors. To assess similarity between origins, 6 attributes were considered: sour, caramel, fruity, floral, earthy and herbaceous on a 0 to 5 point scale.
For benchmarking a paired comparison test was conducted. The objective of this sensory evaluation was to determine whether there was a perceivable difference between two samples of cocoa liquors. The order of sample presentation was randomized to avoid positional bias.

3. RESULTS AND DISCUSSION

In the following sections, analytical target markers are determined based on chemometrics and ML for the routine control of the origin of the products entering the factory. Through ML raw data from chemical analyses are transformed into predictive models, enabling smarter and data-driven decisions. In this context, ML tools were used i) to support origin certification, ii) to define the molecular blueprint of the cocoa terroir and iii) to assess the chemical-sensory proximity of the analysed origins to the industrial reference (IREF). Information is then used to formulate different blends whose flavour similarity to the IREF is further estimated through ML.

3.1 The cocoa volatilome as chemical "terroir" blueprint

The industry's interest in the "terroir" of cocoa to match for consumer demand is constantly increasing. This trend needs a thorough knowledge of the raw materials that come from different origins (Brazil, 2023; Engeseth & Ac Pangan, 2018; Hernandez & Granados, 2021). Consumer acceptance of quality based on terroir involves the appreciation of the unique regional characteristics that contribute to the overall sensory experience of the final product and are generally associated with i) genetic factors that influence organoleptic characteristics, ii) environmental conditions including topography and climate, iii) post-harvest practises such as fermentation and drying methods, and iv) local knowledge and cultural practises (John et al., 2020; Kongor et al., 2016; Kumar et al., 2021; Kumari et al., 2018; McClure, Spinka, & Grün, 2021). For industrial chocolate production, terroir serves a multiple purpose: firstly, it enables the

selection of regions where cocoa is produced with uniformly muted flavours, which is essential for the production of standardised products with neutral sensory profiles. Secondly, it provides a basis to develop new blends that balance cocoa scarcity. Finally, geographical indications can encourage collective efforts in the areas of quality assurance, trade promotion and marketing, while promoting innovations that utilise local resources and regional biodiversity. This approach also supports both product consistency and sustainability efforts throughout the supply chain (Hernandez & Granados, 2021; Lucini et al., 2020).

The volatilome is part of the metabolomic expression of cocoa and encompasses also the volatile compounds responsible for its aroma. Unroasted beans and liquors of different origins were here analysed to evaluate the informative value of volatiles in distinguishing origin and to define models for authentication. The volatile fingerprint through untargeted and targeted features distribution encrypts information on this primary characteristic (*i.e.*, origin). The unsupervised data analysis of the untargeted chromatographic fingerprints shows a stronger origin discrimination in beans compared to liquors (Figure 2a and 2b), with an explained variance of 74.3% and 59.7% respectively. This result is remarkable as liquors undergo additional processing steps that normally reduce the origin-specific discrimination. The normalised responses of the main volatiles in beans and liquors (X variates = 50 targets) were log-transformed and pre-processed using autoscaling before PCA analysis. Exploratory data analysis through PCA of the target volatiles demonstrated comparable discriminative power regarding origin for both beans and liquors across the first three principal components (PCs) with explained variances respectively of 56.21 and 60.5%, confirming the correlation between their volatile profiles (Figure 2c and 2d). Supervised ML by PLS-DA or OPLS-DA, depending form the dataset, show good authentication models of origins with both approaches (i.e., untargeted and targeted features) and for the different cocoa products (i.e., beans and liquors). Figure S1 in the electronic supplementary material shows 3D views of the score plots from OPLS-DA (S1a, S1c and S1d) and PLS-DA (S1b) models of both untargeted (S1a – S1b) and targeted (S1c – S1d) features collected by HS-SPME-GC-MS of beans and liquors. Suitable information related to data pre-processing, algorithm used, validation, correct prediction rate, and amount (%) of samples within each test set are provided in figure S1 and in table S1 in the Supplementary material. Classification of beans using untargeted chromatographic profiles, based on a training set of 60 samples with internal five-fold cross-validation (CV 5) and an external test set of 15 samples, showed 83.3% correct classification by origin. Similarly, a PLS-DA model based on a training set of 58 liquors validated with five cross-validation (CV 5) showed an overall classification accuracy of 98.10% with an external test set of 15 samples. The models showed high specificity and sensitivity, but were slightly overfitting for beans Figure S1a and S1b and table S1.

OPLS-DA on targeted fingerprints provided a classification accuracy exceeded 92%—specifically, 92.80% for beans and 92.31% for liquors (Figures S1c and S1d). For external test set selection, about the 20% of the samples from each dataset (beans and liquors) were chosen using the Kennard-Stone algorithm. OPLS-

DA models were developed on the training sets (61 bean and 60 liquor samples), log10-transformed, autoscaled, and validated using ten-fold cross-validation (CV 10). These models were then applied to the corresponding external test sets (14 beans and 13 liquors). The results showed a similar performance to the untargeted fingerprinting approach, with improved classification rates—especially for beans (Table S1). The chocolate industry is particularly interested in identifying the country of origin for both beans and liquors, as either can be processed by the factory depending on their origin. The role of the volatilome signatures to code the country of origin in various cocoa products was already investigated by other authors (Acierno, Alewijn, Zomer, & van Ruth, 2018; Bagnulo et al., 2023; Marseglia, Musci, Rinaldi, Palla, & Caligiani, 2020). While untargeted strategies are useful for rapid screening, detailed diagnostic patterns, essential for correlating chemical profiles with sensory traits, require analytical platforms capable of identifying and quantifying the key volatiles that define the sensory characteristics of cocoa origin or chemical-sensory terroir (Bressanello et al., 2018; Fanning et al., 2023; Hernandez & Granados, 2021). Figure 3 reports the VIP scores from the O(PLS-DA) for beans (Fig. 3a) and liquors (Fig. 3b) highlighting the most relevant chemicals related to the origins. The volatile blueprint of the terroir of cocoa samples is highlighted by applying a VIP>1 cut off. Terpenes as α-Pinene (*harsh, terpene-like, minty*), δ-3-Carene (*sweet, turpentine-like*), *trans*-β-Ocimene (*Citrus, terpen like*), β-Myrcene (*balsamic, must, spicy, sweet*), limonene (*citrus-like, mint*) result to be important chemical marker for beans. These components originate from the monoterpene biosynthetic pathway and are present in higher concentrations in beans than in pulp during fermentation (Bastos et al., 2019; Chetschik et al., 2018; Colonges et al., 2021). Their higher occurrence in liquors could be related to processing steps such as roasting and grinding where they are freed from their glycosidically bound forms. Their presence depends on the origin and thereby on the post-harvest processing (Aprotosoaie et al., 2016). (Bastos et al., 2019) found that β-Myrcene and Limonene were consistently detected in the beans and pulp throughout the fermentation process. Among these compounds, β-Myrcene was predominantly found in the beans at the end of fermentation, while β-Ocimene reached its highest concentration after 5 days of fermentation. CAM, WA and COL origins were richer in terpenes than ECU. Several key aroma compounds with a VIP score greater than 1, e.g. Acetophenone (f*loral, fruity*), Benzaldehyde (*almond, burnt sugar*), and Tetramethyl pyrazine (*earthy*), are present in both cocoa beans and liquors; they play a crucial role in origin classification and show similar levels across samples Figure 3a and 3b. These aroma-active compounds together with isobutyl acetate (*sweety, fruity, ethereal*), 2- and 3-Methylbutanal (*malty*), β-Myrcene and 2-Heptanone (*balsamic, sweety, fruity*) were confirmed also by other authors as characteristic compounds of both unroasted and roasted beans, and liquors(Liu et al., 2017; Marseglia et al., 2020).

These two analytical data processing have made it possible to develop an “identity” reference tool for origin authentication for beans and liquors in the incoming factory. At the same time, the targeted analysis provides chemical patterns to support the terroir certification (Lucini et al., 2020).

3.2 Evaluation of the similarity of the different origins with a reference

Consider a chest of drawers, each containing cocoa from a distinct origin. If the raw material from the desired origin is unavailable, it is possible to select the most similar option from the other drawers or combine cocoa from different origins to achieve the required sensory properties. A major analytical challenge in quality control is to ensure that the final product is in line with reference (sensory) specifications. When supply is hampered by climate change and/or adverse geopolitical situations in some cocoa-producing areas and, at the same time, demand is increasing the availability of origin substitutes/replacers could help to partially address the above-mentioned shortage.

The investigation of chemical similarity of different origins with the industrial reference could be exploited to address this challenge. The comparison of the detailed chemical terroir fingerprints between the different origins and the industrial reference sample (IREF) is shown in Figure 4a, which visualises the ranking of the proximity index. The similarity index is based on the measure of the normalised geometric Euclidean distance between the volatile chemical space of the analysed samples. This metric, also known as NEAR index (Equation 1), describes the equivalence between two vectors within the range [0, 1] and simplifies the interpretation of the results (Valverde-Som, Ruiz-Samblás, Rodríguez-García, & Cuadros-Rodríguez, 2018). The similarity index is therefore measured as:

$$NEARij = 1 - \frac{dij}{dMAX} = 1 - \frac{\sqrt{\sum_{j=1}^{i}(xsi - xtj)2}}{dMAX} \qquad \text{Equation 1}$$

Where *dij* is the Euclidean distance between two objects and dMAX is the maximum distance that exists between two objects in the data set. Two objects that have the distance of 1 would have a perfect similarity.

The similarity between beans and liquors likely reflecting variations in chemical composition due to processing is shown in Figure 4a. Additionally, the chemical profile influenced by terroir must also be considered in relation to sensory characteristics, as the flavour is a key attribute of cocoa products. For the sensory evaluation of the flavour, the beans must be roasted, ground and processed into cocoa mass. It is essential to use liquors rather than beans for sensory evaluation of incoming samples, as their flavour profile is more directly relevant to the final product quality. However, the chemical similarity of samples does not always correspond to the sensory similarity. This is due to the complex interaction of volatile compounds and their different physico-chemical properties, which influence human perception (El Mountassir, Belloir, Briand, Thomas-Danguin, & Le Bon, 2016). This potential discrepancy highlights the importance of complementing chemical analysis with sensory evaluation. Liquors are therefore here

considered to establish possible relationships of the chemical terroir blueprint and the sensory characteristics of the cocoa samples *versus* the industrial reference.

3.3 The relationship of the chemical terroir with the sensory characteristics of the origins

In the present analytical approach, we used a simplified method for high-throughput analysis suitable for quality control. HS-SPME coupled with GC-MS provides an efficient means for characterising aroma-active compounds by concentrating volatiles from the vapour phase onto a polymer fibre coating based on their partition coefficients (Cordero et al., 2019; Liberto et al., 2020). The sensory impact of volatile compounds depends primarily on their concentration and odour threshold (OT). For rapid screening of large sample sets, the relative odour activity values (ROAV) were calculated. ROAVs assess the potential contribution of key compounds to the overall aroma profile (Cordero et al., 2019; Frauendorfer & Schieberle, 2019, Lu, Liu, Xu, & Xie, 2022; Perotti et al., 2020; Sgorbini et al., 2019; Zhu, Chen, Chen, Chen, & Deng, 2020). This approach recognises that aroma perception arises from complex interactions in which even odourless compounds can modulate sensory experience in mixtures. This provides valuable information for linking chemical data to sensory evaluations (Chambers & Koppel, 2013; El Mountassir et al., 2016; Wang, Chambers, & Kan, 2018). The ROAV of the different compounds were calculated from the chromatographic normalised responses to the internal standard divided by their odour threshold in oil (Frauendorfer & Schieberle, 2006; Lu et al., 2022; National Library of Medicine, 2020; Rychlik, Schieberle, & Grosch, 1998; The Good Scients Company Information System, 2018; Zhu et al., 2020). Sensory description of the compounds was taken from the literature (Table 2) (Cordero et al., 2019; Frauendorfer & Schieberle, 2006, 2019; Magagna et al., 2017; The Good Scients Company Information System, 2018). Six sensory notes were assessed and compared with the industrial reference: fruity, floral, caramel, vegetable, earthy, sour. The similarity in the expression of these notes was measured by the normalised Euclidean distances and reported in Figure 4b**.** The ECU origin proved to be the most similar to the industrial reference (Figure 5). This ranking was also confirmed by the industrial panel evaluation. This origin was therefore selected for further investigation to produce new blends.

3.4 The benchmark in the flavour quality with the industrial reference driver

Independent batches of ECU and IREF were monitored over two years to improve samples collection in order to model an algorithm able to classify pure classes based on ROAVs of the 19 volatiles describing the binary origin model obtained from the PLS-DA targeted liquors (Figure 3). Different percentages of ECU origin were used to create blends with the IREF and tested by the discriminant test for similarity. However, the tasting process requires a panel of experienced tasters and is time-consuming, given the limited number of samples that can be evaluated in each session. The economic commitment, involving expert trained tasters and the significant amount of time needed, poses a challenge for routine quality

controls (Bateman et al., 2023). Objective evaluation methods using analytical tools coupled to machine learning capable of measuring chemical traits can provide reliable prediction of sensory properties and confident objective decisions (Bressanello et al., 2018; Liberto et al., 2019; Scavarda et al., 2021; Strocchi et al., 2022). ML in analytical chemistry is a subcategory of AI in which developed algorithms are potentially able to learn and predict from a dataset. In this context, an AI-based tool such as an artificial smelling machine that predicts the similarity in flavour quality of new blends through their chemical composition can successfully be used to support the panel in speeding up the creation of new products. In this context, a ML algorithm based on OPLS-DA model has been drawn up on the ROAVs of the chemical blueprints of both ECU and industrial reference. The model was built with 128 samples, included 19 variables (ROAVs) and 2 classes (ECU and IREF); data were pre-processed by autoscaling and cross-validated (CV=8). The results of the cross-validated model showed high accuracy (97%) and precision (Table 3a) with optimal sensitivity and specificity. The model was then applied to an external test set of 66 unknown samples containing blends with different percentages of ECU. Table 3b shows the OPLS-DA classification results for the external test set. The results are good in terms of overall accuracy and IREF class sensitivity, considering the sampling variability used to build the model. The specificity in predicting the flavour similarity of the 66 blends with the industrial reference is only slightly lower, with only two false positives and 4 mismatched samples. With these results, the amount up to a maximum percentage of ECU cocoa in substitution of other origins in the original industrial recipe was determined to obtain blends with similar flavour quality. This approach will be extended to the other origins tested to support the development of new blends with targeted flavours as an objective tool to ensure consistency in industrial quality production.

4. CONCLUSIONS

Analytical decision-makers based on ML tools have been developed for routine control of incoming industrial products. Fingerprinting approaches provide a variety of data on all characteristics of the samples and can be used to define quality, authenticity, and for technological applications. In a difficult supply situation, the chemical fingerprints of industrial cocoa can be associated with different origins to obtain a new standard for bulk cocoa or, conversely, to qualify the stability of a product within the same origin. Targeted fingerprinting combined with machine learning helped to define the terroir that can be used for origin certification. Moreover, including origin traits and processing practices, chemical-sensory blueprints of the terroir support products authentication and traceability. The knowledge of the chemical-sensory proximity of the investigated origins to the industrial reference revealed the ECU origin as the most similar to the industrial reference. The OPLS-DA model applied on the ROAV of markers describing the chemical terroir afforded to have a ML tool to determine the limit of substitution of other cocoa origins with ECU in the original recipe. ML tool based on the chemical-sensory blueprints of the terroir

can concur to an automatic and objective decision in benchmarking and qualifying cocoa flavour. The approach developed in this study has significant implications for the cocoa industry, which is facing increasing supply problems. As climate change continues to impact traditional growing regions and consumer demand for consistent quality products increases, data-driven decision support becomes essential to accelerate product development cycles based on scientific and objective evidence. In addition, this methodology also contributes to sustainability efforts by enabling a more efficient use of available cocoa resources through optimised blending strategies.

Acknowledgements

The authors acknowledge the financial support of PhD program of the University of Turin and Soremartec Italia Srl. The attribution is based on the PNRR, Mission 4, component 2 "From Research to Business" - Investment 3.3 "Introduction of innovative doctorates that respond to the innovation needs of companies and promote the recruitment of researchers from companies" DM 352/2022.

Table captions

Table 1 Liquor and bean samples investigated

Table 2 Targeted volatiles, identity confirmation (reference standard (A), Relative retention index (RI), Mass spectrum (MS)), together with calculated and literature retention indices I^Ts, mass spectral similarity index (SI), Target ion (Ti) and qualifier ions (Qis), Odour quality and Threshold (in oil media, otherwise indicated^w).

Table 3 OPLS-DA results of the benchmarking modelling of the flavour quality of new blends: a) confusion matrix of training and external test set, b) accuracy, sensitivity, specificity and precision of training, cross-validated and external test set

Figure Captions

Figure 1 The worldwide cocoa production

Figure 2 PCA score plots of the untargeted chromatographic profiles of beans and liquors (a – b), and score plots of normalised targeted responses of beans and liquors respectively (c - d)

Figure 3 VIP scores of origin classification a) for fermented unroasted beans and b) for liquors

Figure 4 Chemical similarity a) between origins and industrial reference and b) similarity of some sensory notes between origins and industrial reference

Figure 5 Comparison of chemical-sensory pattern between ECU and industrial reference based on the ROAV of markers describing the chemical terroir of IREF and ECU (CCN51)

Supplementary Material

Figure S1 PLS-DA and OPLS-DA score plots and results of origin classification of fermented unroasted beans and liquors by a and b) untargeted fingerprints and c and d) targeted fingerprints

Table S1 OPLS-DA and PLS-DA model parameters of untargeted and targeted data for beans and liquors

References

Acierno, V., Alewijn, M., Zomer, P., & van Ruth, S. M. (2018). Making cocoa origin traceable: Fingerprints of chocolates using Flow Infusion - Electro Spray Ionization - Mass Spectrometry. *Food Control*, *85*, 245–252. https://doi.org/10.1016/j.foodcont.2017.10.002

Andruszkiewicz, P. J., Corno, M., & Kuhnert, N. (2021). HPLC-MS-based design of experiments approach on cocoa roasting. *Food Chemistry*, *360*, 129694. https://doi.org/10.1016/J.FOODCHEM.2021.129694

Aprotosoaie, A. C., Vlad Luca, S., & Miron, A. (2016). Flavor Chemistry of Cocoa and Cocoa Products-An Overview. *Comprehensive Reviews in Food Science and Food Safety*, *15*(1), 73–91. https://doi.org/10.1111/1541-4337.12180

Ayres, L. B., Gomez, F. J. V., Linton, J. R., Silva, M. F., & Garcia, C. D. (2021). Taking the leap between analytical chemistry and artificial intelligence: A tutorial review. *Analytica Chimica Acta*, *1161*, 338403. https://doi.org/10.1016/j.aca.2021.338403

Bagnulo, E., Scavarda, C., Bortolini, C., Cordero, C., Bicchi, C., & Liberto, E. (2023). Cocoa quality: Chemical relationship of cocoa beans and liquors in origin identitation. *Food Research International*, *172*, 113199. https://doi.org/10.1016/j.foodres.2023.113199

Bagnulo, E., Strocchi, G., Bicchi, C., & Liberto, E. (2024). Industrial food quality and consumer choice: Artificial intelligence-based tools in the chemistry of sensory notes in comfort foods (coffee, cocoa and tea). *Trends in Food Science and Technology*, *147*(March), 104415. https://doi.org/10.1016/j.tifs.2024.104415

Balcázar-Zumaeta, C. R., Castro-Alayo, E. M., Cayo-Colca, I. S., Idrogo-Vásquez, G., & Muñoz-Astecker, L. D. (2023). Metabolomics during the spontaneous fermentation in cocoa (Theobroma cacao L.): An exploraty review. *Food Research International*, *163*(September 2022). https://doi.org/10.1016/j.foodres.2022.112190

Bastos, V. S., Uekane, T. M., Bello, N. A., de Rezende, C. M., Flosi Paschoalin, V. M., & Del Aguila, E. M. (2019). Dynamics of volatile compounds in TSH 565 cocoa clone fermentation and their role on chocolate flavor in Southeast Brazil. *Journal of Food Science and Technology*, *56*(6), 2874–2887. https://doi.org/10.1007/s13197-019-03736-3

Bateman, R., Bellón, A., Davila, C. F., Gaca, P., Joncheere, M., Kadow, D., … Laliberte, B. (2023). Cocoa Beans : Chocolate & Cocoa Industry Quality Requirements Edition 2 : December 2023, (December).

Beckett, S. T. (2009). *Non-Conventional Machines and Processes*. *Industrial Chocolate Manufacture and Use: Fourth Edition*. https://doi.org/10.1002/9781444301588.ch17

Black, E., Pinnington, E., Wainwright, C., Lahive, F., Quaife, T., Allan, R. P., … Vidale, P. L. (2020). Cocoa plant productivity in West Africa under climate change: A modelling and experimental study. *Environmental Research Letters*, *16*(1). https://doi.org/10.1088/1748-9326/abc3f3

Boeckx, P., Bauters, M., & Dewettinck, K. (2020). Poverty and climate change challenges for sustainable intensification of cocoa systems. *Current Opinion in Environmental Sustainability*, *47*, 106–111. https://doi.org/10.1016/j.cosust.2020.10.012

Brazil, R. (2023). How Science Can Perfect Fine-Flavor Chocolate. *ACS Central Science*. https://doi.org/10.1021/ACSCENTSCI.3C01191

Bressanello, D., Liberto, E., Cordero, C., Sgorbini, B., Rubiolo, P., Pellegrino, G., … Bicchi, C. (2018). Chemometric Modeling of Coffee Sensory Notes through Their Chemical Signatures: Potential and Limits in Defining an Analytical Tool for Quality Control. *Journal of Agricultural and Food Chemistry*, *66*(27), 7096–7109. https://doi.org/10.1021/acs.jafc.8b01340

Bressanello, D., Marengo, A., Cordero, C., Strocchi, G., Rubiolo, P., Pellegrino, G., … Liberto, E. (2021). Chromatographic Fingerprinting Strategy to Delineate Chemical Patterns Correlated to Coffee Odor and Taste Attributes. *Journal of Agricultural and Food Chemistry*, *69*(15), 4550–4560. https://doi.org/10.1021/acs.jafc.1c00509

Chambers IV, E., & Koppel, K. (2013). Associations of volatile compounds with sensory aroma and flavor: The complex nature of flavor. *Molecules*, *18*(5), 4887–4905. https://doi.org/10.3390/molecules18054887

Chetschik, I., Kneubühl, M., Chatelain, K., Schlüter, A., Bernath, K., & Hühn, T. (2018). Investigations on the Aroma of Cocoa Pulp (Theobroma cacao L.) and Its Influence on the Odor of Fermented Cocoa Beans. *Journal of Agricultural and Food Chemistry*, *66*(10), 2467–2472. https://doi.org/10.1021/acs.jafc.6b05008

Colonges, K., Jimenez, J. C., Saltos, A., Seguine, E., Loor Solorzano, R. G., Fouet, O., … Lanaud, C. (2021). Two Main Biosynthesis Pathways Involved in the Synthesis of the Floral Aroma of the Nacional Cocoa Variety. *Frontiers in Plant Science*, *12*(September), 1–24. https://doi.org/10.3389/fpls.2021.681979

Cordero, C., Guglielmetti, A., Sgorbini, B., Bicchi, C., Allegrucci, E., Gobino, G., … Merle, P. (2019). Odorants quantitation in high-quality cocoa by multiple headspace solid phase micro-extraction: Adoption of FID-predicted response factors to extend method capabilities and information potential. *Analytica Chimica Acta*. https://doi.org/10.1016/j.aca.2018.11.043

Cuadros-Rodríguez, L., Ortega-Gavilán, F., Martín-Torres, S., Arroyo-Cerezo, A., & Jiménez-Carvelo, A. M. (2021). Chromatographic Fingerprinting and Food Identity/Quality: Potentials and Challenges. *Journal of Agricultural and Food Chemistry*, *69*(48), 14428–14434. https://doi.org/10.1021/acs.jafc.1c05584

De Vuyst, L., & Leroy, F. (2020). Functional role of yeasts, lactic acid bacteria and acetic acid bacteria in cocoa fermentation processes. *FEMS Microbiology Reviews*, *44*(4), 432–453. https://doi.org/10.1093/femsre/fuaa014

Delgado-Ospina, J., Molina-Hernández, J. B., Chaves-López, C., Romanazzi, G., & Paparella, A. (2021). The role of fungi in the cocoa production chain and the challenge of climate change. *Journal of Fungi*, *7*(3). https://doi.org/10.3390/jof7030202

El Mountassir, F., Belloir, C., Briand, L., Thomas-Danguin, T., & Le Bon, A.-M. (2016). Encoding odorant mixtures by human olfactory receptors. *Flavour and Fragrance Journal*, *31*(5), 400–407. https://doi.org/10.1002/ffj.3331

Engeseth, N. J., & Ac Pangan, M. F. (2018). Current context on chocolate flavor development — a review. *Current Opinion in Food Science*, *21*, 84–91. https://doi.org/10.1016/J.COFS.2018.07.002

Euromonitor International. (2023). *Chocolate Confectionery*. Retrieved from https://www.euromonitor.com/chocolate-confectionery

Fanning, E., Eyres, G., Frew, R., & Kebede, B. (2023). Linking cocoa quality attributes to its origin using geographical indications. *Food Control*, *151*(April), 109825. https://doi.org/10.1016/j.foodcont.2023.109825

Frauendorfer, F., & Schieberle, P. (2006). Identification of the key aroma compounds in cocoa powder based on molecular sensory correlations. *Journal of Agricultural and Food Chemistry*, *54*(15), 5521–5529. https://doi.org/10.1021/jf060728k

Frauendorfer, F., & Schieberle, P. (2008). Changes in key aroma compounds of Criollo cocoa beans during roasting. *Journal of Agricultural and Food Chemistry*, *56*(21), 10244–10251. https://doi.org/10.1021/jf802098f

Frauendorfer, F., & Schieberle, P. (2019). Key aroma compounds in fermented Forastero cocoa beans and changes induced by roasting. *European Food Research and Technology*, *245*(9), 1907–1915. https://doi.org/10.1007/s00217-019-03292-2

Gutiérrez-Ríos, H. G., Suárez-Quiroz, M. L., Hernández-Estrada, Z. J., Castellanos-Onorio, O. P., Alonso-Villegas, R., Rayas-Duarte, P., … González-Rios, O. (2022). Yeasts as Producers of Flavor Precursors during Cocoa Bean Fermentation and Their Relevance as Starter Cultures: A Review. *Fermentation*, *8*(7). https://doi.org/10.3390/fermentation8070331

Hernandez, C. E., & Granados, L. (2021a). Quality differentiation of cocoa beans: implications for geographical indications. *Journal of the Science of Food and Agriculture*. John Wiley and Sons Ltd. https://doi.org/10.1002/jsfa.11077

Hernandez, C. E., & Granados, L. (2021b). Quality differentiation of cocoa beans: implications for geographical indications. *Journal of the Science of Food and Agriculture*, *101*(10), 3993–4002. https://doi.org/10.1002/jsfa.11077

Hinneh, M., Abotsi, E. E., Van de Walle, D., Tzompa-Sosa, D. A., De Winne, A., Simonis, J., … Dewettinck, K. (2019). Pod storage with roasting: A tool to diversifying the flavor profiles of dark chocolates produced from 'bulk' cocoa beans? (part I: aroma profiling of chocolates). *Food Research International*, *119*(January), 84–98. https://doi.org/10.1016/j.foodres.2019.01.057

Houhou, R., & Bocklitz, T. (2021). Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. *Analytical Science Advances*, *2*(3–4), 128–141. https://doi.org/10.1002/ANSA.202000162

ICCO. (2022). Statistics - International Cocoa Organization. Retrieved 5 September 2022, from https://www.icco.org/

Innovamarketinsights. (2023). Chocolate Trends: Global Market Overview. Retrieved from https://www.innovamarketinsights.com/trends/chocolate-trends/

Jaimez, R., Barragan, L., Fernández-Niño, M., Wessjohann, L. A., Cedeño-Garcia, G., Cantos, I. S., & Arteaga, F. (2022). Theobroma cacao L. cultivar CCN 51: a comprehensive review on origin, genetics, sensory properties, production dynamics, and physiological aspect. *PeerJ*, *10*, e12676. https://doi.org/10.7717/peerj.12676

John, W. A., Böttcher, N. L., Behrends, B., Corno, M., D'souza, R. N., Kuhnert, N., & Ullrich, M. S. (2020). Experimentally modelling cocoa bean fermentation reveals key factors and their influences. *Food Chemistry*, *302*, 125335. https://doi.org/10.1016/J.FOODCHEM.2019.125335

Kolotzek, C., Helbig, C., Thorenz, A., Reller, A., & Tuma, A. (2018). A company-oriented model for the assessment of raw material supply risks, environmental impact and social implications. *Journal of Cleaner Production*, *176*, 566–580. https://doi.org/10.1016/j.jclepro.2017.12.162

Kongor, J. E., Hinneh, M., de Walle, D. Van, Afoakwa, E. O., Boeckx, P., & Dewettinck, K. (2016). Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile - A review. *Food Research International*, *82*, 44–52. https://doi.org/10.1016/j.foodres.2016.01.012

Kumar, S., D'Souza, R. N., Behrends, B., Corno, M., Ullrich, M. S., Kuhnert, N., & Hütt, M. T. (2021). Cocoa origin classifiability through LC-MS data: A statistical approach for large and long-term datasets. *Food Research International*, *140*, 109983. https://doi.org/10.1016/J.FOODRES.2020.109983

Kumari, N., Grimbs, A., D'Souza, R. N., Verma, S. K., Corno, M., Kuhnert, N., & Ullrich, M. S. (2018). Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans. *Food Research International*, *111*, 137–147. https://doi.org/10.1016/J.FOODRES.2018.05.010

Lahive, F., Hadley, P., & Daymond, A. J. (2019). The physiological responses of cacao to the environment

551 and the implications for climate change resilience. A review. *Agronomy for Sustainable*
552 *Development*, *39*(1), 5. https://doi.org/10.1007/s13593-018-0552-0

553 Lefeber, T., Papalexandratou, Z., Gobert, W., Camu, N., & De Vuyst, L. (2012). On-farm implementation
554 of a starter culture for improved cocoa bean fermentation and its influence on the flavour of
555 chocolates produced thereof. *Food Microbiology*, *30*(2), 379–392.
556 https://doi.org/10.1016/J.FM.2011.12.021

557 Lemarcq, V., Tuenter, E., Bondarenko, A., Van de Walle, D., De Vuyst, L., Pieters, L., … Dewettinck, K.
558 (2020). Roasting-induced changes in cocoa beans with respect to the mood pyramid. *Food*
559 *Chemistry*, *332*(June), 127467. https://doi.org/10.1016/j.foodchem.2020.127467

560 Liberto, E., Bicchi, C., Cagliero, C., Cordero, C., Rbiolo, P., & Sgorbini, B. (2020). Headspace sampling: An
561 'evergreen' method in constant evolution to characterize food flavors through their volatile
562 fraction. In P. Tranquida (Ed.), *Advanced Gas Chromatography in Food Analysis* (pp. 3–37). Royal
563 Society of Chemistry. https://doi.org/10.1039/9781788015752-00001

564 Liberto, E., Bressanello, D., Strocchi, G., Cordero, C., Ruosi, M. R., Pellegrino, G., … Sgorbini, B. (2019).
565 HS-SPME-MS-enose coupled with chemometrics as an analytical decision maker to predict in-cup
566 coffee sensory quality in routine controls: Possibilities and limits. *Molecules*, *24*(24).
567 https://doi.org/10.3390/molecules24244515

568 Liu, M., Liu, J., He, C., Song, H., Liu, Y., Zhang, Y., … Su, X. (2017). Characterization and comparison of key
569 aroma-active compounds of cocoa liquors from five different areas. *International Journal of Food*
570 *Properties*, *20*(10), 2396–2408. https://doi.org/10.1080/10942912.2016.1238929

571 Lu, K., Liu, L., Xu, Z., & Xie, W. (2022). The analysis of volatile compounds through flavoromics and
572 machine learning to identify the origin of traditional Chinese fermented shrimp paste from
573 different regions. *LWT*, *171*, 114096. https://doi.org/10.1016/J.LWT.2022.114096

574 Lucini, L., Rocchetti, G., & Trevisan, M. (2020). Extending the concept of terroir from grapes to other
575 agricultural commodities: an overview. *Current Opinion in Food Science*, *31*, 88–95.
576 https://doi.org/10.1016/j.cofs.2020.03.007

577 Magagna, F., Guglielmetti, A., Liberto, E., Reichenbach, S. E., Allegrucci, E., Gobino, G., … Cordero, C.
578 (2017). Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of
579 Processing: Effectiveness of Combined Untargeted and Targeted Approaches for Classification and
580 Discrimination. *Journal of Agricultural and Food Chemistry*, *65*(30), 6329–6341.
581 https://doi.org/10.1021/acs.jafc.7b02167

582 Magagna, F., Liberto, E., Reichenbach, S. E., Tao, Q., Carretta, A., Cobelli, L., … Cordero, C. (2018).
583 Advanced fingerprinting of high-quality cocoa: Challenges in transferring methods from thermal to
584 differential-flow modulated comprehensive two dimensional gas chromatography. *Journal of*
585 *Chromatography A*, *1536*, 122–136. https://doi.org/10.1016/j.chroma.2017.07.014

586 Marseglia, A., Musci, M., Rinaldi, M., Palla, G., & Caligiani, A. (2020). Volatile fingerprint of unroasted
587 and roasted cocoa beans (Theobroma cacao L.) from different geographical origins. *Food Research*
588 *International*, *132*, 109101. https://doi.org/10.1016/j.foodres.2020.109101

589 McClure, A. P., Spinka, C. M., & Grün, I. U. (2021). Quantitative analysis and response surface modeling
590 of important bitter compounds in chocolate made from cocoa beans with eight roast profiles
591 across three origins. *Journal of Food Science*, *86*(11), 4901–4913. https://doi.org/10.1111/1750-
592 3841.15924

593 Moreno-Zambrano, M., Grimbs, S., Ullrich, M. S., & Hütt, M. T. (2018). A mathematical model of cocoa
594 bean fermentation. *Royal Society Open Science*, *5*(10). https://doi.org/10.1098/RSOS.180964

595 National Library of Medicine. (2020). NlH. Retrieved from https://pubchem.ncbi.nlm.nih.gov/

596 Nguyen, Q. T., Nguyen, T. T., Le, V. N., Nguyen, N. T., Truong, N. M., Hoang, M. T., … Bui, Q. M. (2023).
597 Towards a Standardized Approach for the Geographical Traceability of Plant Foods Using
598 Inductively Coupled Plasma Mass Spectrometry (ICP-MS) and Principal Component Analysis (PCA).
599 *Foods*, *12*(9), 1848. https://doi.org/10.3390/FOODS12091848

600 Perotti, P., Cordero, C., Bortolini, C., Rubiolo, P., Bicchi, C., & Liberto, E. (2020). Cocoa smoky off-flavor:
601 Chemical characterization and objective evaluation for quality control. *Food Chemistry*, *309*,
602 125561. https://doi.org/10.1016/j.foodchem.2019.125561

603 Pieter van Donk, D., Akkerman, R. and van der Vaart, T. (2008). Opportunities and realities of supply
604 chain integration: the case of food manufacturers. *British Food Journal*, *110*(2), 218–235.
605 https://doi.org/https://doi.org/10.1108/00070700810849925

606 Putri, D. N., De Steur, H., Juvinal, J. G., Gellynck, X., & Schouteten, J. J. (2024). Sensory attributes of fine
607 flavor cocoa beans and chocolate: A systematic literature review. *Journal of Food Science*, *89*(4),
608 1917–1943. https://doi.org/10.1111/1750-3841.17006

609 Rottiers, H., Tzompa Sosa, D. A., Van de Vyver, L., Hinneh, M., Everaert, H., De Wever, J., … Dewettinck,
610 K. (2019). Discrimination of Cocoa Liquors Based on Their Odor Fingerprint: a Fast GC Electronic
611 Nose Suitability Study. *Food Analytical Methods*, *12*(2), 475–488. https://doi.org/10.1007/S12161-
612 018-1379-7/FIGURES/5

613 Rychlik, M., Schieberle, P., & Grosch, W. (1998). *Compilation of ordor thresholds, odor qualities and*
614 *retention indices of key food odorants*. Dt. Forschungsanst. für Lebensmittelchemie.

615 Saunshia, Y., Sandhya, M. K. V. S., Lingamallu, J. M. R., Padela, J., & Murthy, P. (2018). Improved
616 Fermentation of Cocoa Beans with Enhanced Aroma Profiles. *Food Biotechnology*, *32*(4), 257–272.
617 https://doi.org/10.1080/08905436.2018.1519444

618 Scavarda, C., Cordero, C., Strocchi, G., Bortolini, C., Bicchi, C., & Liberto, E. (2021). Cocoa smoky off-
619 flavour: A MS-based analytical decision maker for routine controls. *Food Chemistry*, *336*, 127691.
620 https://doi.org/10.1016/j.foodchem.2020.127691

621 Sentellas, S., & Saurina, J. (2023). Authentication of Cocoa Products Based on Profiling and
622 Fingerprinting Approaches: Assessment of Geographical, Varietal, Agricultural and Processing
623 Features. *Foods*, *12*(16). https://doi.org/10.3390/FOODS12163120

624 Sgorbini, B., Cagliero, C., Liberto, E., Rubiolo, P., Bicchi, C., & Cordero, C. (2019). Strategies for Accurate
625 Quantitation of Volatiles from Foods and Plant-Origin Materials: A Challenging Task. *Journal of*
626 *Agricultural and Food Chemistry*, acs.jafc.8b06601. https://doi.org/10.1021/acs.jafc.8b06601

627 Somarriba, E., Peguero, F., Cerda, R., Orozco-Aguilar, L., López-Sampson, A., Leandro-Muñoz, M. E., …
628 Sinclair, F. L. (2021). Rehabilitation and renovation of cocoa (Theobroma cacao L.) agroforestry
629 systems. A review. *Agronomy for Sustainable Development*, *41*(5), 64.
630 https://doi.org/10.1007/s13593-021-00717-9

631 Squara, S., Caratti, A., Fina, A., Liberto, E., Spigolon, N., Genova, G., … Cordero, C. (2023). Artificial
632 Intelligence decision-making tools based on comprehensive two-dimensional gas chromatography
633 data: the challenge of quantitative volatilomics in food quality assessment. *Journal of*
634 *Chromatography A*, *1700*, 464041. https://doi.org/10.1016/J.CHROMA.2023.464041

635 Strocchi, G., Bagnulo, E., Ruosi, M. R., Ravaioli, G., Trapani, F., Bicchi, C., … Liberto, E. (2022). Potential
636 Aroma Chemical Fingerprint of Oxidised Coffee Note by HS-SPME-GC-MS and Machine Learning.
637 *Foods*, *11*(24), 1–14. https://doi.org/10.3390/foods11244083

638 The Good Scients Company Information System. (2018). The Good Scents Company (tgsc.). Retrieved 20
639 August 2009, from http://www.thegoodscentscompany.com/

640 Tuenter, E., Delbaere, C., De Winne, A., Bijttebier, S., Custers, D., Foubert, K., … Pieters, L. (2020). Non-

volatile and volatile composition of West African bulk and Ecuadorian fine-flavor cocoa liquor and chocolate. *Food Research International*, *130*, 108943. https://doi.org/10.1016/j.foodres.2019.108943

Valverde-Som, L., Ruiz-Samblás, C., Rodríguez-García, F. P., & Cuadros-Rodríguez, L. (2018). Multivariate approaches for stability control of the olive oil reference materials for sensory analysis – part I: framework and fundamentals. *Journal of the Science of Food and Agriculture*, *98*(11), 4237–4244. https://doi.org/10.1002/jsfa.8948

Wang, H., Chambers, E., & Kan, J. (2018). Sensory Characteristics of Combinations of Phenolic Compounds Potentially Associated with Smoked Aroma in Foods. *Molecules (Basel, Switzerland)*, *23*(8). https://doi.org/10.3390/molecules23081867

Wen, S., Jiang, R., An, R., Ouyang, J., Liu, C., Wang, Z., … Liu, Z. (2023). Effects of pile-fermentation on the aroma quality of dark tea from a single large-leaf tea variety by GC × GC-QTOFMS and electronic nose. *Food Research International*, *174*(P1), 113643. https://doi.org/10.1016/j.foodres.2023.113643

Wongnaa, C. A., & Babu, S. (2020). Building resilience to shocks of climate change in Ghana's cocoa production and its effect on productivity and incomes. *Technology in Society*, *62*. https://doi.org/10.1016/j.techsoc.2020.101288

Yu, W., Ouyang, Z., Zhang, Y., Lu, Y., Wei, C., Tu, Y., & He, B. (2025). Research progress on the artificial intelligence applications in food safety and quality management. *Trends in Food Science and Technology*, *156*(November 2024). https://doi.org/10.1016/j.tifs.2024.104855

Zhu, Y., Chen, J., Chen, X., Chen, D., & Deng, S. (2020). Use of relative odor activity value (ROAV) to link aroma profiles to volatile compounds: application to fresh and dried eel (Muraenesox cinereus). *International Journal of Food Properties*, *23*(1), 2257–2270. https://doi.org/10.1080/10942912.2020.1856133

Table (Editable version)

Table 1 Liquor and bean samples investigated

Origins	Acronym	Beans for origin identitation	Liquors for origin identitation	Liquors for blends modellisation
Colombia	COL	20	14 FORASTERO	-
West Africa	WA (IREF)	19	16 FORASTERO	68
Ecuador	ECU	15	10 ARRIBA 10 CCN51	74
Cameron	CAM	21	23 FORASTERO	-
Blends (mix)	M			66

Table (Editable version)

Table 2 Targeted *volatiles, identity confirmation (reference standard (A), Relative retention index (RI), Mass spectrum (MS)), together with calculated and literature retention indices I^Ts, mass spectral similarity index (SI), Target ion (Ti) and qualifier ions (Qis), Odour quality and Threshold (in oil media, otherwise indicated[w]).*

ID	Identified compounds	Compounds Confirmation	Calc I^T_s	Lit I^T_s	SI	Target and Qualifiers ions		Odour description	Odor Threshold (μg/Kg) in oil
1	Acetone	A; RI; MS	820	821	94	43	58-41	Ethereal	$100x10^3$
2	Methyl acetate	A; RI; MS	836	832	99	43	74-59	Green, pungent	200
3	2-Methylfuran	A; RI; MS	866	871	89	82	53-39	Ethereal	$27x10^3$
4	Ethyl Acetate	A; RI; MS	878	895	98	43	61-70	Fruity, aromatic	$10x10^3$
5	2-Butanone	A; RI; MS	886	908	95	43	72-57	Slightly fruity, balsamic	$10x10^3$
6	2-Methylbutanal[§]	A; RI; MS	898	942	97	57	41-86	Malty	10
7	3-Methylbutanal[§]	RI; MS	902	917	99	44	41-58	Malty	13
8	2-Pentanone	A; RI; MS	957	988	88	43	86-58	Fruity	288
9	Isobutyl acetate	RI; MS	992	1029	96	43	56-73	Sweet, fruity, ethereal	440
10	α-Pinene	A; RI; MS	1006	1016	91	93	91-77	Harsh, terpene-like, minty	274
11	Toluene	A; RI; MS	1039	1042	95	91	39-65	Sweet	$95x10^3$
12	Ethyl 2-methylbutanoate[§]	RI; MS	1034	1062	98	57	102-41	Fruity	0.26
13	Ethyl 3-methylbutanoate	RI; MS	1050	1064	98	88	57-41	Fruity	0.6
14	2-Methyl-1-butanol acetate	RI; MS	1055	1075	98	43	87-70	Fruity	200
15	Hexanal	A; RI; MS	1060	1095	97	56	44-41	Tallowy, leaf-like	120
16	2-Methyl-1-propanol	A; RI; MS	1081	1101	91	43	41-42	Sweet, musty	$1x10^3$
17	2-Pentanol	RI; MS	1107	1125	91	45	55-41	Fruity	$8x10^{3w}$
18	3-Methyl-1-butanol acetate	A; RI; MS	1107	1125	96	43	70-87	Banana like	30
19	1,2-Dimethylbenzene	RI; MS	1118	1140	92	91	106-77	Sweet	450^w
20	δ-3-Carene	A; RI; MS	1130	1143	92	93	136-121	Sweet, turpentine-like	$4x10^{3w}$
21	β-Myrcene	A; RI; MS	1137	1150	95	93	69-79	Balsamic	13^w
22	2-Heptanone	A; RI; MS	1157	1174	93	43	58-71	Sweet, fruity	$1.5x10^3$
23	Limonene	A; RI; MS	1166	1181	91	68	93-136	Citrus, mint	250
24	3-Methyl-1-butanol	RI; MS	1187	1185	95	55	41-70	Pungent, fusel, wine, cocoa	100
25	*Trans*-β-ocimene[*]	A; RI; MS	1218	1250	91	93	79-105	Citrus, terpen like	34^w
26	Styrene	A; RI; MS	1250	1259	97	104	78-51	Balsamic	$3.13x10^3$

ID	Identified compounds	Compounds Confirmation	Calc I^T_s	Lit I^T_s	SI	Target and Qualifiers ions		Odour description	Odor Threshold (μg/Kg) in oil
27	2-Heptanol acetate	A; RI; MS	1260	1266	88	43	87-56	Woody, alcoholic	87[w]
28	3-Hydroxy-2-Butanone	A; RI; MS	1256	1259	98	45	88-73	Butter	$0.8x10^3$
29	2-Heptanol[§]	A; RI; MS	1294	1294	97	45	55-83	Citrus	10
30	4-Heptanol	A; RI; MS	1300	1308	85	55	43-73	Alcoholic	410 [w]
31	2-Nonanone	RI; MS	1363	1385	92	58	43-142	Fruity	100
32	2,3,5-Trimethylpyrazine[§]	A; RI; MS	1380	1391	94	122	42-81	Earthy	290
33	Acetic acid[§]	A; RI;MS	1410	1408	96	60	43-110	Sharp, pungent, sour	124
34	Furfural	A; RI; MS	1432	1448	95	96	39-67	Sweet, bread-like	200[w]
35	Tetramethylpyrazine	RI; MS	1442	1466	96	136	54-94	Earthy	$38x10^3$
36	Benzaldehyde	A; RI; MS	1483	1508	96	106	77-51	Almond, burnt sugar	60
37	meso-2,3-Butanediol	A; RI; MS	1507	1537	93	45	57-75	Creamy type	$668x10^3$
38	Propanoic acid	A; RI; MS	1500	1508	91	74	57-45	Pungent, acidic and dairy-like	384
39	1-Methoxy-2-propyl acetate	MS	1532	-	97	43	88-70	Sweet, ether-like	-
40	4-Hydroxybutanoic acid	MS	1581	-	92	42	86-56	-	-
41	Butanoic acid[§]	A; RI; MS	1629	1637	93	60	73-43	Sweaty	109
42	Acetophenone	A; RI; MS	1606	1627	94	105	77-120	Floral, fruity	562
43	3-Methylbutanoic acid[§]	A; RI; MS	1682	1704	95	60	43-87	Sweaty	22
44	2-Phenylethyl acetate[§]	A; RI; MS	1773	1785	98	104	91-78	Flowery	233
45	Hexanoic Acid	RI; MS	1814	1816	91	60	73-87	Pungent, sweat	$5.4x10^3$
46	Phenylethyl Alcohol[§]	A; RI; MS	1863	1912	99	91	122-65	Honey-like	211
47	2-Acetylpyrrole	A; RI; MS	1966	1971	95	94	109-66	Musty, nutty-like	$170x10^{3w}$
48	5,6-dihydro-6-pentyl-2H-Pyran-2-one	A; RI; MS	2240	2246	95	97	68-41	Sweet, creamy type	$1.1x10^{3w}$
49	n-Octanoic Acid	A; RI; MS	2045	2050	95	60	73-101	Sweaty	300
50	n-Nonanoic acid	A; RI; MS	2168	2174	95	60	73-115	Sweaty, waxy	$2.4x10^{3w}$
51	n-Decanoic acid	A; RI; MS	2270	2279	88	60	73-129	Soap-like, fatty	$>1 x10^6$

[§] Key aroma compounds. Odour description from literature (Frauendorfer & Schieberle, 2006; National Library of Medicine, 2020; The Good Scients Company Information System, 2018). Odor threshold from Rychlik, Schieberle, & Grosch, 1998,[w] Odour threshold in water.

Table (Editable version)

Table 3 OPLS-DA results of the benchmarking modelling of the flavour quality of new blends: a) confusion matrix of training and external test set, b) accuracy, sensitivity, specificity and precision of training, cross-validated and external test set

a)

Confusion Matrix

Training set cross-val	ECU	IREF	No match
ECU	74	3	3
IREF	1	46	1

Test set	ECU	IREF	No match
ECU	17	2	4
IREF	2	40	1

b)

Classification results

Training Set		***Sens***	***Spec***	***Prec***
err rate 0.03	ECU	0.96	0.98	0.99
accuracy 0.97	IREF	0.98	0.96	0.94

Cross CV=8		***Sens***	***Spec***	***Prec***
	ECU	0.96	0.98	0.99
	IREF	0.98	0.96	0.94

External Test Set		***Sens***	***Spec***	***Prec***
err rate 0.18	ECU	0.84	0.90	0.84
accuracy 0.88	IREF	0.90	0.84	0.90

Figure 1

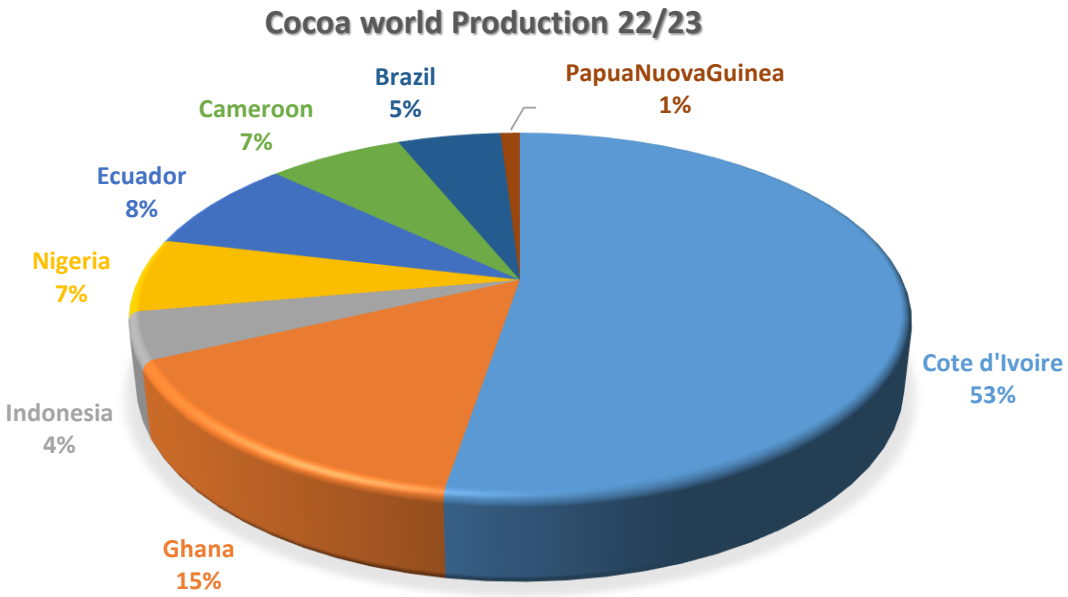
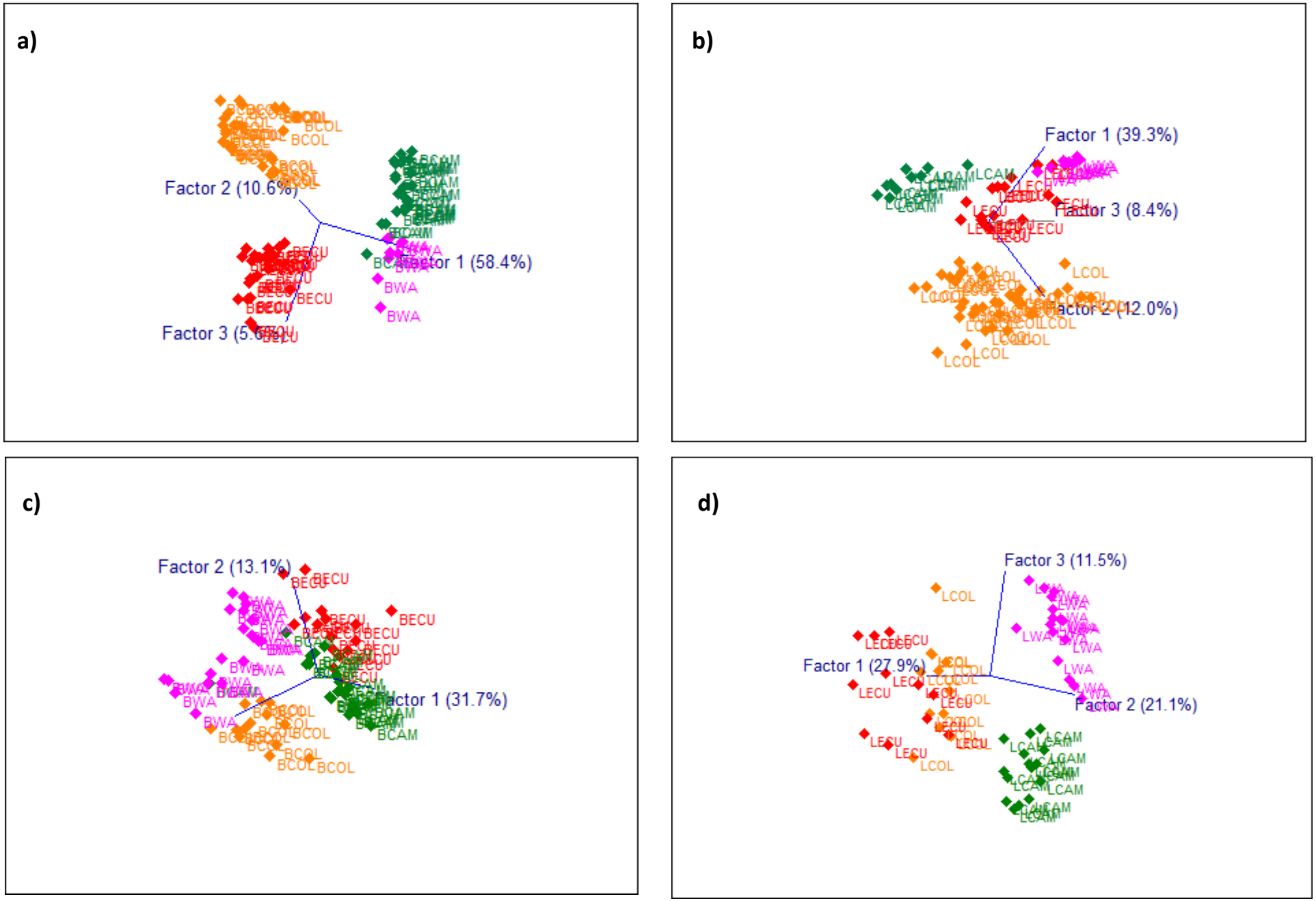


Figure 1 The worldwide cocoa production

Figure 2



Figure(s)

Figure 3

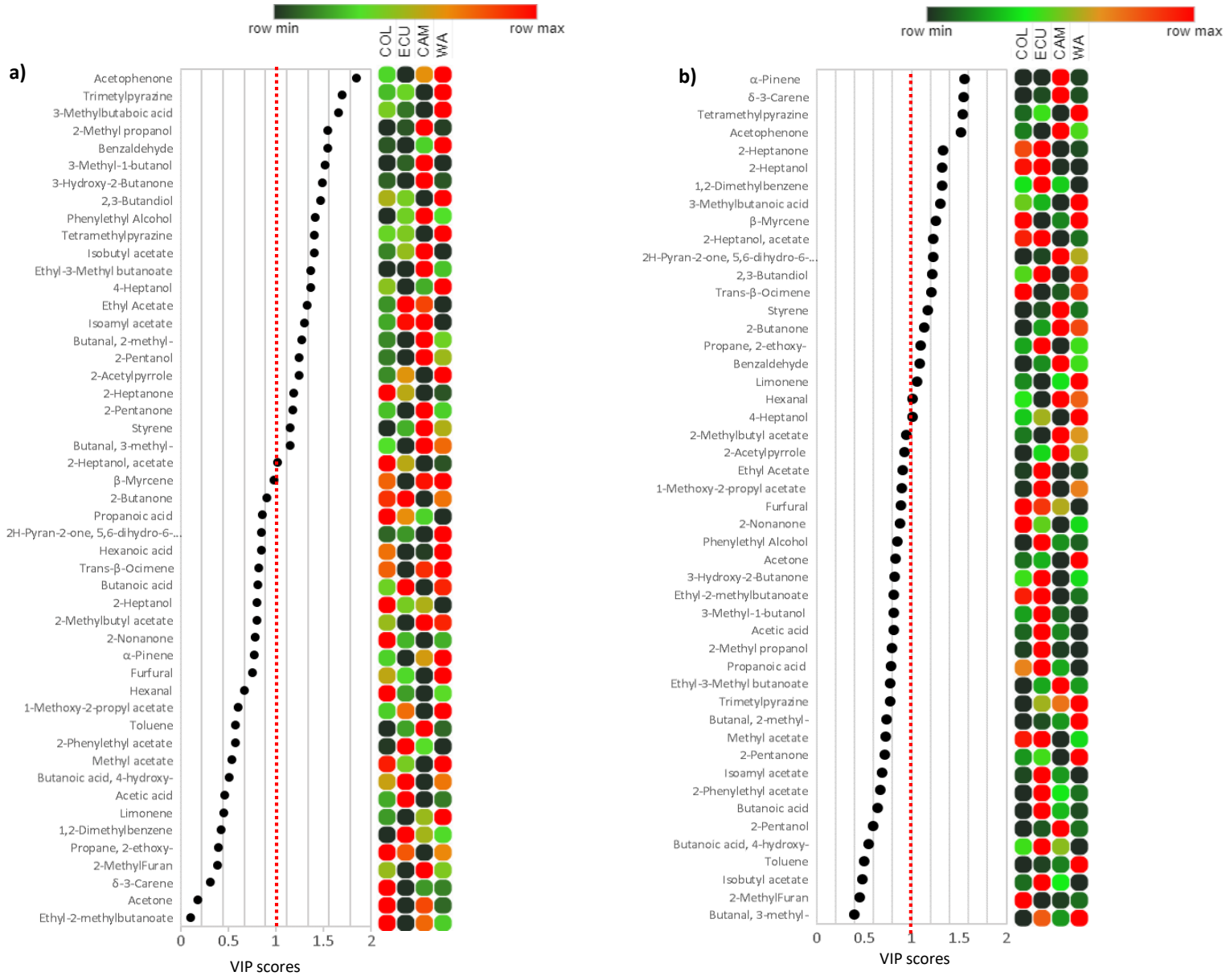


Figure 3 VIP scores of origin classification a) for fermented unroasted beans and b) for liquors

Figure(s)

Click here to access/download;Figure(s);Figure_4.pptx

Figure 4

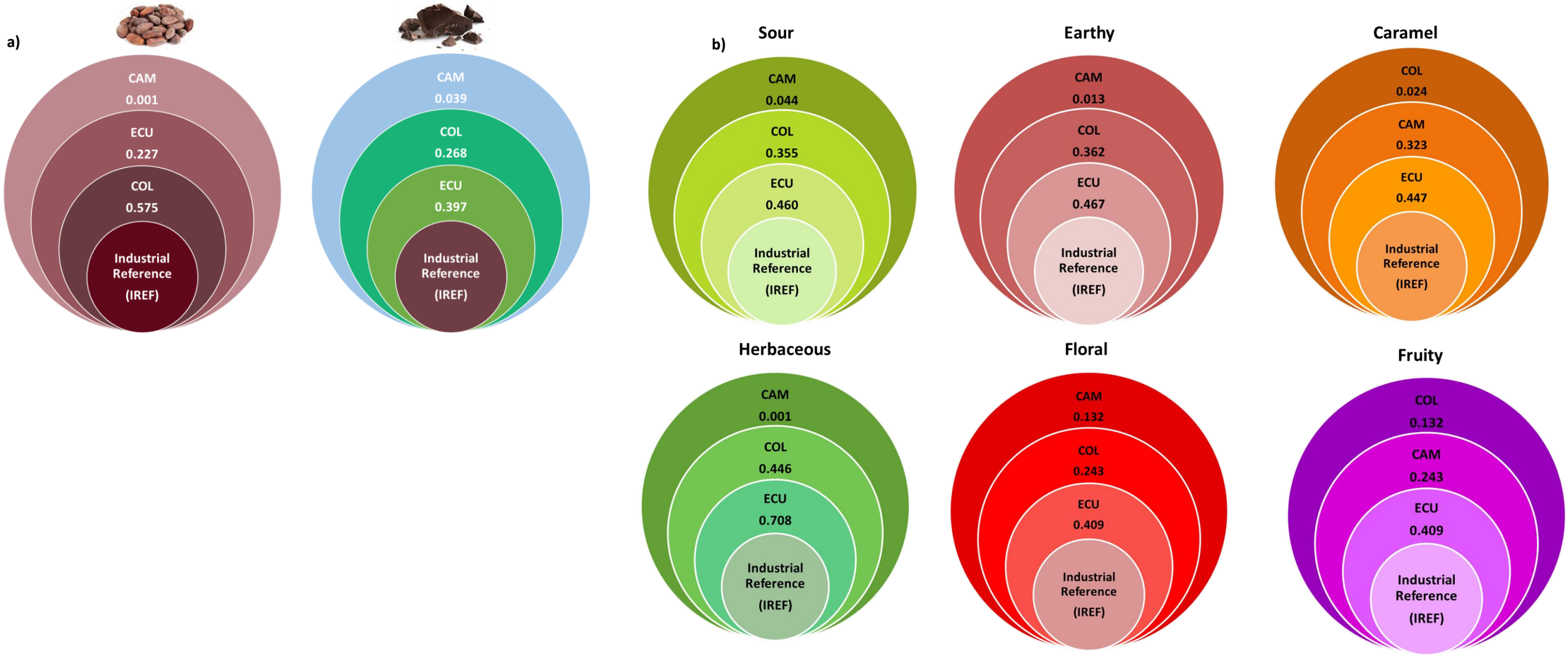


Figure 4 Chemical similarity a) between origins and industrial reference and b) similarity of some sensory notes between origins and industrial reference

Figure(s)

Click here to access/download;Figure(s);Figure_5.pptx

Figure 5

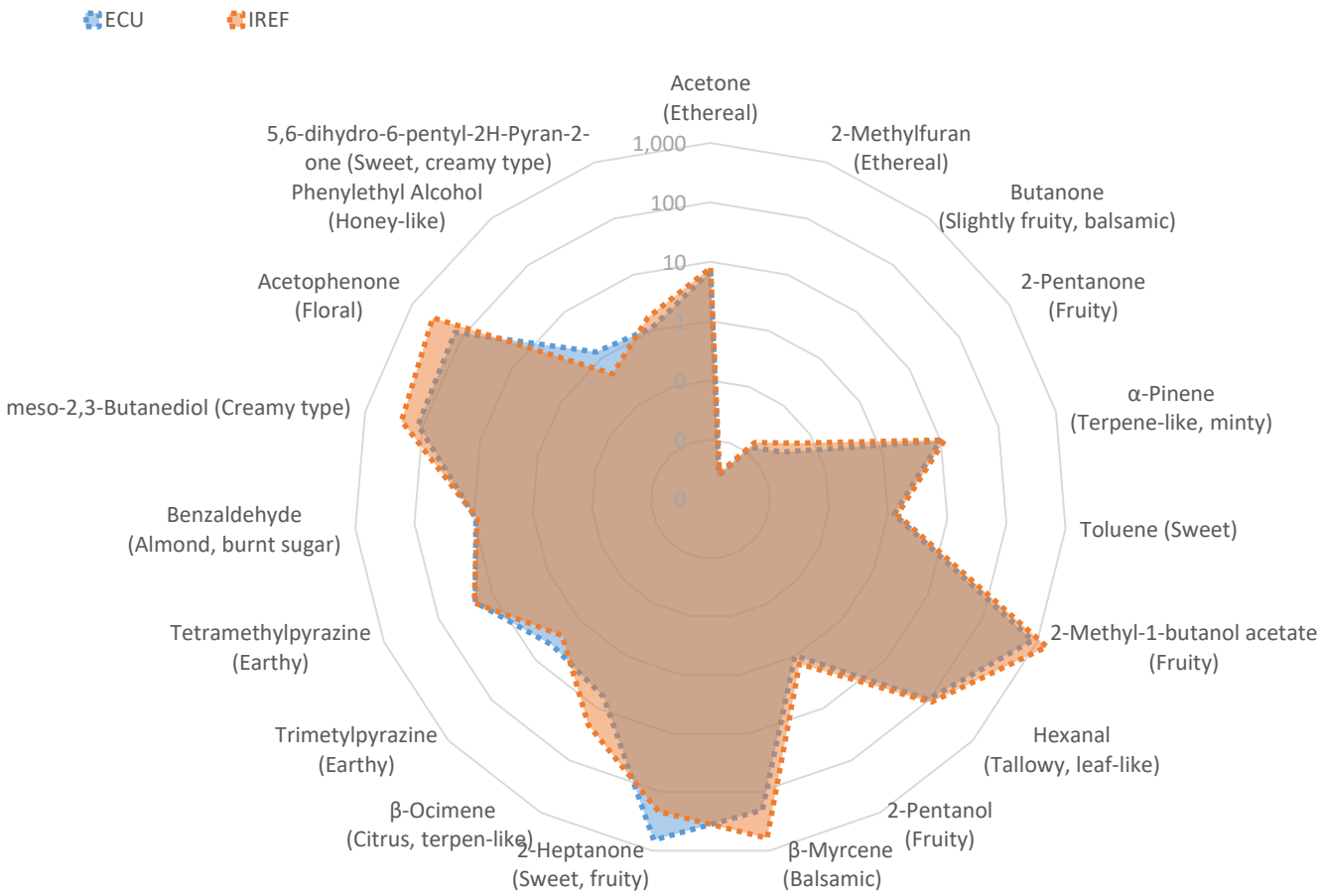


Figure 5 Comparison of chemical-sensory pattern between ECU and industrial reference based on the ROAV of markers describing the chemical terroir of IREF and ECU (CCN51)

Declaration of interests

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

☒The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Eloisa Bagnulo reports financial support was provided by SOREMARTEC srl. The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dr. Cristian Bortolini is presently employees of Soremartec Italia s.r.l Alba, Italy. Other 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.

Graphical Abstract