Identifying predictive markers in complex samples of biogenic volatile compounds using GC×GC-TOFMS and machine learning

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae (Chemistry)

In the faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

Daniel T. Pretorius

Supervisor: Dr Yvette Naudé

Co-supervisor: Prof Egmont Rohwer

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Declaration by the author

The author, Daniel T. Pretorius, hereby declares that this thesis is his own original work, and that it has not been previously submitted for degree qualification.

Signature

12/08/2022

Date

Research outputs

Oral presentation for the ChromSA Chromatography Postgraduate Student Workshop (October 2021), entitled: “The identification of foliar volatile organic compounds, from the leaves of southern African species of *Plectranthus* and *Coleus* (family: Lamiaceae), as putative biomarkers of genus, using comprehensive GC×GC-TOFMS, multivariate statistics and machine learning”.

Oral presentation for the 19th International GC×GC Symposium (June 2022), entitled: “Identifying markers of genus for *Plectranthus* and *Coleus* in complex samples of foliar volatile compounds using GC×GC-TOFMS and machine learning”.

Acknowledgements

I am indebted to my supervisors, Dr Yvette Naudé and Prof Egmont Rohwer, for their guidance, instruction and support throughout the course of my research, which took more than one unexpected turn.

To my parents, Elizabeth and Louwrens Pretorius, all my gratitude — naught of this would have come to pass without their support and counsel.

To the late Prof Anton Stoltz, my acknowledgement and fond memory are due.

My thanks to Taneshka Kruger for guiding me into the Vhembe district and helping me find my feet there; to Salphina for her assistance, and to the workers of the Masisi and Madimbo clinics for making provision for me amidst their daily routines.

And finally, my gratitude to Damian Vaz de Sousa, for helping me formulate the idea behind the research on *Plectranthus* and *Coleus*, and for his assistance and support.

Table of Contents

List of Figures

List of Tables

Abstract

Samples of biogenic VOCs are varied and complex, presenting a significant challenge to analytical scrutiny. This dual study investigates the applicability of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry (GC×GC-TOFMS), in combination with machine learning, in identifying chemical markers — in the form of biogenic volatile organic compounds (VOCs) — as a tool of classification and prediction of discrete biological states.

The first study (*Identifying predictive volatile markers of genus for southern African Plectranthus and Coleus using GC×GC-TOFMS and machine learning*) investigates foliar VOCs as markers of genus for southern African *Plectranthus* and *Coleus* species. The second study (*Identifying predictive volatile markers of malaria infection from human skin using GC×GC-TOFMS and machine learning*) investigates cutaneous VOCs from the human epidermis as markers of malaria-infection. GC×GC-TOFMS was used to analyse the relevant VOC analytes, and three machine learning algorithms (an elastic-net regression, a random forest and a support-vector machine) were used to construct models of the acquired data from a training set, and to make predictions — of genus, in the case of the first study, and on malaria-infection status, in the case of the second study — on samples from a testing set.

For the first study (N=45 samples), a predictive accuracy as high as 90% was obtained (with a sensitivity of up to 100%), and a suite of sesquiterpenes (including α- and β-cubebene, β-ylangene, β-copaene, γ-cadinene and isogermacrene D) were identified as putative markers of genus *Coleus*. Though predictive models were not obtained in the case of the second study (N=52 samples), certain compounds were identified as being potential markers of a participant's malaria-status. These include alcohols (such as (E)-2-octen-1-ol), sulphur species (such as isoamyl cyanide and isothiazole), and short- to long-chain aliphatic carboxylic acids (such as *n*-decanoic acid and 9-hexadecenoic acid).

Preface: Purpose and structure of the dissertation

This dissertation consists of the results and findings of two separate studies investigating the potential of volatile organic compounds (VOCs) of biological origin as chemical markers for prediction and classification using comprehensive two-dimensional gas chromatography-time-of-flight-mass spectrometry (GC×GC-TOFMS), multivariate statistics and machine learning. In the first study, the foliar VOC profiles of species of the genera *Plectranthus* and *Coleus* are analysed for putative chemical markers of genus; in the second study, the cutaneous VOC profiles from the epidermis of human participants are assessed for putative chemical markers of malaria infection. For the experimental and extraction phases of both studies, similar sampling strategies and extraction techniques were used, adapted for the particular type of sample (leaf or human epidermis), and both studies followed the same experimental and statistical methodology. The structure of the dissertation is summarised in Figure 1:

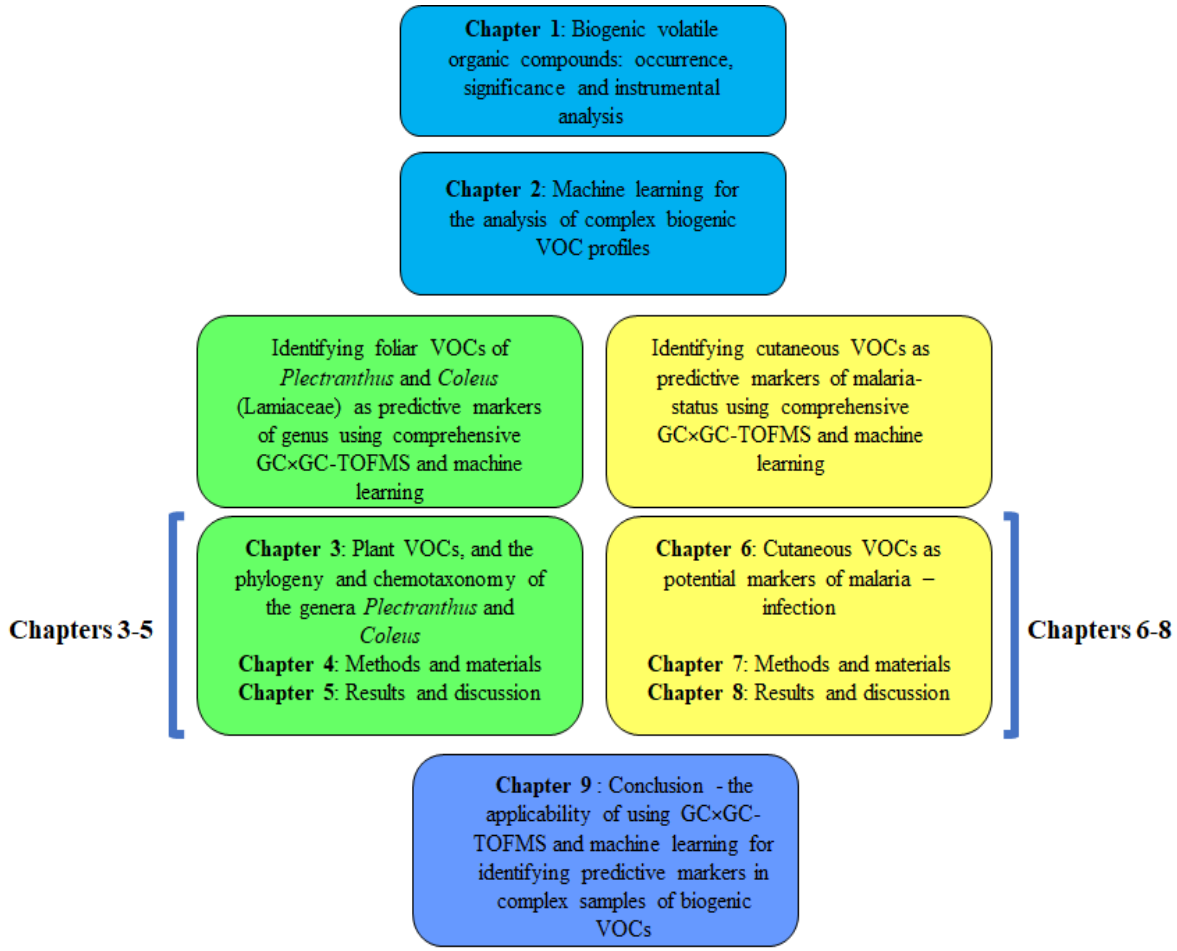


Figure 1: Structure of the dissertation. Chapters 3-5 and Chapters 6-8 are represented here to lie parallel since they respectively cover two independent studies within the broader topic of biogenic VOCs, complex sample analysis and machine learning.

Chapter 1 introduces the topic of biogenic VOCs: their occurrence and significance, as well as their extraction and instrumental analysis. Chapter 2 discusses the theory and principles of multivariate statistics and machine learning, and their application to the analytical study of complex samples in general, and biogenic VOCs in particular.

Chapters 3 to 6 cover the study of foliar VOCs from members of *Plectranthus* and *Coleus* for the prediction of genus. Chapter 3 provides background to the phylogeny and taxonomy of the two genera, outlines the ecological role of foliar VOCs, and discusses their relevance to chemotaxonomy and metabolomics. Chapter 4 consists of the experimental methods and materials for the sampling and analysis of foliar VOCs. Chapter 5 presents and discusses the results— firstly, the chromatographic data; secondly, the results of the multivariate statistical and machine learning analysis— and provides a conclusion.

Chapter 6 provides background to the topic of malaria caused by infection with the *Plasmodium* parasite, and discusses the role of cutaneous VOCs from the human epidermis as potential biomarkers of malaria-infection status. Chapter 7 consists of the experimental methods and materials for the sampling and analysis of cutaneous VOCs, and outlines the methodology and methods of the multivariate statistical and machine learning analysis. Chapter 8 presents and discusses the results— firstly, the chromatographic data; secondly, the results of the multivariate statistical and machine learning analysis— and provides a conclusion.

Chapter 9 concludes the dissertation by discussing the broader implications and applicability of using VOCs as markers for prediction and classification.

Chapter 1: Biogenic volatile organic compounds: occurrence, significance and instrumental analysis

Chapter summary

Volatile organic compounds (VOCs) are low molecular-weight organic chemical species of high vapour pressure at ambient conditions. The sources of environmental and atmospheric VOCs are anthropogenic (derived from human and industrial activity) and biogenic (derived from biological metabolism). Biogenic VOCs form a part of the metabolome of an organism, and thus have the potential to be used as biomarkers for metabolomic analysis; more particularly, for prediction and classification. Biogenic VOCs can be extracted using solid-phase microextraction/sorptive extraction. Such extracts are complex and composed of a large number of chemical constituents, analysis of which requires an instrumental technique of high sensitivity and specificity, requirements satisfied in practice by GC-MS. With the additive modification of a second stage of chromatographic separation, GC×GC-TOFMS enables a more detailed analysis of VOC samples, at lower levels of detection, than single-stage GC-MS.

1.1) Volatile organic compounds (VOCs)

1.1.1) Background and definition of VOCs

A large number of the organic molecules present in the atmosphere, whether they are of natural or anthropogenic origin, belong to the category of volatile organic compounds (VOCs) — a broad designation encompassing any organic molecule, saturated or unsaturated, that has a sufficiently low boiling point as to predominate in the vapour phase at the prevailing ambient temperature and pressure (not withstanding effects of intermolecular interactions). However, the measure of volatility is not uniformly defined [1, 2]. A study by the United States South Coast Air Quality Management District [2] defines VOCs to be those organic compounds with vapour pressures of 0.01-33 mmHg (1.3×10^{-5} to 4.3×10^{-2} atm), and showing greater than 95% evaporation by mass after six months, whereas semivolatile organic compounds (sVOCs) are defined to be those compounds with vapour pressures of 0.001-0.01 mmHg (1.3×10^{-6} to 1.3×10^{-5} atm), and showing 5-95% evaporation after six months. The subcategory of very volatile organic compounds (vVOCs) has also been proposed; the United States Environmental

Protection Agency (US EPA) places vVOCs in the boiling point range of less than 0 to 50-100 °C; VOCs in the range of 50-260°C; and sVOCs between 240 and 400°C [3].

Though there is no fixed numerical definition of volatility, most molecules that are considered to be volatile, semi-volatile or very volatile have a boiling point in the range of 0°C (or less) to 280°C, and a vapour pressure greater than 10^{-5} atmosphere [1]. The volatility of organic compound depends on the number of carbon atoms composing the molecule and, if the compound contains oxygen, its oxidation state [1]. The fewer constituent carbons, the lower the molecular mass, and the greater the volatility of the molecule. Oxidised organic molecules in solution form intermolecular hydrogen bonds which stabilise them in the liquid phase. The extent of hydrogen bond formation increases with an increase in the oxidation state, so that, for example, carbonyls tend to have lower vapour pressures than those of alcohols [1]. Multi-oxidised compounds, such as diols and dicarbonyls, have significantly lower vapour pressures than non-oxidised compounds of a similar number of, or even more, carbon atoms. For example: alkanes of fewer than twenty carbons partition from an aerosol phase to the gas phase, whereas dicarboxylic acids with only three or more carbon atoms cannot obtain sufficient energy to vaporise [1].

1.1.2) Occurrence and sources of VOCs: anthropogenic and biogenic VOCs

On earth, VOCs occur ubiquitously in air and in water. Accurate average global atmospheric VOC-level estimations are difficult to make, due to temporal and local/geographical variations in ambient concentrations; for example, the concentration of VOCs at sources and sinks may vary from less than 100 ppb to less than 1 ppb [4]. The sources of global VOCs are the biosphere (biogenic VOCs) as well as the sphere of human activity (anthropogenic VOCs) [1, 3-10]. The biosphere has been reported to be the major source of atmospheric VOCs, with an estimated emission rate of 760 Tg carbon per year [4, 5]. However, the anthropogenic contribution, previously estimated at 142 Tg (1) carbon per year, is significant [1, 4, 6-8].

Anthropogenic VOCs play a major role in atmospheric chemistry as photochemical reactants in tropospheric ozone formation [3-7]. VOCs are found in many industrial and commercial chemical products, including paints, solvents, glues, adhesives, varnishes, lacquers, waxes, incense, tobacco, e-cigarette vaporising liquids, detergents, dry-cleaning agents, perfumes, aerosol sprays, fumigants, pesticides and insect repellents. VOCs occur naturally in petroleum fuels and natural gases, the combustion and storage of which are significant sources of volatile pollutants [1, 4, 6, 7]. The widespread presence, in water and air,

of anthropogenic VOCs is an important point of environmental and health concern. Many VOCs are irritants of the respiratory tract, nose, eyes and skin, and may cause headaches, dizziness and fatigue; exposure to high concentrations of some VOCs may cause damage to the lungs, liver, kidneys or central nervous system [10].

Biogenic VOCs are secondary metabolites produced by the biochemical pathways of organisms, including microbes (bacteria and fungi), plants and animals [11-16]. In this context, they are also known as volatile organic metabolites (VOMs) [17]. The major portion of biogenic VOCs in the atmosphere are emitted by terrestrial plants in the form of isoprene and isoprenoid derivatives, the most notable of which are a group of terpene constitutional isomers: monoterpenes, sesquiterpenes and terpenoids [1, 5-8, 12]. Biogenic VOCs form the chemical basis of chemoreception, olfaction and chemical signalling, and thus play a critical role in ecological systems, mediating interactions between organisms of different trophic levels [18, 19].

1.1.3) The application of biogenic VOCs as biomarkers of biological features and states in metabolomics

The full set of metabolites produced by an organism is known as the metabolome of that organism, and the field of metabolomics is concerned with mapping and characterising whole metabolomes, or metabolomic profiles [20]. The VOCs/VOMs produced and emitted by an organism thus form a part of the metabolome of the organism, or more specifically, the volatilome [17]. Since VOMs are part of the normal metabolism of an organism, abnormal patterns in VOM production can be indicative of a pathological or disease state, and thus, in the context of medical diagnosis, VOMs can be used as biomarkers of pathology and disease [16, 17].

In general, a particular composition and ratio of a set of VOCs, from a biological sample, collectively forming a VOC profile (or volatilome), could stand as a signature, or pattern, indicative of a particular biological feature, state, or category. The premise of both studies presented in this dissertation is that VOC profiles can be used to predict a set of features pertaining to a sample from which a profile is obtained, and to categorise or classify the sample accordingly.

This dissertation is concerned with two types of biogenic VOCs: 1) foliar VOCs from the leaves of species of the genera *Plectranthus* and *Coleus*, as potential biomarkers of genus, and 2) cutaneous VOCs from the human epidermis, as potential biomarkers of status-of-infection

with the malaria-causing parasite, *Plasmodium*. Accordingly, plant VOCs are considered in more detail in Chapter 3, whereas cutaneous VOCs are discussed in Chapter 6.

1.2) The extraction of biogenic VOCs: solid-phase microextraction (SPME)/sorptive extraction

1.2.1) Principles of solid-phase microextraction (SPME)/sorptive extraction

Volatile organic compounds (VOCs) can be extracted from a variety of sources — biological, environmental, synthetic and industrial. Traditional liquid-liquid extraction and distillation techniques are effective for the isolation of bulk volatile fractions, however, the trace analysis of complex samples requires a method that can effectively preconcentrate a wide variety of trace level volatiles to the exclusion of heavier bulk components. The analytical extraction of VOCs is achieved using solid-phase microextractions (SPME) [21-24], in which a sorbent extraction phase (liquid, solid or polymeric), fixed to a solid support, is introduced to the sample, for a specified period of time, initiating the mass transfer of molecules from the sample matrix to the extraction phase, either by absorption into the body of the sorbent, or by adsorption onto the surface of the sorbent. Depending on the specific application, the extraction phase can be inserted into a liquid or vapour sample, or brought into direct contact with a solid sample, or it may be suspended in the headspace of the sample— a space of enclosed ambient air surrounding the sample, and in which volatiles accumulate [22-24].

1.2.2) Theory of SPME

The theory of SPME is presented here as derived by Pawliszyn [24]. The technique is based on the partitioning of molecules, between the sample matrix and the solid sorbent, at equilibrium, according to the equation:

$$C_0 \cdot V_s = C_s^{\infty} V_s + C_f^{\infty} V_f \qquad (1)$$

where C_0 is the initial concentration of the analyte in the sample, V_s the volume of the sample, C_s^{∞} the concentration of the analyte at equilibrium, C_f^{∞} the concentration of analyte in the sorbent at equilibrium, and V_f, the volume of the sorbent [24]. The extent to which analyte

molecules partition from the sample into the sorbent is governed by K_{fs}, the distribution coefficient:

$$K_{fs} = \frac{C_f^{\infty}}{C_s^{\infty}} \qquad (2)$$

An expression for C_f^{∞} is obtained by combining and re-expressing Equations 1 and 2:

$$C_f^{\infty} = C_0 \cdot \frac{K_{fs}V_s}{K_{fs}V_f + V_s} \qquad (3)$$

From this, an equation describing a linearly proportional relationship between the quantity of analyte extracted by the sorbent (n) and the initial concentration of the analyte in the sample (C_0), can be derived [24]:

$$n = C_f^{\infty}V_f = C_0 \cdot \frac{K_{fs}V_sV_f}{K_{fs}V_f + V_s} \qquad (4)$$

Equation 4 assumes a sample consisting of a single homogenous phase with no headspace, however, it can be modified to apply to more than one phase by factoring in the necessary distribution coefficients. Furthermore, for a sample of a large volume ($V_s >> K_{fs} \cdot V_f$), Equation (4) simplifies to [24]:

$$n = K_{fs} \cdot V_f \cdot C_0 \qquad [5]$$

Equation 5 applies to any sample whose volume is not known, such as the headspace of liquid or a solid surface [24].

SPME samplers and sampling devices may take on any form suited to a particular application—a protractible fibre, a membrane, a coated stir-bar, or a loop [22, 24-27]. This makes SPME a flexible and broadly applicable device, and can be used for many types of biological samples.

The sampling methods for both studies in this dissertation implemented SPME/sorptive extraction, albeit in different forms, for the extraction of biogenic VOCs. For the extraction of foliar VOCs from leaves of species of *Plectranthus* and *Coleus*, a protractible polydimethylsilicone (PDMS)-fibre device was employed (Chapter 4); and for the extraction

of cutaneous VOCs from the epidermis of human participants, PDMS tube samplers shaped into loops were employed (Chapter 7). Chapter 3 provides further details about the sampling and extraction of plant VOCs, whereas Chapter 6 discusses in further detail applications of sorptive extraction to the sampling and extraction of cutaneous VOCs from the epidermis.

1.3) Instrumental analysis of biogenic VOCs:

1.3.1) Gas chromatography-mass spectrometry (GC-MS)

Samples of biogenic VOCs extracted from biological sources, whether they originate from microbes, plants or animals, can be composed of a variety of trace chemical constituents, some of which may be structural (as well as stereochemical) isomers. Analysis of such complex extracts, particularly at trace-level limits of detection, requires an instrumental method of high sensitivity and specificity, which can be achieved using gas chromatography-mass spectrometry (GC-MS) [14, 28-30], coupling chromatographic separation with mass spectrometric detection. For non-targeted analysis, separation of the sample components prior to analysis is essential, and the high vapour pressure of VOCs makes GC the most suitable technique for this [14, 28]. GC-MS addresses the complexity of biogenic VOC samples in two ways: 1) by separating the sample components prior to analysis, and; 2) by combining chromatographic retention index data with mass spectral data in order to characterise, identify, and if required, quantify sample analytes (with the use of appropriate standards) that otherwise would not be easily resolved, or readily distinguishable, on the basis of chromatographic or mass spectral data alone [28, 29, 31]. GC-MS data are complementary in the sense that poorly separated compounds, represented by overlapping chromatographic peaks, can be distinguished on the basis of mass spectral information.

1.3.2) Instrumentation for GC-MS in the analysis of biogenic VOCs

The typical GC column setup for VOC investigations is an open-tubular capillary column system with a nonpolar, usually organosiloxane, stationary phase. Mass analysers used in GC-MS applications involving biogenic VOCs include: the magnetic sector [32, 33]; quadrupole [14, 30, 33-36]; the ion trap, including the quadrupole ion trap (QIT) and the toroidal ion trap

[14, 29, 37-39]; time-of-flight (TOF) [25-27, 40]; and quadrupole time-of-flight (QTOF) [41, 42]. Other GC detection methods, such as flame ionisation detection (FID) and flame photometric detection (FPD), for the detection of sulphur VOCs, are also used [43].

In a quadrupole mass analyser, increasing AC/DC radio frequency (RF) potentials, kept at a constant ratio, are applied to four rods arranged in parallel. As the RF potentials change, ions of different mass-to-charge ratios (*m/z*) are steered towards the detector [30]. Robust and cost-effective, the quadrupole is the most widely used mass spectrometer for qualitative and quantitative purposes. A quadrupole can scan the range of mass-to-charge ratios with a narrow-band filter, or it can be operated in selected-ion-monitoring (SIM) mode, which improves the sensitivity of targeted analysis, as well as the accuracy of quantification [30, 44].

The quadrupole ion trap mass spectrometer (QIT) is based on the principles of a quadrupole, but applied to a Paul-type ion trap, which is able to generate, store and selectively detect ions of different *m/z* ratios [30, 45, 46]. QIT mass spectrometers are more sensitive than normal quadrupoles, as they do not need to sacrifice full mass spectral coverage, by operating in SIM mode, in order to gain sensitivity. The major feature of QIT-MS, however, is that it offers the possibility of tandem *n*-stage mass spectrometric detection (MS/MS; or MS^n), with n = 2-6 stages. In such multi-stage analyses, precursor ions of selected *m/z* values can be made to undergo collision-induced dissociation (CID), and the mass analysis of the product fragments permits the enhanced distinction between precursor ions of similar *m/z* ratios [30, 45].

Time-of-flight mass spectrometry (TOFMS) determines the mass-to-charge ratios of ions on the basis of their transit times in a constant-field flight tube [30]. Ions of the same charge are produced in pulses/transients, and accelerated, at equal kinetic energy, by an electric field of predetermined strength, towards a detector. The flight time of an ion to the detector is measured, and used to determine its *m/z* ratio. Since all *m/z* ratios produced in a single pulse are measured, a TOF analyser does not need to linearly scan the mass range, and unlike a quadrupole, there is no need for SIM mode operation in order to quantify a particular analyte [44]. TOF-MS has a wide linear dynamic range, and shows decreased spectral bias [44]. Furthermore, TOF mass analysers have a higher spectral acquisition rate (up to 500 Hz) than a quadrupole [44]. This enables the acquisition of multiple data-points along the widths of even narrow peaks, which in turn permits the deconvolution of peaks unresolvable by a quadrupole-based GC-MS [44].

Quadrupole time-of-flight (QTOF) mass spectrometry combines the capabilities and qualities of the quadrupole and TOF: high sensitivity, good resolution and rapid spectral acquisition, as well as the option of MS-MS analysis [46]. Methods which involve the direct introduction of sample analytes into a mass spectrometer (discussed in Chapters 3B.4 and

6B.5.2), are well served by QTOF analysers [41, 42]— since there is no chromatographic separation of the sample components prior to mass analysis, the fragmentation spectra of CID products are relied upon to resolve signals produced by species of similar *m/z* ratios.

1.3.3) Comprehensive two-dimensional gas chromatography-mass spectrometry (GC×GC-MS)

Although GC-MS is well suited to the analysis of complex samples, the coelution of unresolved analytes is still a complicating factor. A more comprehensive analysis, with improved resolution of coeluting components, and thus broader analyte coverage, can be achieved by introducing a second stage, or dimension, of separation. In two-dimensional gas chromatography-mass spectrometry (GC×GC-MS), separation occurs successively in two connected columns which are coupled together by a modulator device, so that components that are not separated in the first column, or the first dimension (^{1}D), are separated in the second column, or second dimension (^{2}D), without loss of ^{1}D separation [47-49]. GC×GC is thus capable of resolving peaks that ^{1}D GC cannot, and yields both ^{1}D and ^{2}D retention data for all resolved peaks, with up to 30% more analytes detected compared to one-dimensional GC-MS [50, 51].

The two columns used in GC×GC differ in polarity, and may either be used in a nonpolar-polar combination, or a polar-nonpolar combination, the former case being the more prevalent [47, 50-53]. The two columns may be housed within the same oven, or the second column may occupy its own secondary oven. The first dimension (^{1}D) column is longer than the second dimension (^{2}D) column, and though originally it had a wider internal diameter and film thickness [53], modern secondary columns are of the same internal diameter and film thickness as the primary column. The ^{1}D separation is typically temperature-programmed, causing separation by differences in component volatility. If the second column is in the primary oven, it is operated at the temperature of the temperature programme, whereas if it is housed separately in a secondary oven, it is operated at a temperature higher by a constant amount than that of the first. However, despite temperature programming, ^{2}D separation is in effect isothermal due to the rapid ^{2}D run times [50, 53, 54]. Consequently, components that are left unresolved in ^{1}D, due to similarities in volatility, can be separated in ^{2}D on the basis of polarity. Such a twofold chromatographic separation, that is capable of resolving, on the basis of two different physical and/or chemical properties, components that cannot be resolved on the basis of one property, is termed *orthogonal*. The term *comprehensive* refers to the fact that the full

volume of each eluting band undergoes ^{2}D separation[1] [47, 53]. The use of two coupled columns means that, in principle, the total peak capacity is the product of the peak capacity of each column [54], furnishing GC×GC with overall peak capacities appreciably larger than those possible with 1D GC [54].

The two columns of a GC×GC are coupled by a device called a modulator, which cryogenically focusses, with liquid carbon dioxide or nitrogen[2], eluent from the ^{1}D column, and subsequently releases it, *via* heating with hot air, into the inlet of the ^{2}D column. Modulator designs are of two broad types: valve-based and thermal. In common use are thermal cryogenic modulators, which involve single- or dual-stage cryo-focussing, with single-, double- or quad-jet systems [52].

The modulator has three functions: 1) to focus eluent exiting the ^{1}D column into narrow bands; 2) to split these bands into narrower bands for introduction into the second column; 3) to control the timing of the freeze/heat cycle. The narrow bands produced by cryo-focussing lower the limits of detection relative to single-stage GC by up to five-fold [51, 53]. In order to maintain the separation already achieved in ^{1}D, the separation time for each band introduced into the ^{2}D column should not exceed the modulation period, otherwise late eluting peaks will appear as "wrap-around" peaks in the following modulation period. This necessitates rapid ^{2}D separation (which is why the ^{2}D column is shorter than the^{1}D column). As a result, retention times in ^{2}D are shorter than those in ^{1}D by a factor of about 10^{3} [51]. Rapid ^{2}D run times mean that the ^{2}D separation is effectively isothermal; despite programmed temperature change in the primary oven, each ^{2}D separation occurs at the current temperature of the programme [51, 53, 54].

The narrow and rapidly eluting ^{2}D bands of GC×GC necessitate detectors with small internal volumes and high detection rates (20-100 Hz) [55]. Suitable detectors include TOF mass spectrometers (500 Hz), FIDs (50-100 Hz), micro electron-capture detectors (μECDs; 50-250 Hz), atomic emission detectors (AEDs; 50-100 Hz) and element-specific detectors such as sulphur chemiluminescence detectors (SCD; 50-100 Hz) [55]. Studies that use GC×GC to detect and resolve biogenic VOCs use TOFMS [25-27]. The coupling of TOFMS to GC×GC (GC×GC-TOFMS) enhances the resolution of the technique, since overlapping peaks from the ^{2}D column can be distinguished from each other on the basis of different mass spectral fragmentation patterns [55].

[1] This is in contrast to multidimensional gas chromatography (MDGC), where only selected fractions, or heart-cuts, of eluent from the ^{1}D column are introduced into the ^{2}D column [51].
[2] For the focussing of highly volatile species, nitrogen, which is a liquid at -196˚C, is preferred to carbon dioxide, which is a liquid at -70 ˚C [51].

Chromatographic data produced by orthogonal separation is represented by a 2D chromatogram, or a 2D contour plot, with overlapping peaks on the ^{1}D axis splitting into two or more peaks on the ^{2}D axis. A 3D surface plot, with signal intensity plotted on the third axis, can also be constructed from 2D plots. On the plots of higher-dimensional chromatograms, congeners of structural, chemical or functional similarity can be seen to be distributed together along lines, or bands, oblique to the 2D axes [51-53], providing feature data unique to 2D chromatography.

References

[1] Goldstein, A.H., Galbally, I.E. 2007. *Known and unexplored organic constituents in the earth's atmosphere*. Environ. Sci. Technol., 41(5): 1514-1521. https://doi.org/10.1021/es072476p.

[2] Võ, U-U.T., Morris, M.P. 2014. *Non-volatile, semi-volatile, or volatile: redefining volatile for volatile organic compounds*. J. Air Waste Manag. Assoc., 64(6): 661-669. https://doi.org/10.1080/10962247.2013.873746.

[3] United States Environmental Protection Agency. 2017. *Technical overview of volatile organic compounds* [Internet]. Accessed: 22/04/2021. https://www.epa.gov/indoor-air-quality-iaq/technical-overview-volatile-organic-compounds.

[4] Helmig, D. Bottenheim, J., Galbally, I.E., Lewis, A., Milton, M.J.T., Penkett, S., Plass-Duelmer, C., Reimann, S., Tans, P., Thiel, S. 2009. *Volatile organic compounds in the global atmosphere*. EOS Trans. Am. Geophys. Union., 90(52): 513-520. https://doi.org/10.1029/2009EO520001.

[5] Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrakou, T., Müller, J-F., Kuhn, U., Stefani, P., Knorr, W. 2014. *Global data set of biogenic VOC emissions calculated by the MEGAN model over the last 30 years*. Atom. Chem. Phys., 14(17): 9317-9341. https://doi.org/10.5194%2Facp-14-9317-2014.

[6] Hester, R.E., Harrison, R. M., Derwent, R.G. 1995. *Sources, distributions and fates of VOCs in the atmosphere*. In: Harrison, R.M., Hester, R.E. (eds.). Volatile Organic Compounds in the atmosphere. RSC: 1-17. https://doi.org/10.1039/9781847552310-00051.

[7] Hester, R.E., Harrison, R. M., Derwent, R.G., Atkinson, R. 1995. *Gas phase tropospheric chemistry of organic compounds*. EOS Trans. Am. Geophys. Union., 90(52): 65-90. https://doi.org/10.1039/9781847552310-00065.

[8] Williams, J., Koppmann, R. 2007. *Volatile organic compounds in the atmosphere: an overview*. In: Koppmann, R. (ed.). Volatile organic compounds in the atmosphere. Blackwell: 1-32. https://doi.org/10.1002/9780470988657.ch1.

[9] Reimann, S., Lewis, A.C. 2007. *Anthropogenic VOCs*. In: Koppmann, R. (ed.). Volatile organic compounds in the atmosphere. Blackwell: 33-81. https://doi.org/10.1002/9780470988657.ch2.

[10] United States Environmental Protection Agency. *Volatile organic compounds' impact on indoor air quality* [Internet]. Accessed: 27/04/2021. https://www.epa.gov/indoor-air-quality-iaq/volatile-organic-compounds-impact-indoor-air-quality.

[11] Steiner, A.H., Goldstein, A.L. 2007. *Biogenic VOCs*. In: Koppmann, R. (ed.). Volatile organic compounds in the atmosphere. Blackwell: 82-128. https://doi.org/10.1002/9780470988657.ch3.

[12] Kesselmeier, J., Staudt, M. 1999. *Biogenic volatile organic compounds (VOC): an overview on emission, physiology and ecology*. J. Atmos. Chem., 33(1): 23-88. https://doi.org/10.1023/A:1006127516791.

[13] Korpi, A., Järnberg, J., Pasanen, A-L. 2009. *Microbial volatile organic compounds*. Crit. Rev. Toxicol., 39(2): 139-193. https://doi.org/10.1080/10408440802291497.

[14] Zhang, Z., Li, G. 2010. *A review of advances and new developments in the analysis of biological volatile organic compounds*. Microchem. J., 95(2): 127-139. https://doi.org/10.1016/j.microc.2009.12.017.

[15] Morath, S.U., Hung, R., Bennett, J.W. 2012. *Fungal volatile organic compounds: a review with emphasis on their biotechnological potential*. Fungal Biol. Rev., 26(2-3): 73-83. https://doi.org/10.1016/j.fbr.2012.07.001.

[16] De Lacy Costello, B., Amann, A., Al-Kateb, H., Flynn, C., Filipiak, W., Khalid, T., Osborne, D., Ratcliffe, N.M. 2014. *A review of the volatiles of the healthy human body*. J. Breath Res., 8(1): 014001. https://doi.org/10.1088/1752-7155/8/1/014001.

[17] Opitz, P., Herbarth, O. 2018. *The volatilome — investigation of volatile organic metabolites (VOMs) as potential tumor markers in patients with head and neck squamous cell carcinoma (HNSCC)*. J. Otolaryngol. Head Neck Surg., 47(1): 42. https://doi.org/10.1186/s40463-018-0288-5.

[18] Maffei, M.E. 2010. *Sites of synthesis, biochemistry and functional role of plant volatiles*. S. Afr. J. Bot., 76(4): 612-631. https://doi.org/10.1016/j.sajb.2010.03.003.

[19] Kost, C., Heil, M. 2006. *Herbivore-induced plant volatiles induce an indirect defence in neighbouring plants*. J. Ecol., 94(3): 619-628. https://doi.org/10.1111/j.1365-2745.2006.01120.x.

[20] Idle, J.R., Gonzalez, F.J. 2007. *Metabolomics*. Cell Metab., 6(5): 348-351. https://doi.org/10.1016/j.cmet.2007.10.005.

[21] Arthur, C.L., Pawliszyn, J. 1990. *Solid phase microextraction with thermal desorption using fused silica optical fibers*. Anal. Chem., 62(19): 2145-2148. https://doi.org/10.1021/ac00218a019.

[22] Zhang, Z., Pawliszyn, J. 1993. *Headspace solid-phase microextraction*. Anal. Chem., 65(14): 1843-1852. https://doi.org/10.1021/ac00062a008.

[23] Pawliszyn, J. 2000. *Theory of solid-phase microextraction*. J. Chromatogr. Sci., 38(7): 270-278. https://doi.org/10.1093/chromsci/38.7.270.

[24] Pawliszyn, J. *Theory of solid-phase microextraction*. In: Handbook of solid-phase microextraction [e-book]. Pawliszyn, J (ed.). Elsevier: 2011. Accessed (09/2021): 13-59. https://www.elsevier.com/books/handbook-of-solid-phase-microextraction/pawliszyn/978-0-12-416017-0.

[25] Roodt, A. P., Naudé, Y., Stoltz, A., Rohwer, E. 2018. *Human skin volatiles: passive sampling and GC × GC-ToFMS analysis as a tool to investigate the skin microbiome and interactions with anthropophilic mosquito disease vectors*. J. Chromatogr B, 1097-1098: 83-93. https://doi.org/10.1016/j.jchromb.2018.09.002.

[26] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Non-invasive sorptive extraction for the separation of human skin surface chemicals using comprehensive gas chromatography coupled to time-of-flight mass spectrometry: a mosquito-host biting site investigation*. J. Sep. Sci., 43(22): 4202-4215. https://doi.org/10.1002/jssc.202000522.

[27] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Chemical profiling of the human skin surface for malaria vector control via a non-invasive sorptive sampler with GC×GC-TOFMS*. Anal. Bioanal. Chem., 412(23): 5759-5777. https://doi.org/10.1007/s00216-020-02799-y.

[28] Hübschmann, H-J. 2008. *Handbook of GC/MS: fundamentals and applications (2nd ed.)*. Wiley-VCH Verlag GmbH & Co. KGaA. https://doi.org/10.1002/9783527625215.

[29] Materić, D., Turner, C., Morgan, G., Mason, N., Gauci, V. 2015. *Methods in plant foliar volatile organic compounds research*. Appl. Plant Sci., 3(12): 1500044. https://doi.org/10.3732/apps.1500044.

[30] Skoog, D.A., Holler, F.J., Crouch, S.R. 2007. *Principles of instrumental analysis (6th ed.)*. Brooks/Cole, Belmont.

[31] Logan, J.G., Birkett, M.A., Clark, S.J., Powers, S., Seal, N.J., Wadhams, L.J., Mordue (Luntz), A.J., Pickett, J.A. 2008. *Identification of human-derived chemicals that interfere with attraction of Aedes aegypti mosquitoes*. J. Chem. Ecol., 34(3): 308-322. https://doi.org/10.1007/s10886-008-9436-0.

[32] Robinson, A. Busula, A.O., Voets, M.A., Beshir, K.B., Caulfield, J.C., Powers, S.J., Verhulst, N.O., Winskill, P., Muwanguzi, J., Birkett, M.A., Smallegange, R.C., Masiga, D.K., Mukabana, W.R., Sauerwein, R.W., Sutherland, C.J., Bousema, T., Pickett, J.A., Takken, W., Logan, J.G., de Boer, J.G. 2017. *Plasmodium-associated changes in human odor attract mosquitoes*. PNAS, 115(18): E4209-E4218. https://doi.org/10.1073/pnas.1721610115.

[33] Soini, H.A., Bruce, K.E., Klouckova, I., Brereton, R.G., Penn, D.J., Novotny, M.V. 2006. *In situ surface sampling of biological objects and preconcentration of their volatiles for chromatographic analysis*. Anal. Chem., 78(20): 7161-7168. https://doi.org/10.1021/ac0606204.

[34] Penn, D.J., Oberzaucher, E., Grammer, K., Fischer, G., Soini, H.A., Wiesler, D., Novotny, M.V., Dixon, S.J., Xu, Y., Brereton, R.G. 2007. *Individual and gender fingerprints in human body odour*. J. R. Soc. Interface, 4(13): 331-340. https://doi.org/10.1098/rsif.2006.0182.

[35] Cagliero, C., Mastellone, G., Marengo, A., Bicchi, C., Sgorbini, B., Rubiolo, P. 2021. *Analytical strategies for in-vivo evaluation of plant volatile emissions— a review*. Anal. Chim. Acta, 1147(1): 240-258. https://doi.org/10.1016/j.aca.2020.11.029.

[36] Tholl, D., Hossain, O., Weinhold, A., Röse. U.S.R., Wei, Q. 2021. *Trends and applications in plant volatile sampling and analysis*. Plant J., 106(2): 314-325. https://doi.org/10.1111/tpj.15176.

[37] Dormont, L., Bessière, J-M., McKey, D., Cohuet, A. 2013. *New methods for field collection of human skin volatiles and perspectives for their application in the chemical ecology of human-pathogen-vector interactions*. J. Exp. Biol., 216(15): 2783-2788. https://doi.org/10.1242/jeb.085936.

[38] Natsch, A., Derrer, S., Flachsmann, Schmid, J. 2006. *A broad diversity of volatile carboxylic acids, released by a bacterial aminoacylase from axilla secretions, as candidate molecules for the determination of human body-odor type*. Chem. Biodivers., 3(1): 1-20. https://doi.org/10.1002/cbdv.200690015.

[39] Caroprese, A., Gabbanini, S., Beltramini, C., Lucchi, E., Valgimigli, L. 2009. *HS-SPME-GC-MS analysis of body odour to test the efficacy of foot deodorant formulations*. Skin Res. Technol., 15(4): 503-510. https://doi.org/10.1111/j.1600-0846.2009.00399.x.

[40] Mochalski, P., Unterkofler, K., Hinterhuber, Amann, A. 2014. *Monitoring of skin-borne volatile markers of entrapped humans by selective reagent ionization time of flight mass spectrometry in NO^+ mode*. Anal. Chem., 86(8): 3915-3923. https://doi.org/10.1021/ac404242q.

[41] Martínez-Lozano, P. 2009. *Mass spectrometric study of cutaneous volatiles by secondary electrospray ionization*. Int. J. Mass Spectrom., 282(3): 128-132. https://doi.org/10.1016/j.ijms.2009.02.017.

[42] Martínez-Lozano, P., de la Mora, J.F. 2009. *On-line detection of human skin vapors*. J. Am. Soc. Mass Spectrom., 20(6): 1060-1063. https://doi.org/10.1016/j.jasms.2009.01.012.

[43] Gallagher, M., Wysocki, C.J., Leyden, J.J., Spielman, A.I., Sun X., Preti, G. 2008. *Analyses of volatile organic compounds from human skin*. BJD, 159(4): 780-791. https://doi.org/10.1111/j.1365-2133.2008.08748.x.

[44] Binkley, J., Libarondi, M. 2010. *Comparing the capabilities of time-of-flight and quadrupole mass spectrometers*. LC GC Spec. Issue, 8(3): 28-33. https://www.chromatographyonline.com/view/comparing-capabilities-time-flight-and-quadrupole-mass-spectrometers-0.

[45] March, R.E. 2000. *Ion trap mass spectrometers*. In: Lindon, J.C., Tranter, G.E., Koppenaal, D.W. (eds.). Encyclopedia of spectroscopy and spectrometry (3rd ed). Elsevier: 330-337. https://doi.org/10.1016/B978-0-12-409547-2.12675-7.

[46] Chernushevich, I.V., Loboda, A.V., Thomson, B.A. 2001. *An introduction to quadrupole-time-of-flight mass spectrometry*. J. Mass Spectrom., 36(8): 849-865. https://doi.org/10.1002/jms.207.

[47] Phillips, J.B., Beens, J. 1999. *Comprehensive two-dimensional gas chromatography: a hyphenated method with strong coupling between the two dimensions*. J. Chromatogr. A, 856(1-2): 331-347. https://doi.org/10.1016/S0021-9673(99)00815-8.

[48] Phillips, J.B., Gaines, R.B., Blomberg, J., van der Wielen, F.W.M., Dimandja, J-M., Green, V., Granger, J., Patterson, D., Racovalis, L. de Geus, H-J., de Boer, J., Haglund, P., Lipsky, J., Sinha, V., Ledford, E.B. *A robust thermal modulator for comprehensive two-dimensional gas chromatography*. J. High Resol. Chromatogr., 22(1): 3-10. https://doi-org/10.1002/(SICI)1521-4168(19990101)22:1%3C3::AID-JHRC3%3E3.0.CO;2-U.

[49] Truong, T.T., Marriott, P.J., Porter, N.A. 2001. *Analytical study of comprehensive and targeted multidimensional gas chromatography incorporating modulated cryogenic trapping*. J. AOAC Int., 84(2): 323-336. https://doi.org/10.1093/jaoac/84.2.323.

[50] Winnike, J.H., Wei, X., Knagge, K.J., Colman, S.D., Gregory, S.G., Zhang, X. 2015. *Comparison of GC-MS and GC×GC-MS in the analysis of human serum samples for biomarker discovery*. J. Proteome Res., 14(4): 1810-1817. https://doi.org/10.1021/pr5011923.

[51] LECO Corporation. 2016. *Pegasus® HT-C and Pegasus 4D-C* [Brochure]. LECO, St. Joseph, form no. 209-252.

[52] Mondello, L. 2012. *GC×GC handbook: fundamental principles of comprehensive 2D GC* [brochure]. Shimadzu Corp., C146-E177: 4-10. https://www.shimadzu.com/an/literature/gcms/jpo212150.html.

[53] Ramos, L. (ed.), Brinkman, U.A.Th. 2009. *Multidimensionality in gas chromatography: general concepts*. In: Comprehensive two dimensional gas chromatography. Ramos, L. (ed.). Elsevier, Oxford: 3-14. https://doi.org/10.1016/S0166-526X(09)05501-9.

[54] Semard, G., Adahchour, M., Focant J-F. 2009. *Basic instrumentation for GC×GC*. In: Comprehensive two dimensional gas chromatography. Ramos, L. (ed.). Elsevier, Oxford: 15-44. https://doi.org/10.1016/S0166-526X(09)05502-0.

[55] Sanz, J. 2009. *Theoretical considerations*. In: Comprehensive two dimensional gas chromatography. Ramos, L. (ed.). Elsevier, Oxford: 49-75. https://doi.org/10.1016/S0166-526X(09)05503-2.

Chapter 2: Machine learning for the analysis of complex biogenic VOC profiles

Chapter Summary

Machine learning is suited to the analysis of complex samples of biogenic VOCs. The multivariate data obtained from GC-MS and GC×GC-TOFMS can be used by algorithms to construct models of the data, and to make predictions on data from new samples. The two types of machine learning are supervised and unsupervised learning. Supervised learning can be performed using simple classification or regression. The relative bias and variance of a model affects predictive performance, and thus an accurate model must find an optimum bias-variance trade-off. The machine learning pipeline consists of four steps: 1) data preparation and pre-processing; 2) training/tuning; 3) testing/prediction, and; 4) variable ranking and feature selection. The effects of outliers and highly correlated variables on the bias are accounted for by using regularisation techniques. The three algorithms implemented in this study are briefly discussed: 1) elastic-net regression (ridge-lasso logistic regression); 2) random forest; 3) support-vector machine.

2.1) Background: supervised and unsupervised learning

As discussed in Chapter 1.3, biogenic VOC samples typically contain many compounds in highly variable relative abundance. Analysis of VOC profiles, by GC-MS or GC×GC-TOFMS, yields large and complex datasets consisting of multiple predictors that require multivariate statistical treatment. The analysis of such datasets is made possible by machine learning algorithms that construct mathematical models of the data, optimise the modelling process in a systematic fashion, and make predictions on future samples.

There are two broad categories of machine learning: supervised and unsupervised learning [1, 2]. Supervised learning involves the classification of each sample of a dataset in a dual or multi-class system. For example, in this study, a sample of foliar VOCs collected from a leaf may be classified according to genus (*Plectranthus* or *Coleus*), or a sample of cutaneous VOCs, collected from the epidermis of a patient suspected of having malaria, may be categorised as either *positive* or *negative* with regards to the malaria-status of the patient. In supervised learning, the class of each sample is known to the algorithm, whereas in unsupervised learning,

there are no data labels, and the algorithm searches for associated features that cause clustering in the data [1, 2].

There are two types of supervised learning: simple classification and regression [1, 2]. The aim of both types is prediction, however whereas simple classification uses discrete values to denote each class, regression uses continuous numerical input to output continuous probability values [1, 2].

2.2) The bias-variance trade-off and regularisation

A good model is one that both accurately describes the training data and that generalises well to new data. There are two properties of a model that determine its predictive capabilities: the bias and variance.

The bias describes how closely the model is fit to the dataset; a model with high bias has a poor fit, and will not produce accurate predictions [1]. A decrease in the bias equates to an increase in the variance (dispersion) of the model, which increases the accuracy of the fit [1, 2]. However, an overly high variance leads to overfitting of the model to the data, which ultimately results in poorer predictive performance on new data [1]. In order to improve predictive performance, a compromise between the bias and the variance has to be reached, a situation called the bias-variance trade-off [1].

The bias can be sensitive to outliers and correlated predictors, and in regression equations this may result in regression coefficients of large magnitude [2]. A model can be rendered more robust to extreme values with regularisation— the process of penalising, or shrinking, estimated coefficients [1, 2]. The implementation of the bias-variance trade-off and regularisation is discussed further in Chapter 2.3.2.

2.3) The machine learning pipeline

The general method for supervised learning involves deriving a model in a training step, and then assessing the predictive performance of the model on new data in a testing step. A schematic pipeline for the machine learning process is depicted in Figure 2, as adapted from the method of Kuhn, 2008 [3].

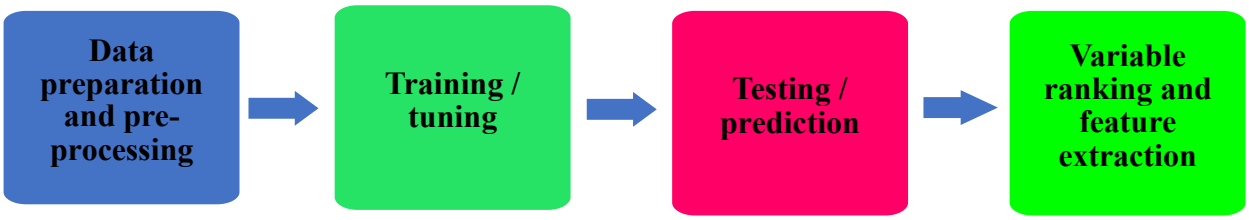


Figure 2: A pipeline for machine learning analysis for prediction (classification and regression). Adapted from the method of Kuhn, 2008 [3].

Details of the method are outlined in Chapter 4 (for foliar VOCs of *Plectranthus* and *Coleus*) and in Chapter 7 (for cutaneous VOCs). The pipeline consists of four steps: 1) data preparation and pre-processing; 2) training/tuning; 3) testing/prediction, and; 4) variable ranking and feature selection.

2.3.1) Data preparation and pre-processing

In most situations, the data may require processing prior to modelling. This may involve transformation and normalisation of the raw data (centring and scaling), the imputation of missing values (if applicable), and the removal of variables which have a variance close or equal to zero [3, 4]. Such near-zero variance predictors do not contribute any significant variation to the data, and can thus be removed.

2.3.2) Training

Training is the central step in the machine learning pipeline, in which the algorithm constructs a model of the data which can be used to make predictions on the samples from the testing set [1-4]. Since the distribution and structure of the data is not necessarily known, and since not all models will be equally suited to the problem — nor necessarily give the same results — more than one algorithm may be trained. The statistical models implemented by algorithms are controlled by parameters, such as a coefficient in the equation of the statistical model in question, or a rule for executing a particular algorithmic procedure [1-4]. For regularised models, these parameters determine the types and magnitudes of the regularisation penalties. The ideal parameter values are not known, and different values result in models with variable predictive performance during testing. Consequently, these parameters need to be tuned in a systematic fashion in order to find their optima, and thus the training step is synonymous with the tuning of machine learning models.

Tuning is a reiterative process in which a range of parameter values are tested, producing a set of models with differing outputs. The accuracy/fit of each model is measured and optimised using a statistical resampling method, such as bootstrapping or cross-validation [1-4]. The resampling procedure splits the training set into multiple training/testing splits, and for each split, a model, controlled by the set of tuning values chosen for the split, is constructed and used to make in-training predictions.

The accuracy of in-training predictions by resampling is evaluated on the basis of a metric which can be plotted in terms of a loss function: the area under the curve (AUC) of the receiver operating characteristic (ROC). The ROC curve plots the sensitivity and specificity values for predictions, and the optimal model is that with the highest AUC of the ROC, as determined by bootstrapping or cross-validation, for a given parameter, or combination of parameters [2-5]. The final model selected is used to make predictions on the testing set, and to rank the importance, or weight, of the variables describing the data.

The process of tuning is a practical solution to the bias-variance trade-off (Chapter 2.2). Inputting different coefficient values, or the implementation of a different algorithmic rule, may produce models of differing bias and variance, and which are thus comparatively under- or overfit to the training data. Tuning permits a wide range, or grid, of parameters to be optimised — without the need of assessing the accuracy of each model individually, which would be an impractical and time-consuming task — and the final model, with the optimal parameters, strikes the optimal trade-off between bias and variance.

2.3.3) Testing/prediction

The final model constructed and selected during training is used to classify samples in the testing set [1-4]. Data for each sample is used as input, with the output being discrete for simple classification problems, and continuous for regression problems. The probability of the sample belonging to one or the other of a given set of classes is then calculated, and the class with the highest probability is selected to be the predicted case [3, 4]. At this point, if only a narrow tuning range was explored, and if it is reasonable to suspect that the predictive performance of the model could be improved by more extensive tuning, the training step can be repeated with a wider tuning range or grid.

2.3.4) Variable ranking and feature selection

The final step in the pipeline consists of ranking the variables in terms of their importance, or weight, in the model [1, 3, 4]. In principle, high ranking variables correspond to select features, and are thus good predictors of a particular category. The method for ranking variables depends on the algorithm employed. For example, a regression model will rank variables according to the values of their coefficients, or a metric such as the AUC of the ROC is used to calculate the predictive reliability of a variable [3, 4].

2.4) Types of machine learning algorithms

There are different types of models for machine learning, each with its own variations and modifications, and different algorithms for implementing them. Many of the statistical models for machine learning (ML) fall under the family of Generalised Linear Models (GLMs), in which the data is assumed to follow a distribution of the family of exponential distributions [1, 5-7]. A GLM is constructed from a probability distribution function (f) of a response variable (Y) and a linear combination of predictors (η) related to the expected value (μ) by a link function (g) [5]. *Generalised* linear model is a term distinct from *general* linear model, which refers to linear regression models, including least-squares linear regression, which itself is a particular example of a generalised linear model in which a normal distribution is assumed. However, the *linear* in GLM refers to the linear combination of predictors, not the linearity of a function. GLMs may use a combination of linear functions to approximate an exponential dispersion, or it may use an exponential function [1, 5-7].

The three supervised models used in both studies included in this dissertation are discussed: 1) elastic-net regression (ridge-lasso logistic regression); 2) random forest; and 3) support-vector machine.

2.4.1) Elastic-net regression / ridge-lasso logistic regression

Elastic-net regression is a special application of regularised logistic regression. Logistic regression is a GLM that uses a logistic function to model the probability of a certain event being true, such as the probability that a sample belongs to a particular class, using maximum likelihood estimation [1, 2, 5-7]. Binary logistic regression is used for sample data which may fall in one of two categories, with the binary value (0 or 1) of the dependent variable indicating

its category; multinomial logistic regression is used for sample data of multiple possible categories, and thus assigns multiple indicator values to represent the sample categories [5].

In many applications, regularisation is applied to logistic regression in order to increase the variance of the model and prevent overfitting. For logistic models, this involves applying shrinkage penalties to the values of the regression coefficients of the predictors [1, 2, 5]. There are two forms of regularisation used in logistic regression: 1) least absolute shrinkage and selection operator (lasso) regression and ridge regression. For both ridge and lasso regression, the magnitude of the penalty is determined by the parameter lambda (λ). In lasso regression the penalty is taken as the absolute value of the coefficients (the L_1 norm or rectilinear distance), whereas in ridge regression, it is taken as the square of the magnitude of the regression coefficients (the L_2 norm, or the Euclidean norm) [1]. Consequently, coefficients in a lasso regression can reach zero and become excluded, producing a sparser model, whereas those of a ridge regression can only tend towards zero, and are thus retained [1, 8-10].

In elastic-net regression, the ridge and lasso operators are combined according to an optimal ratio determined during tuning [8-10]. By attenuating the magnitude of the regression coefficients, the model is desensitised to outliers and to multicollinearity between predictors, increasing the variance, and leading to improved predictions on the testing set.

2.4.2) Random forest

Given multiple predictors for a sample, a decision tree algorithm can be used to predict the category of a new observation by inputting values of the new observation into the decision tree [1, 2]. Though individual trees can accurately describe the data on which they are constructed (which is to say, trained), they tend to be inaccurate at future classification; however, the aggregate predictions of a large number, or forest, of decision trees tend to be more accurate [11, 12]. A random forest is an ensemble technique that uses a multitude of (semi)-random individual trees in the modelling process. An important difference between traditional decision trees and those constructed in a random forest is that, in the former, all the predictors, or *features*, occurring at the nodes, are included in the tree, whereas the trees of a random forest are constructed from a randomly selected set, or *subspace*, of the predictors [1, 2, 11-13]. Since the individual trees of a random forest contain a random subspace of predictors, they are simpler than standard trees; this represents a trade-off between the bias and the variance.

2.4.3) Support-vector machine

Support-vector machines are effective and robust algorithms for the classification of dual- and multi-class data with a high degree of accuracy, even in cases where there is overlap between datapoints of different classes [14-17].

Given a set of observations that fall into one of two categories that are linearly separable — there is no overlap in the dispersion of two groups of observations — a margin, called a support vector classifier, that separates the observations of different categories, can be computed [1, 2, 15]. Provided the sets of observations are linearly separable, an optimal classifier is one in which the distance between the classifier and the most proximal datapoints to it is maximised, without any being separated (misclassified) into the incorrect category. However, in practice, most datasets are not linearly separable without error. A support-vector machine is able to separate linearly inseparable data by projecting datapoints into an *n*-dimensional space, and computing a hyperplane, or support vector classifier, that optimises the separation between distinguishable groups of datapoints within that space [1, 2, 15-17]. The optimal hyperplane is one that results in the lowest misclassification error (as determined by a resampling technique such as bootstrapping or cross-validation), and is termed a soft margin classifier [1, 2, 16].

Because transformation of linearly inseparable data into a high dimensional space is not computationally feasible, a kernel function can be used to compute high dimensional coordinates of the data, without performing the transformation— the so-called kernel trick [1, 2]. This is done by taking the dot product of pairs of observations in order to find a soft margin classifier. Linear and polynomial kernel functions are used to compute classifiers of low- to high-dimensions, whereas the radial basis kernel function is used to compute classifiers of infinite dimensions [1, 2, 14, 15].

All kernel functions have the tuning parameter designated C (also referred to as the cost) [1, 2, 14, 15]. C applies a penalty to the value of a misclassified observation, and the tuning process involves minimising it while maintaining a maximal distance between the margin and the observation. A polynomial kernel function requires an additional two parameters: the degree of the polynomial function, which defines its dimensionality, and the scale of the function, which is a regularisation parameter that determines the weighted influence of an observation on resampling classification [1-4].

References

[1] Hastie, T., Tibshirani, R., Friedman, J. 2017. *The elements of statistical learning: data mining, inference, and prediction (2nd ed.)*. Springer-Verlag, New York: 9-39. https://link.springer.com/book/10.1007/978-0-387-84858-7.

[2] Mohri, M., Rostamizadeh, A., Talwalkar. 2018. *Foundations of machine learning (2nd ed.)*. The MIT Press, London. https://mitpress.mit.edu/books/foundations-machine-learning-second-edition.

[3] Kuhn, M. 2008. *Building predictive models in R using the caret package*. J. Stat. Softw., 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

[4] Kuhn, M., Johnson, K. 2013. *Applied predictive modeling*. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3.

[5] Agresti, A. 2006. *An introduction to categorical data analysis (2nd ed.)*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/0470114754.

[6] Myers, R.H., Montgomery, D.C., Vining, G.G., Robinson, T.J. 2010. *Generalized linear models: with applications in engineering and the sciences (2nd ed.)*. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470556986.

[7] Gbur, E.E., Stroup, W.W., McCarter, K.S., Durham, S., Young, L.J. Christman, M., West, M., Kramer, M. 2012. *Generalized linear models*. In: Analysis of generalized linear mixed models in the agricultural and natural resources sciences. ASA, CSSA & SSSA. https://doi.org/10.2134/2012.generalized-linear-mixed-models.c3.

[8] Zou, H., Hastie, T. 2003. *Regression shrinkage and selection via the elastic net, with applications to microarrays*. Stanford University. DOI not available.

[9] Zou, H., Hastie, T. 2005. *Regularization and variable selection via the elastic net*. J. R. Statist. Soc. B, 67(2): 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[10] Tibshirani, R. 2011. *Regression shrinkage and selection via the lasso: a retrospective*. J. R. Statist. Soc. B, 73(3): 273-282. https://statweb.stanford.edu/~tibs/ftp/lasso-retro.pdf.

[11] Breiman, L. 2001. *Random forests*. Mach. Learn., 45(1): 5-32. https://doi.org/10.1023/A:1010933404324.

[12] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P. 2003. *Random forest: a classification and regression tool for compound classification and QSAR modeling*. J. Chem. Inf. Comput. Sci., 43(6): 1947-1958. https://doi.org/10.1021/ci034160g.

[13] Biau, G., Scornet, E. 2016. *A random forest guided tour*. TEST, 25(2): 197-227. https://doi.org/10.1007/s11749-016-0481-7.

[14] Cortes, C., Vapnik, V. 1995. *Support-vector networks*. Mach. Learn., 20(1): 273-297. https://doi.org/10.1007/BF00994018.

[15] Boser, B.E., Guyon, I.M., Vapnik, V.N. 1996. *A training algorithm for optimal margin classifiers*. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory. https://doi.org/10.1145/130385.130401.

[16] Crammer, K., Singer, Y. 2001. *On the algorithmic implementation of multiclass kernel-based vector machines*. J. Mach. Learn. Res., 2(1): 265-292. https://dl.acm.org/doi/10.5555/944790.944813.

[17] Doğan, Ü., Glasmachers, T., Igel, C. 2016. *A unified view on multi-class support vector classification*. J. Mach. Learn. Res., 17(1): 1-32. https://dl.acm.org/doi/10.5555/2946645.2946690.

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

Chapter 3: Plant VOCs, and the phylogeny and chemotaxonomy of the genera *Plectranthus* and *Coleus*

Chapter 3A: Summary and background

3A.1) Summary

Two prominent plant genera, *Plectranthus*, and *Coleus*, many species of which are indigenous to southern Africa, have been previously classified as a single genus of the name *Plectranthus.* However, phylogenetic analysis of markers of the plastid genome of subtribe *Plectranthinae* (family: Lamiaceae) has led to the recognition of *Coleus* as a sister taxon to *Plectranthus* [1, 2]. The purpose of this study is to analyse the profiles of foliar volatile organic compounds (VOCs), from the leaves of southern African species of *Plectranthus* and *Coleus*, as predictive markers of genus, using GC×GC-TOFMS and machine learning.

Foliar VOCs from fresh crushed leaves of representative species of each genera (nine *Plectranthus*, six *Coleus*) were extracted using static HS-SPME, and analysed using GC×GC-TOFMS. Profiles of the foliar VOCs for each sample were constructed from their total ion chromatograms, and machine learning algorithms were used to model the data, to make predictions on the genus of new samples, and to tentatively identify putative markers of genus for *Plectranthus* and *Coleus*.

3A.2) Background: the taxonomy of *Plectranthus* and *Coleus*

The genera *Plectranthus* L'Hérit. and *Coleus* Lour. are part of family Lamiaceae, a large and diverse family of angiosperms which includes many commonly known herbs and shrubs of horticultural and phytochemical significance [2-7]. Within Lamiaceae is tribe Ocimeae, and the largest subtribe within Ocimeae is *Plectranthinae* [8, 9]. The latter contains eleven genera, including *Plectranthus* and *Coleus*. Most species of *Plectranthus* and *Coleus* are herbaceous perennials (some semi-succulent) indigenous to paleotropical regions of the globe [3-5].

The taxonomical relationship of the genus *Coleus* in relation to *Plectranthus* is a matter that only recently, in light of genomic evidence, has seen overall consensus [1, 2]. The first descriptions of specimens of *Plectranthus* and *Coleus* were published independently in the late eighteenth century [2], and on the basis of observations on stamen morphology, the two genera

were classified separately [2]. However, in 1962, a revised analysis of stamen morphology proposed subsuming genus *Coleus* under *Plectranthus* [2, 10]. This reclassification was not met with broad consensus [2, 11-22]. The recent phylogenetic data from a number of markers of the plastid genome of subtribe *Plectranthinae*, has led to a more definitive phylogeny in which the genus *Coleus* (along with four other genera) is delimited from *Plectranthus* (along with five other genera) as its own clade [1, 2, 23].

More recently, chemotaxonomical evidence consistent with the new phylogeny has been brought to light in the form of a variety of diterpenoids ($C_{20}H_{32}$), particularly a group of C-14-deoxy abietanes (including spirocoleons, royleanones, *p*-quinomethanes and extended quinone abietanes) characteristic of genus *Coleus* [24]. Though there have been numerous studies and analyses on the volatile composition of individual specimens of *Plectranthus* and *Coleus* [25-28], there appears currently to be no genus-wide data on the VOC metabolomes of members of these two genera, a gap which this study intends to address in a preliminary investigation.

References

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

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

[3] Brits, G. J. 2001. *Indigenous Plectranthus (Lamiaceae) from South Africa as new flowering pot plants*. Acta Horticulturae, 552(18): 165-170. https://doi.org/10.17660/ActaHortic.2001.552.18.

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

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

[6] Venkateshappa, S.M., Sreenath, K.P. 2013. *Potential medicinal plants of Lamiaceae*. AIJRFANS, 3(1): 82-87.

[7] Michel, J., Rani, N.Z.A., Husain, K. 2020. *A review on the potential use of medicinal plants from Asteraceae and Lamiaceae plant family in cardiovascular diseases*. Front Pharmacol., 11(1): 852. https://doi.org/10.3389%2Ffphar.2020.00852.

[8] Pastore, J.F.B., Harley, R.M., Forest, F., Paton, A., van den Berg, C. 2018. *Phylogeny of the subtribe Hyptidinae (Lamiaceae tribe Ocimeae) as inferred from nuclear and plastid DNA*. Taxon, 60(5): 1317-1329. https://doi.org/10.1002/tax.605008.

[9] Zhong, J-S., Li, J., Li, L., Conran, J.G., Li, H-W. 2010. *Phylogeny of Isodon (Schrad. ex. Benth) Spach (Lamiaceae) and related genera inferred from nuclear ribosomal ITS, trnL-trnF region, and rps16 intron sequences and morphology*. Syst. Bot., 35(1): 207-219. https://doi.org/10.1600/036364410790862614.

[10] Morton, J.K. 1962. *Cytotaxonomic studies on the West African Labiatae*. Bot. J. Linn. Soc., 58(372): 231-283. https://doi.org/10.1111/j.1095-8339.1962.tb00896.x.

[11] Blake, S.T. 1971. *A revision of Plectranthus (Labiatae) in Australasia*. Contr. Queensland Herb., 9: 1–120. https://doi.org/10.5479/si.00810282.75.

[12] Codd, L.E. 1975. *Plectranthus (Labiatae) and allied genera in Southern Africa*. Bothalia, 11(4): 371–442. https://doi.org/10.4102/abc.v11i4.1482.

[13] Codd, L.E. 1985. *Plectranthus*. In: Leistner, O.A., (ed.). Flora of Southern Africa. Lamiaceae. Bot. Res. Inst., 28(4): 137–172.

[14] Forster, P.I. 1992. *Five new species of Plectranthus L. Hérit (Lamiaceae) from Queensland*. Austrobaileya 3(4): 729–740. http://www.jstor.org/stable/41738814.

[15] Forster, P.I. 1994. *Ten new species of Plectranthus L'Her. (Lamiaceae) from Queensland*. Austrobaileya, 4(2): 159–186. http://www.jstor.org/stable/41738850.

[16] Forster, P. I. 1997. *Plectranthus amoenus and P. thalassoscopicus (Lamiaceae), new species from north-eastern Queensland, Australia, new species from north-eastern Queensland, Australia*. Austrobaileya, 4(4): 653-660. https://www.jstor.org/stable/41738898.

[17] Paton, A.J., Bramley, G., Ryding, O., Polhill, R.M., Harvey, Y.B., Iwarsson, M., Willis, F., Phillipson, P.B., Balkwill, K., Lukhoba, C.W., Oteino, D., Harley, R.M. 2009. *Lamiaceae (Labiatae)*. In: Beentje, H.J., Ghazanfar, S.A., Polhill, R.M., (eds). Flora of Tropical East Africa. R. Bot. Gard. Kew, London: 430.

[18] Forster, P.I. 2011. *Five new species of Plectranthus L. Hér.(Lamiaceae) from New South Wales and Queensland*. Austrobaileya, 8(3): 387–404. http://www.jstor.org/stable/41965592.

[19] Paton, A.J., Bramley, G., Ryding, O., Polhill, R.M., Harvey, Y.B., Iwarsson, M., Willis, F., Phillipson, P.B., Balkwill, K, Oteino, D., Harley, R.M. 2013. *Lamiaceae*. In: Timberlake, J., (ed.). Flora Zambesiaca, R. Bot. Gard. Kew, 8(8): 346.

[20] Wu, C.Y., Huang, Y.C. 1977. *Coleus*. In: Flora Reipublicae Popularis Sinicae, 66: 536–544.

[21] Cramer, L.H. 1978. *A revision of Coleus (Labiatae) in Sri Lanka (Ceylon)*. Kew Bull., 33(3): 551–561. https://doi.org/10.2307/4109658.

[22] Li, H.W, Hedge I.C. 1994. *Lamiaceae*. In: Wu, Z.Y., Raven, P.H., (eds.). Flora of China, Missouri Botanic Garden, 17: 50–299.

[23] Paton, A.J., Springate, D., Suddee, S., Otieno, D., Grayer, R.J., Harley, M.M., Willis, F., Simmonds, M.S., Powell, M.P., Savolainen, V. 2004. *Phylogeny and evolution of basils and allies (Ocimeae, Labiatae) based on three plastid DNA regions*. Mol. Phylogenet. Evol., 31(1): 277–299. https://doi.or/10.1016/j.ympev.2003.08.002.

[24] Grayer, R.J., Paton, A.J., Simmonds, M.S.J., Howes, M-J.R. 2019. *Differences in diterpenoid diversity reveal new evidence for separating the genus Coleus from Plectranthus*. Nat. Prod. Rep., 38(10): 1720-1728. https://doi.org/10.1039/D0NP00081G.

[25] Ngassoum, M.B., Jirovetz, L., Buchbauer, G., Fleischhacker, W. 2000. *Investigation of essential oils of Plectranthus glandulosus Hook f. (Lamiaceae) from Cameroon*. J. Essent. Oil Res., 13(2): 73-75. https://doi.org/10.1080/10412905.2001.9699615.

[26] Alasbahi, R.H., Melzig, M.F. 2010. *Plectranthus Barbatus: a review of phytochemistry, ethnobotanical uses and pharmacology – part 1*. Planta Med., 76(7): 653-661. Emir. J. Food Agric., 24(2): 137-141. [DOI not available].

[27] Mota, L., Figueiredo, C., Pedro, L.G., Barroso, J.G., Miguel, M.C., Faleiro, M.L., Ascensão. 2014. *Volatile-oils composition, and bioactivity of the essential oils of Plectranthus barbatus, P. neochilus, and P. ornatus grown in Portugal*. Chem. Biodivers., 11(5): 719-732. https://doi.org/10.1002/cbdv.201300161.

[28] Aziz, P., Muhammed, N., Intisar, A., Abid, M.A., Din, M.I., Yaseen, M. 2020. *Constituents and antibacterial activity of leaf essential oil of Plectranthus scutellarioides*. Plant Biosyst., *Preprint: 1-6. https://doi.org/10.1080/11263504.2020.1837279.

Chapter 3B: Plant VOCs

3B.1) Background to plant VOCs

Biogenic VOCs constitute the greater portion of global atmospheric VOCs, with estimated emission rates, from the past 26 years, of approximately 760 Tg [1] to 1150 Tg carbon per year [2]. Of all organisms, terrestrial plants produce VOCs in the greatest abundance — an estimated 90% of VOCs in the atmosphere, equivalent to approximately 400-800 Tg carbon per year [2]. This fact is reflected in the VOC composition of the atmosphere, which is dominated by compounds of plant origin. Isoprene, a simple unsaturated VOC of molecular formula C_5H_8, which is produced in abundance by trees, comprises an estimated 40-70% of total biogenic VOCs [1-3]. The isoprenoids, or terpenes, of molecular formula $(C_5H_8)_n$, are derivatives of isoprene, with monoterpenes (n=2) consisting of 11-15%, and sesquiterpenes (n=3) 3%, of the total biogenic VOC emissions [1-3]. Lower molecular mass VOCs, such as ethanol, acetaldehyde and acetone, make up about 6-10% of the total, and other species, 20-30% of the total [1-3]. Forested regions, in particular, are substantial sources of isoprene and derivative isoprenoids.

Plants produce VOCs as secondary metabolites via a large number of biochemical pathways, and they are emitted from all the major plant organs — roots, leaves and flowers. Plant VOCs may function as molecular signals, or cues, mediating ecological interactions of the plant with other organisms in the environment, such as insects or other plants, of the same or different species [4-10]. For example, the fragrant vapours, or so-called bouquets, of VOCs emitted by flowers, upon anthesis, serve to attract pollinator insects to feed on the flower nectar and thereby promote pollen dispersal [4, 10].

This study is focussed on foliar VOCs, which are produced, stored and emitted by leaves. Foliar VOCs are continuously emitted into the air, however the so-called green-leaf volatiles (GLVs) are only released upon damage of leaf tissue by herbivorous insects, such as caterpillars, aphids or mites, feeding on the leaf material [4, 5, 7]. Once released, they may attract predators to the herbivores, such as wasps, which in turn serve to defend the leaves from consumption [4, 5, 7].

3B.2) Chemotaxonomy and metabolomics: related applications of plant VOC analysis

The VOC profile of a plant (the full set of VOCs produced and emitted by the plant) is typically complex, and the composition of VOCs (types and quantities) produced and emitted by plants is likely to vary to a greater or lesser extent between different taxa [4, 11]. This variability renders plant VOCs potentially useful in the field of chemotaxonomy — the classification of plants according to the chemical profiles of different and/or related specimens. Chemotaxonomic analyses typically focus on a group of a few to many secondary metabolites, from a representative selection of species, genera, or family [11-26]. The general method involves two main steps: 1) secondary molecular metabolites are separated and/or isolated (usually from specific organ tissue), and detected and/or identified, from a selection of related specimens; 2) the relative occurrence of the detected metabolites from one group of samples to the next is compared, and they are organised into groups according to their chemical profiles [12-26]. Often, the results are compared for agreement to known taxonomical models [21-23, 25]. Molecules which are distinctive or highly characteristic of particular taxa are considered to be markers of those taxa [19, 21-25]. A more detailed analysis may involve the mapping of individual markers as phenotypic evolutionary characters on a phylogenetic tree [22, 25]. The major classes of chemotaxonomic compounds include flavonoids, phenols and alkaloids (including tropane alkaloids and iridoids) — also as glycosidic compounds — as well as xanthone derivatives, terpenes and terpenoids [13-26].

Chemotaxonomy overlaps with the field of metabolomics (discussed in Chapter 1.1.3). The full set of metabolites produced by an organism constitute the metabolome of the organism, in the same way that the full set of genes of an organism constitute its genome, and the full set of its proteins its proteome [27, 28]. Metabolomics is concerned with the elucidation of metabolomes, which is to say, the metabolic profiles of organisms. The patterns of secondary metabolite occurrence in plants are a result of evolution and natural selection, and in many cases, overlap between the secondary metabolic profiles of taxa are indicative of biochemical pathways held in common by those taxa, and thus evolutionary relatedness between them [11]. However, such overlap may in some cases arise due to the fact that the same metabolites can be produced by different biochemical pathways evolved independently in unrelated organisms, and are thus a result of convergent evolution [11].

Though chemotaxonomic analysis may provide detailed information on the secondary metabolic profiles of plants, as well as reveal underlying commonalities in terms of the expression of different VOCs by different plant taxa, it is not in itself used for the determination

of evolutionary phylogenetic relationships. Nevertheless, chemotaxonomy provides supporting evidence for phylogenies constructed by genomic systematics, and the chemoinformatic relationships revealed by chemotaxonomic and metabolomic studies may reveal phytochemical groups and relationships of potential practical applicability.

3B.3) The sampling and analysis of plant VOCs

Plant VOCs can be obtained *in situ*, from particular organs (leaves, flowers, roots, or fruit), or *in vivo* from a whole plant [28-33]. Figure 3 summarises different sampling approaches in this regard. Sample analytes can either be extracted, or pre-concentrated, prior to analysis, or they can be introduced directly from the source into an analytical instrument, without prior extraction, for real-time and online analysis. The former approach is more common, and is suited to both targeted and non-targeted analysis using GC-FID, GC-MS [34-64] and GC×GC-TOFMS [28-31, 61]. Direct-introduction techniques do not include a step for chromatographic separation, and are suited to targeted analysis using soft-ionisation mass spectrometric techniques, such as a proton-transfer reaction mass spectrometry (PTR-MS) [28-31, 65-68].

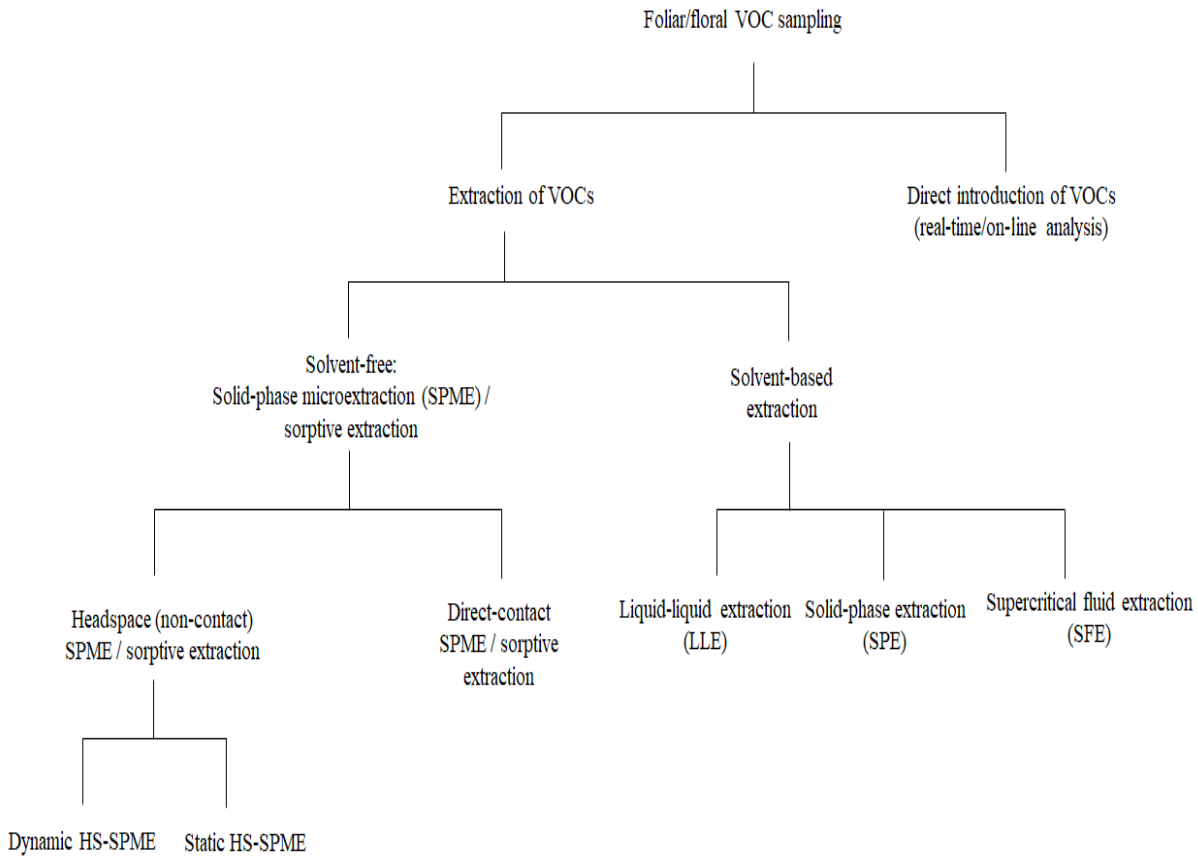


Figure 3: Classification scheme (derived by author) of the methods for sampling plant VOCs.

3B.3.1) Extraction prior to analysis

The practice of plant volatile extraction shares its history in perfumery. The aromatic concentrate that is a precursor material to perfume, called an absolute, is obtained through the traditional technique of *enfleurage*, in which flower petals are pressed and stored in saturated fat, such as lard and tallow, which absorb the aromatic volatile components, forming a waxy pomade, which is then treated with high purity alcohol to extract the volatile components, yielding an *absolute* [32].

In modern laboratory contexts, there are various approaches for the extraction of plant VOCs, depending on the purpose of the experiment. Broadly, these can be divided into solvent-based techniques and solvent-free techniques (Figure 3). The former include liquid-liquid extraction (LLE), steam distillation, supercritical fluid extraction (SFE), and solid-phase extraction (SPE). The latter include solid-phase microextraction, or sorptive extraction.

3B.3.2) Liquid-liquid extraction and steam distillation/hydrodistillation

The modern method for obtaining a concentrated volatile solution is liquid-liquid extraction (LLE) with low-boiling organic solvent [32-34], which is based on the solubility, partitioning and mass transfer of nonpolar solutes in nonpolar solution. Modifications of liquid extraction methods include ultrasound- and microwave-assisted extraction (UAE/MAE), which promote the rapid heating and partitioning of analyte solutes, without degradation, from the sample matrix to solvent, and thus enhance extraction selectivity and efficiency [35-39].

Essential oils are volatile oil mixtures free of involatile vegetal components, such as pigments and tannins, and are of greater purity than absolutes. Essential oils are obtained by methods of distillation — predominantly steam distillation, and its variant: hydrodistillation [32-34, 40]. In steam distillation, the volatile components in the source vegetal material, or extract/mixture derived therefrom, is distilled with pressurised water vapour, and the vapour is condensed as separable nonpolar volatile and hydrosol distillates. The volatile distillate is purified further by LLE, UAE or MAE, or fractional distillation [33-34]. Hydrodistillation is a form of steam distillation in which the source material is present in solution with the heated water steam-source [32-34]. Steam distillation is advantageous in that it affords purer volatile extracts than liquid-liquid extraction, however, it is not suitable for the isolation of labile VOCs, as these are degraded in the distillation process [26, 39].

3B.3.3) Supercritical fluid extraction (SFE)

An important innovation applicable to the extraction of plant VOCs is supercritical fluid extraction (SFE), in which supercritical fluid — a substance within a temperature and pressure range beyond the liquid/vapour critical point — serves as the extraction phase [41, 42]. In most applications, carbon dioxide is used as a supercritical solvent, which has a low critical pressure and temperature, is nonpolar, and easily removed after extraction [34, 43]. SFE permits the adjustment of the density of the supercritical solvent, via temperature-pressure parameters, within the supercritical range, which in turn modifies the analyte selectivity of the solvent. The addition of cosolvents provides an additional means of modifying the properties of the supercritical fluid, thereby permitting finer tuning in terms of analyte selectivity [34, 43].

3B.3.4) Solid-phase extraction (SPE)

Solid-phase extraction (SPE) is a technique based on the relative affinity of analyte molecules for different stationary solid phases and mobile liquid phases, and implements a setup similar to that of a classical chromatographic column — analyte molecules are passed, along with mobile phase, into a solid stationary phase, packed into a vertically-held cartridge, and washed of impurities with solvent systems of graded polarity [44-46]. The clean extracts are usually present at the end of the process either in the mobile or the stationary phase, and require subsequent elution if present in the latter [44].

3B.3.5) Solid-phase microextraction / sorptive extraction

The most widely utilised method for the extraction of plant VOCs for analysis is SPME/sorptive extraction (Chapter 1.2) [47-53]. The advantage of using polymeric sorbents is that they exclude higher molecular weight species, so that there are fewer potential matrix interferents present in the sample. For this reason, sorptive extraction is a more selective technique for the sampling of VOCs. In addition, the preparation and clean-up stages involved in solvent-based methods are obviated.

Sorptive samplers are typically composed of nonpolar and porous organic polymers, such as polydimethylsiloxane (PDMS), alone or in combination with Carboxen (CAR) and/or

divinylbenzene (DVB) [54-56] In a typical application of HS-SPME, the sorbent is a retractable fibre (50-100 microns in diameter) encased in a syringe-like device, which is exposed, after insertion through a septum, to the headspace of the sample, allowing for sufficient accumulation of VOCs into the headspace, and preventing contamination of the sampler by atmospheric components [29-31, 47-53, 57-61]. However, the sampler may take any form suitable to the experimental setup. The headspace conditions can be either static (under equilibrium conditions) or dynamic (under gas-flow conditions) [29-31]. In dynamic headspace sampling, air is pumped through a charcoal filter and into a glass or plastic chamber containing the plant, or parts of the plant, to be sampled, and the VOCs are collected by a sorbent [29-31, 59].

In a contact method, the sampler is brought into physical contact with the vegetal tissue. For foliar VOCs sampling, this can involve, for instance, attaching PDMS tape, or a magnetic stir bar coated with a thin layer of PDMS, to the surface of the vegetal source [62, 63]. If the organ of interest is fruit, the sampler can be inserted into the fruit tissue [65].

3B.4) Real-time and online analysis of plant VOCs

In a direct introduction technique, plant volatiles are neither extracted prior to analysis, nor separated by chromatography, but instead are introduced directly into an analytical instrument for online real-time analysis using proton transfer reaction-mass spectrometry (PTR-MS) [65-68]. In PTR-MS, water vapour is used as a source of hydronium ions (H_3O^+) which undergo protonation reactions with VOCs (those that have a higher proton affinity than water) to form protonated analyte ions, which enter the quadrupole or time-of-flight mass spectrometer [65-68]. The technique is very sensitive, but is limited to targeted analysis of known VOCs with proton affinity, and thus is not suitable for nontargeted analysis of complex samples.

References

[1] Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrakou, T., Müller, J-F., Kuhn, U., Stefani, P., Knorr, W. 2014. *Global data set of biogenic VOC emissions calculated by the MEGAN model over the last 30 years*. Atom. Chem. Phys., 14(17): 9317-9341. https://doi.org/10.5194%2Facp-14-9317-2014.

[2] Guenther, A., Hewitt, C.N., Erickson, D., Fall, R., Geron, C., Graedel, T., Harley, P., Klinger, L., Lerdau, M., Mckay, W.A., Pierce, T., Scholes, B., Steinbrecher, R., Tallamraju, R., Taylor, J., Zimmerman, P. 1995. *A global model of natural volatile organic compound emissions*. J. Geophys. Res. Atmos., 100(D5): 8873-8892. https://doi.org/10.1029/94JD02950.

[3] Guenther, A., Jiang, X., Heald, C.L., Sakulyanontvittaya, T., Duhl, T., Emmons, L.K., Wang, X. 2012. *The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions*. Geosci. Model Dev., 5(2): 1471-1492. http://doi.org/10.5194/gmdd-5-1503-2012.

[4] Maffei, M.E. 2010. *Sites of synthesis, biochemistry and functional role of plant volatiles*. S. Afr. J. Bot., 76(4): 612-631. https://doi.org/10.1016/j.sajb.2010.03.003.

[5] Arimura, G., Ozawa, R., Shimoda, T., Nishioka, T., Boland, W., Takabayashi, J. 2000. *Herbivory-induced volatiles elicit defence genes in lima bean leaves*. Nature, 406(6795): 512-515. https://doi.org/10.1038/35020072.

[6] Knudsen, J.T., Eriksson, R., Gershenzon, J., Ståhl. 2006. *Diversity and distribution of floral scent*. Bot. Rev., 72(1): 1-120. http://doi.org/10.1663/0006-8101(2006)72[1:DADOFS]2.0.CO;2.

[7] Sasso, R., Iodice, L., Digilio, M.C., Carretta, A., Ariati, L., Guerrieri. 2008. *Host-locating response by the aphid parasitoid Aphidius ervi to tomato plant volatiles*. J. Plant Interact., 2(3): 175-183. https://doi.org/10.1080/17429140701591951.

[8] Rasmann, S., Turlings, T.C.J. 2007. *Simultaneous feeding by aboveground and belowground herbivores attenuates plant-mediated attraction of their respective natural enemies*. Ecol. Lett., 10(10): 926-936. https://doi.org/10.1111/j.1461-0248.2007.01084.x.

[9] Rasmann, S., Köllner, T.G., Degenhardt, J., Hiltpold, I., Toepfer, S., Kuhlmann, U., Gershenzon, J., Turlings, T.C.J. *Recruitment of entomopathogenic nematodes by insect-damaged maize roots*. Nature, 434(7034): 732-737. https://doi.org/10.1038/nature03451.

[10] Kost, C., Heil, M. 2006. *Herbivore-induced plant volatiles induce an indirect defence in neighbouring plants*. J. Ecol., 94(3): 619-628. https://doi.org/10.1111/j.1365-2745.2006.01120.x.

[11] Hegnauer, R. 1986. *Phytochemistry and plant taxonomy – an essay on the chemotaxonomy of higher plants*. Phytochemistry, 25(7): 1519-1535. https://doi.org/10.1016/S0031-9422(00)81204-2.

[12] Reynolds, T. 2007. *The evolution of chemosystematics*. Phytochemistry, 68: 2887-2895. https://doi.org/10.1016/j.phytochem.2007.06.027.

[13] Bhargava, V.V., Patel, S.C., Desai, K.S. 2013. *Importance of terpenoids and essential oils in chemotaxonomic approach*. Int. J. Herb Med.

[14] Hao, D.C., Gu, X-J., Xiao, P.G. 2015. *Chemotaxonomy – a phylogeny-based approach*. In: Hao, D.C., Gu, X-J., Xiao, P.G. (eds.). Medicinal plants. Woodhead Publishing, Philadelphia: 1-48. https://doi.org/10.1016/B978-0-08-100085-4.00001-3.

[15] Umoh, O.T. 2020. *Chemotaxonomy: the role of phytochemicals in chemotaxonomic delineation of taxa*. Asian Plant Res. J., 5(1): 43-52. https://doi.org/10.9734/aprj/2020/v5i130100.

[16] Mikanagi, Y., Yokoi, M., Ueda, Y., Saito, N. 1978. *Flower flavonol and anthocyanin distribution in subgenus Rosa*. Biochem. Syst. Ecol. 23(2): 183-200. https://doi.org/10.1016/0305-1978(95)93849-X.

[17] Daniel, M., Sabnis, S.D. 1979. *Chemotaxonomy of Loganiaceae*. Curr. Sci., 48(9): 383-385. https://www.jstor.org/stable/24081112.

[18] Chialva, F., Liddle, P.A.P., Doglia, G. 1983. *Chemotaxonomy of wormwood (Artemisium absinthium* L.*)*. Z Lebensm. Unters. Forsch., 176: 363-366. https://doi.org/10.1007/BF01057728.

[19] Viljoen, A.M., Van Wyk, B-E., Van Heerden, F.R. 1996. *Distribution and chemotaxonomic significance of flavonoids in Aloe (Asphodelaceae)*. Pl. Syst. Evol., 211: 31-42. https://doi.org/10.1007/BF00984910.

[20] Griffin, W.J., Lin, G.D. 2000. *Chemotaxonomy and geographical distribution of tropane alkaloids*. Phytochemistry, 53(6): 623-637. https://doi.org/10.1016/s0031-9422(99)00475-6.

[21] Jensen, S.R., Schripsema, J. 2002. *Chemotaxonomy and pharmacology of Gentianaceae*. In: Struwe, L., Alberts, V. (eds). Gentianaceae – systematics and natural history. Cambridge University Press: 573-631. http://doi.org/10.1017/CBO9780511541865.007.

[22] Kim, S.W., Ban, S.H., Chung, H., Cho, S., Chung, S.J., Choi, P.S., Yoo, O.J., Liu, J.R. 2004. *Taxonomic discrimination of flowering plants by multivariate analysis of Fourier transform infrared spectroscopy data*. Plant Cell Rep., 23: 246-250. https://doi.org/10.1007/s00299-004-0811-1.

[23] Martucci, M.E.P., De Vos, R.C.H., Carollo, C.A., Gobbo-Neto, L. 2013. *Metabolomics as a potential chemotaxonomical tool: applications in the genus Vernonia Schreb*. PLoS One, 9(4): e93149. https://doi.org/10.1371/journal.pone.0093149.

[24] Xiao, C., Wu, M., Chen, Y., Jia, P., Jia, R., Zheng, X. 2014. *Metabolomic analysis provides novel chemotaxonomic characteristics for phenotypic cultivars of tree peony*. Anal. Methods, 6(19): 7854-7864. https://doi.org/10.1039/C4AY01028K.

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

[26] Sarangowa, O., Kanazawa, T., Nishizawa, M., Myoda, T., Bai, C., Yamagishi, T. 2014. *Flavonol glycosides in the petal of Rosa species as chemotaxonomic markers*. Phytochemistry, 107: 61-68. https://doi.org/10.1016/j.phytochem.2014.08.013.

[27] Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L., Bino, R. 2002. *Plant metabolomics: the missing link in functional genomic strategies*. Plant Cell, 14(7): 1437-1440. https://doi.org/10.1105%2Ftpc.140720.

[28] Idle, J.R., Gonzalez, F.J. 2007. *Metabolomics*. Cell Metab., 6(5): 348-351. https://doi.org/10.1016/j.cmet.2007.10.005.

[29] Tholl, D., Boland, W., Hansel, A., Loreto, F., Röse. U.S.R., Schnitzler, J-P. 2006. *Practical approaches to plant volatile analysis*. Plant J., 45(4): 540-560. https://doi.org/10.1111/j.1365-313X.2005.02612.x.

[30] Tholl, D., Hossain, O., Weinhold, A., Röse. U.S.R., Wei, Q. 2021. *Trends and applications in plant volatile sampling and analysis*. Plant J., 106(2): 314-325. https://doi.org/10.1111/tpj.15176.

[31] Cagliero, C., Mastellone, G., Marengo, A., Bicchi, C., Sgorbini, B., Rubiolo, P. 2021. *Analytical strategies for in-vivo evaluation of plant volatile emissions— a review*. Anal. Chim. Acta, 1147(1): 240-258. https://doi.org/10.1016/j.aca.2020.11.029.

[32] Sievers, A.F. 1952. *Methods of extracting volatile oils from plant material and the production of such oils in the United States (5th ed.)*. Technical Bulletin (16), US Department of Agriculture, Washington.

[33] Séquin, M. 2017. *Volatiles of the perfume industry*. In: Encyclopedia of applied plant sciences (2nd ed.). Thomas, B., Murray, B.G., Murphy, D. Elsevier: 393-398. https://doi.org/10.1016/B978-0-12-394807-6.00089-7.

[34] Zhang, Z., Li, G. 2010. *A review of advances and new developments in the analysis of biological volatile organic compounds*. Microchem. J., 95(2): 127-139. https://doi.org/10.1016/j.microc.2009.12.017.

[35] Ferhat, M.A., Tigrine-Kordjani, N., Chemat, S., Meklati, B.Y., Chemat, F. 2007. *Rapid extraction of volatile compounds using a new simultaneous microwave distillation: solvent extraction device*. Chromatographia, 65(3): 217-222. https://doi.org/10.1365/s10337-006-0130-5.

[36] Yong, Y., Wang, Z-M., Wang, Y-T., Li, T-C., Cheng, J-H., Liu, Z-Y., Zhang, H-Q. 2007. *Non-polar solvent microwave-assisted extraction of volatile constituents from dried Zingiber Officinale Rosc*. Chinese J. Chem., 25(3): 346-350. https://doi.org/10.1002/cjoc.200790067.

[37] Delazar, A., Nahar, L., Hamedeyazdan, S., Sarker, S.D. 2012. *Microwave-assisted extraction in natural products isolation*. In: Sarker, S., Nahar, L. (eds.). Natural products isolation. Methods Mol. Biol., 864: 89-115. https://doi.org/10.1007/978-1-61779-624-1_5.

[38] Omar, J., Alonso, I., Garaikoetxea, A., Etxebarria, N. 2013. *Optimization of focused ultrasound extraction and supercritical fluid extraction of volatile compounds and antioxidants from aromatic plants*. Food Anal. Methods, 6(1): 1611-1620. https://doi.org/10.1007/s12161-013-9587-7.

[39] Zhang, H., Yan, H., Li, Q., Lin, H., Wen, X. 2021. *Identification of VOCs in essential oils extracted using ultrasound- and microwave-assisted methods from sweet cherry flower*. Sci. Rep., 11:1167. https://doi.org/10.1038/s41598-020-80891-0.

[40] Scott, R.P.W. 2005. *Essential oils*. In: Worsfold, P., Townshend, A., Poole, C. Encyclopedia of Analytical Science (2nd ed.). Elsevier: 554-561. https://doi.org/10.1016/B0-12-369397-7/00147-3.

[41] Wenclawiak, B. 1992. *SFC and SFE: an introduction for novices*. In: Wenclawiak, B. (ed.). Analysis with supercritical fluids: extraction and chromatography. Springer-Verlag, Berlin: 1-8. https://doi.org/10.1007/978-3-642-77474-4_1.

[42] Clifford, A.A. 1993. *Introduction to supercritical fluid extraction in analytical science*. In: Westwood, S.A. Supercritical fluid extraction and its use in chromatographic sample preparation. Springer-Science, Glasgow: 1-38. https://doi.org/10.1007/978-94-011-2164-4.

[43] Pourmortazavi, S.M., Hajimirsadeghi, S.S. 2007. *Supercritical fluid extraction in plant essential and volatile oil analysis*. J. Chromatogr. A, 1163(1-2): 2-24. https://doi.org/10.1016/j.chroma.2007.06.021.

[44] Fritz, J.S. 1999. *Analytical solid-phase extraction*. Wiley-VCH, New York.

[45] Tinjan, P., Jirapakkul, W. 2007. *Comparative study on extraction methods of free and glycosidically bound volatile compounds from kaffir lime leaves by solvent extraction and solid phase extraction*. Kasetsart J. (Nat. Sci.), 41(5): 300-306. https://li01.tci-thaijo.org/index.php/anres/issue/view/16739.

[46] Riachi, L.G., Abi-Zaid, I.E., Moreira, R.F.A., De Maria, C.A.B. 2012. *Volatile composition of peppermint (Mentha piperita L.) commercial teas through solid phase extraction*. Arch. Latinoam. Nutr., 62(4): 389-392.

[47] Arthur, C.L., Pawliszyn, J. 1990. *Solid phase microextraction with thermal desorption using fused silica optical fibers*. Anal. Chem., 62(19): 2145-2148. http://doi.org/10.1021/ac00218a019.

[48] Zhang, Z., Pawliszyn, J. 1993. *Headspace solid-phase microextraction*. Anal. Chem., 65(14): 1843-1852. https://doi.org/10.1021/ac00062a008.

[49] Pawliszyn, J. 2000. *Theory of solid-phase microextraction*. J. Chromatogr. Sci., 38(7): 270-278. https://doi.org/10.1093/chromsci/38.7.270.

[50] Bicchi, C., Cordero, C., Iori, C., Rubiolo, P. 2000. *Headspace sorptive extraction (HSSE) in the headspace analysis of aromatic and medicinal plants*. J. High Resolut. Chromatogr., 23(9): 539-546. https://doi.org/10.1002/1521-4168(20000901)23:9%3C539::AID-JHRC539%3E3.0.CO;2-3.

[51] Augusto, F., Valente, A.L.P. 2002. *Applications of solid-phase microextraction to chemical analysis of live biological samples*. Trends Analyt. Chem., 21(6-7): 428-438. https://doi.org/10.1016/S0165-9936(02)00602-7.

[52] Belliardo, F., Bicchi, C., Cordero, C., Liberto, E., Rubiolo, P., Sgorbini, B. 2006. *Headspace-solid-phase microextraction in the analysis of the volatile fraction of aromatic and medicinal plants*. J. Chromatogr. Sci., 44(7): 416-429. https://doi.org/10.1093/chromsci/44.7.416.

[53] Zhu, F., Xu, J., Ke, Y., Huang, S., Zeng, F., Luan, T., Ouyang, G. 2013. *Applications of in vivo and in vitro solid-phase microextraction techniques in plant analysis: a review*. Anal. Chim. Acta, 794: 1-14. https://doi.org/10.1016/j.aca.2013.05.016.

[54] Bicchi, C., Drigo, S., Rubiolo, P. 2000. *Influence of fibre coating in headspace solid-phase microextraction-gas chromatographic analysis of aromatic and medicinal plants*. J. Chromatogr. A, 892(1-2): 469-485. https://doi.org/10.1016/S0021-9673(00)00231-4.

[55] Adam, M., Juklová, Bajer, T., Eisner, A., Ventura, K. 2005. *Comparison of three different solid-phase microextraction fibres for analysis of essential oils in yacon (Smallanthus sonchifolius) leaves*. J. Chromatogr. A, 1084(1-2): 2-6. https://doi.org/10.1016/j.chroma.2005.05.072.

[56] Pinheiro, G.P., Galbiatti, M.I., Carneiro, M.J., Sawaya, A.C.H.F. 2019. *Comparison of four different solid-phase microextraction fibers for analysis of Plectranthus amboinicus (Lour.) Spreng. Leaf volatiles*. Ad. Med. Plant Res., 7(2): 38-43. https://doi.org/10.30918/AMPR.72.19.020.

[57] Xiong, G., Goodridge, C., Wang, L., Chen, Y., Pawliszyn, J. 2003. *Microwave-assisted headspace solid-phase microextraction for the analysis of bioemissions from Eucalyptus citriodora leaves*. J. Agric. Food Chem., 51(27): 7841-7847. https://doi.org/10.1021/jf0346105.

[58] Flamini, G., Cioni, P.L., Morelli, I. 2005. *Composition of the essential oils and in vivo emission of volatiles of four Lamium species from Italy: L. purpureum, L. hybridum, L. bifidum and L. amplexicaule*. Food Chem., 91(1): 63-68. https://doi.org/10.1016/j.foodchem.2004.05.047.

[59] Stashenko, E.E., Jaramillo, B.E., Martínez, J.R. 2004. *Analysis of volatile secondary metabolites from Colombian Xylopia aromatica (Lamarck) by different extraction and headspace methods and gas chromatography*. J. Chromatogr. A, 1025(1): 105-113. https://doi.org/10.1016/j.chroma.2003.10.059.

[60] Guo, F-Q., Huang, L-F., Zhou, S-Y., Zhang, T-M., Liang, Y-Z. 2006. *Comparison of the volatile compounds of Atractylodes medicinal plants by headspace solid-phase microextraction-gas chromatography-mass spectrometry*. Anal. Chim. Acta, 570(1): 73-78. https://doi.org/10.1016/j.aca.2006.04.006.

[61] Naudé, Y., Makuwa, R., Maharaj, V. 2016. *Investigating volatile compounds in the vapour phase of (1) a hot water infusion of rhizomes, and of (2) rhizomes of Siphonochilus aethiopicus using head space solid phase microextraction and gas chromatography with time of flight mass spectrometry*. S. Afr. J. Bot., 106: 144-148. https://doi.org/10.1016/j.sajb.2016.07.006.

[62] Boggia, L., Sgorbini, B., Bertea, C.M., Cagliero, C., Bicchi, C., Maffei, M.E., Rubiolo, P. 2015. *Direct contact-sorptive tape extraction coupled with gas chromatography-mass spectrometry to reveal volatile topographical dynamics of lima bean (Phaseolus lunatus L.) upon herbivory by Spodoptera littoralis Boisd.*. BMC Plant Biol., 15: 102. https://doi.org/10.1186/s12870-015-0487-4.

[63] Kfoury, N., Scott, E., Orians, E., Robbat, A., Jr. 2017. *Direct contact sorptive extraction: a robust method for sampling plant volatiles in the field*. J. Agric. Food Chem., 65(38): 8501-8509. https://doi.org/10.1021/acs.jafc.7b02847.

[64] Risticevic, S., Souza-Silva, E.A., DeEll, J.R., Cochran, J., Pawliszyn, J. 2015. *Capturing plant metabolome with direct-immersion in vivo solid phase microextraction of plant tissues*. Anal. Chem., 88(2): 1266-1274. https://doi.org/10.1021/acs.analchem.5b03684.

[65] Hansel, A., Jordan, A., Holzinger, R., Prazeller, Vogel, W., Lindinger, W. 1995. *Proton transfer reaction mass spectrometry: on-line trace gas analysis at the ppb level*. IJMSI, 149-150: 609-619. https://doi.org/10.1016/0168-1176(95)04294-U.

[66] Steeghs, M., Bais, H.P., de Gouw, J., Goldan, P., Kuster, W., Northway, M., Fall, R., Vivanco, J.M. 2004. *Proton-transfer-reaction mass spectrometry as a new tool for real time analysis of root-secreted volatile organic compounds in Arabidopsis*. Plant Physiol., 135(1): 47-58. http://doi.org/10.1104/pp.104.038703.

[67] Maleknia, S.D., Bell, T.L., Adams, M.A. 2007. *PTR-MS analysis of reference and plant-emitted volatile organic compounds*. Int. J. Mass Spectrom., 262(3): 203-210. https://doi.org/10.1016/j.ijms.2006.11.010.

[68] Farré-Armengol, G., Filella, I., Lluisa, J., Primante, C., Peñuelas, J. 2015. *Enhanced emissions of floral volatiles by Diplotaxis erucoides (L.) in response to folivory and florivory by Pieris brassicae (L.)*. Biochem. Syst. Ecol., 63(1): 51-58. https://doi.org/10.1016/j.bse.2015.09.022.

Chapter 4: Methods and materials

4.1.1) Ethical considerations

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

4.1.2) Reagents and chemical standards

Methanol, acetone, acetonitrile and *n*-hexane were purchased from Merck, Pretoria, South Africa. The solution of *n*-alkanes (C_8-C_{28}) used for the calculation of linear retention indices of selected analytes (Chapter 4.5) was purchased from Merck, Pretoria, South Africa.

4.2) Sample population for foliar VOC sampling

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

Fifteen species were included in the study. Of genus *Plectranthus*, nine species were included in the study: *P. ambiguus* (Bolus) Codd, *P. chimanimanensis* S.Moore, *P. ecklonii* Benth., *P. fruticosus* L'Hér., *P. oertendahlii* T.C.E.Fr., *P. saccatus* Benth., *P. strigosus* Benth. ex E.Mey. *P. verticillatus* (L.f.) Druce, and *P. zuluensis* T.Cooke, Bull. Of genus *Coleus*, the six species included were: *C. hadiensis* (Forssk). A.J. Paton, *C. hereroensis* (Engl.) A.J.Paton, *C. livingstonei* A.J.Paton, *C. longipetiolatus* Gürke, *C. madagascariensis* (Pers.) A.Chev., and *C. neochilus* (Schltr.) Codd. Samples of each species were taken in triplicate. Although replicates were taken, each was treated as a discrete sample (i.e.: replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning (Chapter 4.6.4), giving a total sample size of 45. Air blanks and a soil blank were taken to aid in accounting for air and soil VOC contaminants.

Species were identified using species descriptions [1] and herbarium specimens. Voucher specimens of the plants sampled were submitted for record to the H.G.W.J Schweickerdt Herbarium (PRU) of the University of Pretoria.

4.3) HS-SPME of foliar VOCs

Fresh leaves were picked and removed of petioles, weighed to a mass of two grams, and crushed with a mortar and pestle. The crushed leaf matter was enclosed in a glass vial (40 mL) with a screw cap with a central hole (3.2 mm radius) lined with Teflon® septa (Separations, South Africa). The vial was left to stand in a water bath at 40˚C, for 15 minutes, to allow for equilibration of the system to occur. Headspace extraction of the foliar VOCs was performed using a SPME device with a fused-silica fibre coated with a 100-micron-diameter layer of polydimethylsiloxane (PDMS) (Supelco, Sigma-Aldrich®, Kempton Park, South Africa). The extraction time was 15 minutes, and the temperature of extraction 40˚C. Thermal desorption (at 230˚C for 5 minutes) of sorbed analytes from the SPME fibre for instrumental analysis by GC×GC-TOFMS (Chapter 4.4) was performed in the GC inlet directly after sampling. The fibre was conditioned at 280˚C in split mode (50:1) for 20 minutes prior to extraction. The extraction method and conditions were also applied to the air blanks and soil blank.

4.4) Instrumental and analytical methods: comprehensive GC×GC-TOFMS

Comprehensive two-dimensional chromatographic separation and mass spectrometric analysis was performed on a LECO® Pegasus® 4D GC×GC-TOFMS with an Agilent® 7890A chromatograph and a dual quad-jet cryogenic modulator (LECO®, Kempton Park, South Africa), operated by ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®). The hot jets were operated with nitrogen gas produced by a nitrogen gas generator (Peak Scientific, South Africa), and the cold jets were operated with nitrogen gas cooled with liquid nitrogen (Afrox, South Africa). The primary (1D) column was an Rxi-1MS apolar capillary column of length 30 m, 250 µm ID and 0.25 µm film thickness; the secondary (2D) column was an Rxi-17SilMS mid-polar capillary column of length 0.760 m, 250 µm ID and 0.25 µm film thickness (Restek, Bellefonte, PA, USA).The carrier gas (ultra-high purity grade helium [Afrox, Gauteng, South Africa]) flow rate was constant at 1.4 mL/min, with a front inlet septum purge flow rate of 3 mL/min, a splitless time of 30 s, and a front inlet purge flow rate of 30 mL/min , giving a total front inlet flow of 31.4 mL/min.

The initial temperature for the primary oven was held at 40˚C for 1.5 min, and ramped to 280˚C (hold time 3 min) at a rate of 6˚C/min. The total run time was 44.5 min. The temperature programme rate for the secondary oven and the modulator was the same as that of the primary oven, but offset by +5˚C and +15˚C respectively. The transfer line to the TOFMS and the front

inlet of the chromatograph were maintained at a temperature of 280˚C and 230˚C respectively. The modulation period was 2 s, with a hot pulse time of 0.6 s and a cool time between stages of 0.4 s.

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

4.5) Data acquisition and processing

ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®) was used for data acquisition and chromatographic peak alignment. A S/N threshold of 100 was set, and deviations in retention time were bound within the modulation period (2s), for 1D peaks, and 0.1s for 2D peaks. Tentative identification of the analyte compounds was achieved by comparison of experimental mass spectra with reference spectra of the National Institute of Standards and Technology (NIST) library (version 2.2), with the minimum similarity threshold for a match set at 75%. Chromatographic peak areas were normalised during the processing step prior to statistical and machine learning analysis (Chapter 4.6.4).

Retention indices of reported analyte compounds were calculated using the method of the linear temperature programmed retention index, as developed by Van den Dool and Kratz [2], using a series of *n*-alkanes (C_8-C_{28}; Merck, Pretoria, South Africa). Due to column maintenance, and the fact that the *n*-alkane series was injected after replacement of the secondary column, some samples (*C. neochilus*, *C. madagascariensis*, *C. hereroensis*, *C. hadiensis*, *P. fruticosus*, *P. oertendahlii*, *P. verticillatus* were re-run in order to obtain the appropriate retention indices with respect to the *n*-alkanes.

4.6) Data analysis: statistics and machine learning

4.6.1) Dataset processing prior to statistical analysis

The peak area datasets produced by comprehensive GC×GC-TOFMS analysis were processed in four steps prior to statistical analysis:

1) Contaminant compounds were removed from each individual sample data set, including organosiloxanes, halogens, boronic compounds and metallic complexes.

2) Tentatively identified compounds, with corresponding peak area values, from all samples, were combined into a single data set using the VLOOKUP function of Microsoft® Excel® (version 16.0.14228.20204).

3) Peak area values for compounds which were not detected in a particular sample were assigned a value of zero for that sample

4) Blank corrections were performed, for the soil-blank with respect to the air-blank (triplicate-averaged), and for each replicate with respect to the air- and soil-blanks. Compounds with resultant negative values (of which there were 192) were removed, as these were contaminants (of no informative import) occurring to a greater extent in the air-blank than in the samples.

4.6.2) Preliminary statistical analysis: principal component and linear discriminant analysis

The four-step processing resulted in a dataset (N = 45) of 1794 compounds. In order to determine the efficacy of the air- and soil-blank correction, and as a preliminary assessment of the variation across the sample data, a Principal Component Analysis (PCA), with a Wide estimation, and a Linear Discriminant Analysis (LDA), with a Wide Linear fit, were performed on the final dataset, using JMP® (Version 15.0.0) statistical software.

4.6.3) Machine learning (regression and classification)

Machine learning was performed using R© computational and statistical software (version 1.3.959) with the Classification and Regression Training (caret) package (version 6.0-86) [3]. The method was based on that of Kuhn, 2008 [4]. The seed function was set at 95 for all computations. Three algorithms were used to construct regression and classification models of the data: an elastic-net regression (using the glmnet algorithm), a random forest (ranger) and a support vector machine (svmPoly). The analysis pipeline outlined in Chapter 2.3 (Figure 4) was followed for each algorithm:

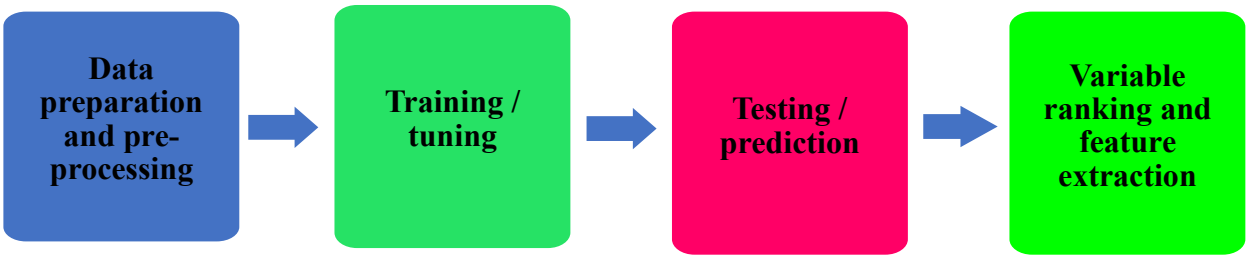


Figure 4: The machine learning pipeline followed for the training and testing of the elastic-net regression (glmnet), random forest (ranger) and support vector machine (svmPoly) algorithms (Chapter 2.3). Adapted from the method of Kuhn, 2008 [5].

4.6.4) Dataset splitting and pre-processing

The dataset (after contaminant removal and blank correction) was shuffled and randomly split with a 0.5 ratio into a training and testing set, such that there was an approximately equal proportion of samples of the same genus in each set. For the results of each of the three models to be comparable, the same train/test split was used for the training and testing of each model. After the preliminary statistical analysis, and prior to machine learning, the training dataset was pre-processed, using pre-processing functions of the caret package [4], in two steps:

1) The data was normalised, i.e.: centred (by subtracting the peak area mean of each variable from the sample peak area values) and scaled (by dividing the result of centring by the standard deviation of the peak area mean).

2) Variables that had zero or near-zero variance were removed from the dataset. Pre-processing resulted in the removal of 1160 compound variables, and 634 variables were retained, centred and scaled. The same pre-processed dataset was used for the training of each of the three models. Although replicates were taken, each was treated as a discrete sample (ie: replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning.

4.6.5) Model tuning and training

The parameters and coefficients of each model were computed and optimised using the resampling trainControl function of caret to perform a five-fold, five-times-repeated cross-validation. The function selects those model parameters with the highest AUC/ROC values as determined by cross-validation.

For the elastic-net regression, using the glmnet algorithm of the caret package, the tuning parameters are alpha (α) and lambda (λ); alpha is equal to the fractional contribution of the lasso-regression penalty to the total ridge-lasso penalty on the regression coefficients of the

model, and lambda is the magnitude of the penalty (Chapter 2.4.1). Five values of alpha, between zero and one, and ten values of lambda, between 0.0001 and one, were used to tune the model.

For the random forest, using the ranger algorithm, the tuning parameters are mtry (the number of randomly selected variables used at each nodal split), the split rule for the splitting of each node, and the minimal node size [5]. The values of mtry used to tune the model were 2, 3 and 5; the split rules used were gini and extratrees, and the minimal node sizes used were 1, 3 and 5.

For the support vector machine, using the svmPoly algorithm, the tuning parameters are the degree and scale of the kernel function, and C (Chapter 2.4.3). The values of C used to tune the model were 0.01, 0.1, 1 and 10, and values of 1-3 were set for the degree and scale.

4.6.6) Variable importance and feature selection

Predictors included in model training were ranked in terms of their relative importance (Chapter 2.3.4). For the glmnet algorithm, variables are ranked according to the values of the regression coefficients of the final tuned model [5]. For the ranger algorithm, the predictors are ranked according to predictive accuracy from out-of-bag resampling [5]. For the svmPoly algorithm, the AUC of the ROC from in-training resampling is used to rank the variables [5]. For each model, the twenty highest ranking compounds are listed, and retention index values (Chapter 4.5) for these compounds, where applicable, were calculated. Peak area values of the top variables were normalised (i.e.: centred and scaled) and visualised as a heatmap, using the heatmap.2 function of the gplots package (version 3.1.3).

References

[1] Van Jaarsveld, E.J., Thomas, V. 2006. *The Southern African Plectranthus and the art of turning shade into glade.* Fernwood Press, Simon's Town.

[2] Van den Dool, H., Kratz, P.D. 1963. *A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography*. J. Chromatogr., 11: 463-471. https://doi.org/10.1016/S0021-9673(01)80947-X.

[3] Kuhn, M. 2019. *The caret Package*. Available from: http://topepo.github.io/caret/index.html [Accessed: 23/11/2021].

[4] Kuhn, M. 2008. *Building predictive models in R using the caret package*. J. Stat. Softw., 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

[5] Kuhn, M. 2018. *Caret: classification and regression training*. Available from: https://cran.r-project.org/web/packages/caret/index.html [Accessed: 23/11/2021].

Chapter 5: Results and discussion

5.1) Chromatographic data from comprehensive GC×GC-TOFMS

Total ion chromatograms (TICs) of foliar compounds from the leaves of southern African *Plectranthus* and *Coleus* were obtained by comprehensive GC×GC-TOFMS. Chromatograms taken from each species included in the study, for both dimensions of retention (1D and 2D), are presented in Figures 7-36.

In order to ascertain whether there is a notable difference between the chromatograms of crushed- and whole-leaf samples, foliar extractions from *C. neochilus*, for both situations, were performed using static HS-SPME. Figure 5 shows overlaid TICs, for the first (1D) and second (2D) dimensions of separation, corresponding to the two situations. The peak distribution for both is very similar, however the crushed-leaf TIC has an overall greater peak intensity, and in addition, has greater peak density (greater number of peaks). Those peaks common to the whole- and crushed-leaf samples may correspond to compounds that are passively and regularly emitted, and thus form part of the normal odour profile of the leaves. The additional peaks observed in the crushed-leaf sample, particularly in the 1D retention time region of ±200-850 seconds, likely correspond to green-leaf volatiles (GLVs), which are sequestered within the leaf, and released upon damage of the tissue (caused, for example, by feeding herbivores) [1-3]. GLVs are reported to be C_6 unsaturated and/or oxygenated species (including compounds such as hexanal and (E)-2-hexenal) and terpenes [1-3]. As discussed in Chapter 5.7, certain C_6 species tentatively identified in this study, including hexanal, as well as certain C_8-C_{10} species, are ranked by the machine learning models employed as top predictors of genus.

Figure 6 presents the contour and surface plots of the 2D chromatograms of a single replicate of *C. neochilus*, which is representative of the chromatographic trend observed across the samples of both genera. There are two regions with prominent peaks— the first lies in the 1D retention time range of ±430-700 seconds, and the second in the range of ±1040-1300 seconds. The first region is populated by peaks of the monoterpenes ($C_{10}H_{16}$) and monoterpenoids, and the second region by peaks of the sesquiterpenes ($C_{15}H_{24}$) and sesquiterpenoids (with respect to the second region). The first region also consists of peaks from C_6-C_{10} species, which are likely to be GLVs (as discussed above), which have been reported to consist of terpenes in addition to C_6 species [1, 3]. No peaks with retention times greater than 1500 seconds are observed, indicating a low occurrence of VOCs of greater than fifteen carbon units.

Comparison of the TICs from species of *Plectranthus* and *Coleus* reveals inter-species differences in terms of chromatographic structure, which is to say, the complexity (number and distribution) and intensity of peaks. Some members of *Coleus* — *C. livingstonei* (Figure 11 and Figure 12), *C. longipetiolatus* (Figure 13 and Figure 14) and *C. neochilus* (Figure 17 and Figure 18) — and one member of *Plectranthus* — *P. fruticosus* (Figure 25 and Figure 26) — have TICs of significant complexity and intensity in both the monoterpene and sesquiterpene regions. *C. hereroensis* (Figure 9 and Figure 10), *P. chimanimanensis* (Figure 21 and Figure 22), *P. saccatus* (Figure 29 and Figure 30) and *P. ecklonii* (Figure 23 and Figure 24) have comparatively moderate, but still significant peak density in these regions. Two of the species from each genus, *C. hadiensis* (Figure 7 and Figure 8) and *P. zuluensis* (Figure 35 and Figure 36), are comparable in so far as they have higher peak density in the region of the sesquiterpenes and sesquiterpenoids. Four species (*C. madagascariensis* [Figure 15 and Figure 16], *P. ambiguus* [Figure 19 and Figure 20], *P. verticillatus* [Figure 33 and Figure 34], and *P. strigosus* [Figure 31 and Figure 32]) have comparatively sparse chromatograms, with *C. madagascariensis*, *P. strigosus* and *P. verticillatus* showing little to no peaks for the sesquiterpenes. The chromatograms of *P. fruticosus* are of a complexity more comparable to members of *Coleus* such as *C. neochilus*, *C. livingstonei*, and *C. longipetiolatus*. It should be noted that column maintenance during the analysis period caused wraparound to occur in the 2D chromatograms of *C. livingstonei*, *C. longipetiolatus*, *P. ambiguus* and *P. saccatus*.

An overloaded peak at ±100 seconds in the first dimension, and between 0.4 and 1 seconds in the second dimension, is present in all samples, and is partly due to acetone, which was used for the cleaning of the SPME vials and implements used to handle and crush the leaves. However, this region appears also to contain highly volatile species such as acetaldehyde, a known GLV [1,3], as can be observed from peaks present in this region for the whole-leaf sample of *C. neochilus* (Figure 5). The chromatograms of all species of both genera show a peak at 500 seconds in the first dimension, and between ±0.8-1.5 seconds in the second dimension, which in many cases correspond to isomers of the monoterpene, pinene, which is commonly found in the plant kingdom.

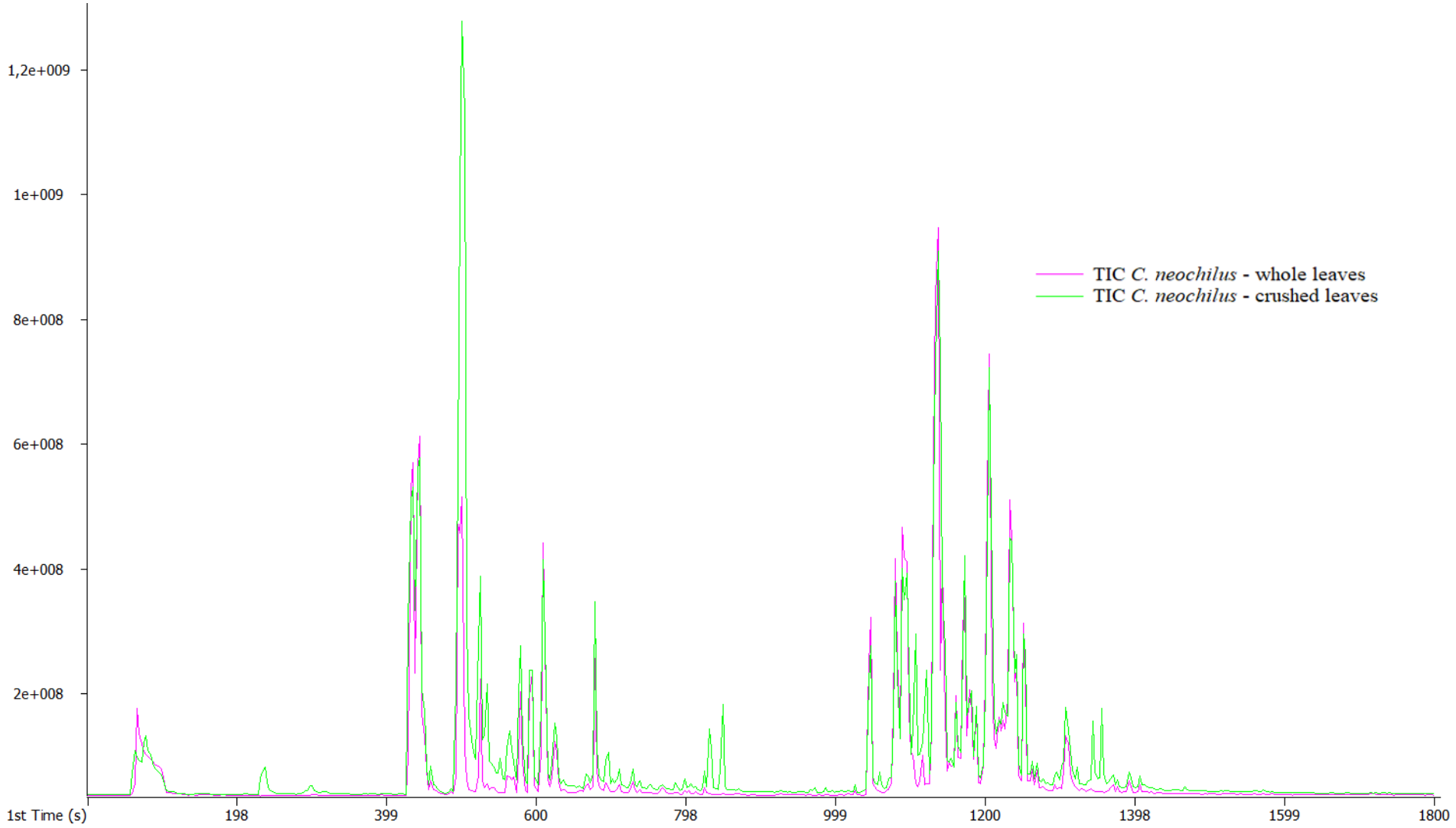


Figure 5: 1D TIC overlay of whole-leaf (pink) and crushed-leaf (green) samples of *C. neochilus*. The crushed-leaf TIC appears to be characterised by peaks of greater height, and an overall greater peak abundance.

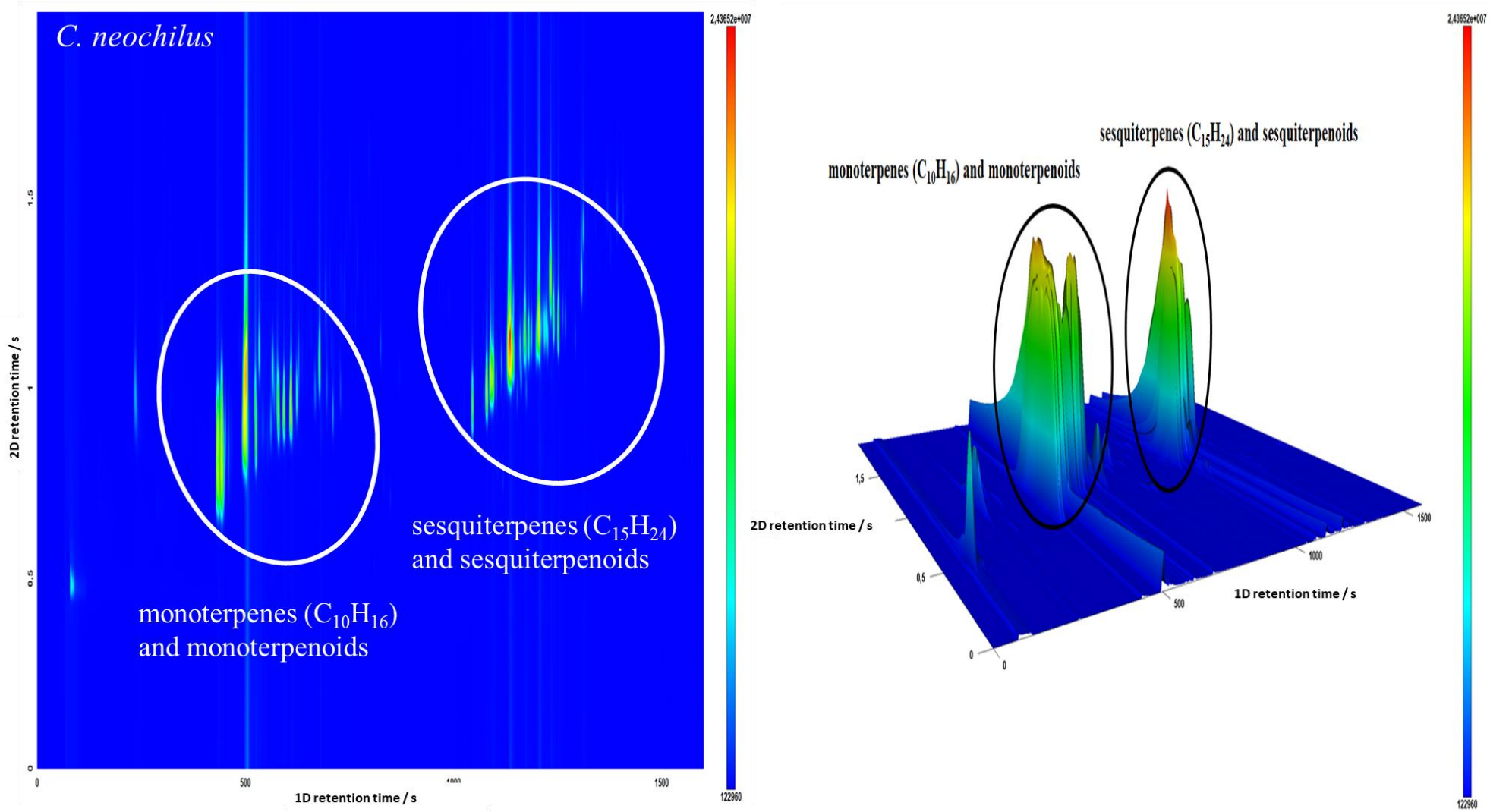


Figure 6: 2D TIC contour (left) and surface (right) plots of replicate extraction 1 of *C. neochilus*, illustrating the two peak regions, which occur to some extent in all the samples, characteristic of monoterpenes/monoterpenoids and the sesquiterpenes/sesquiterpenoids.

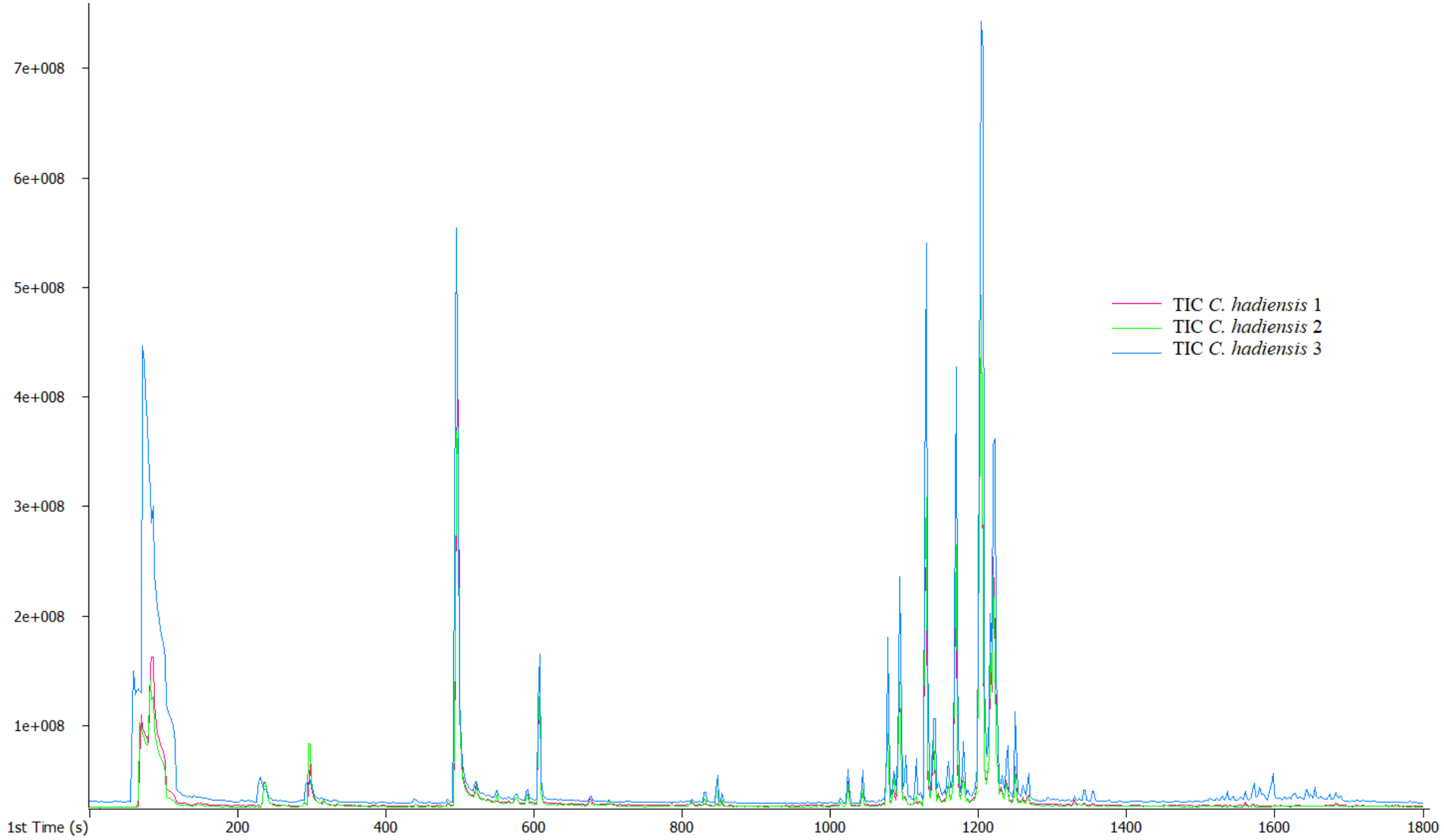


Figure 7: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. hadiensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

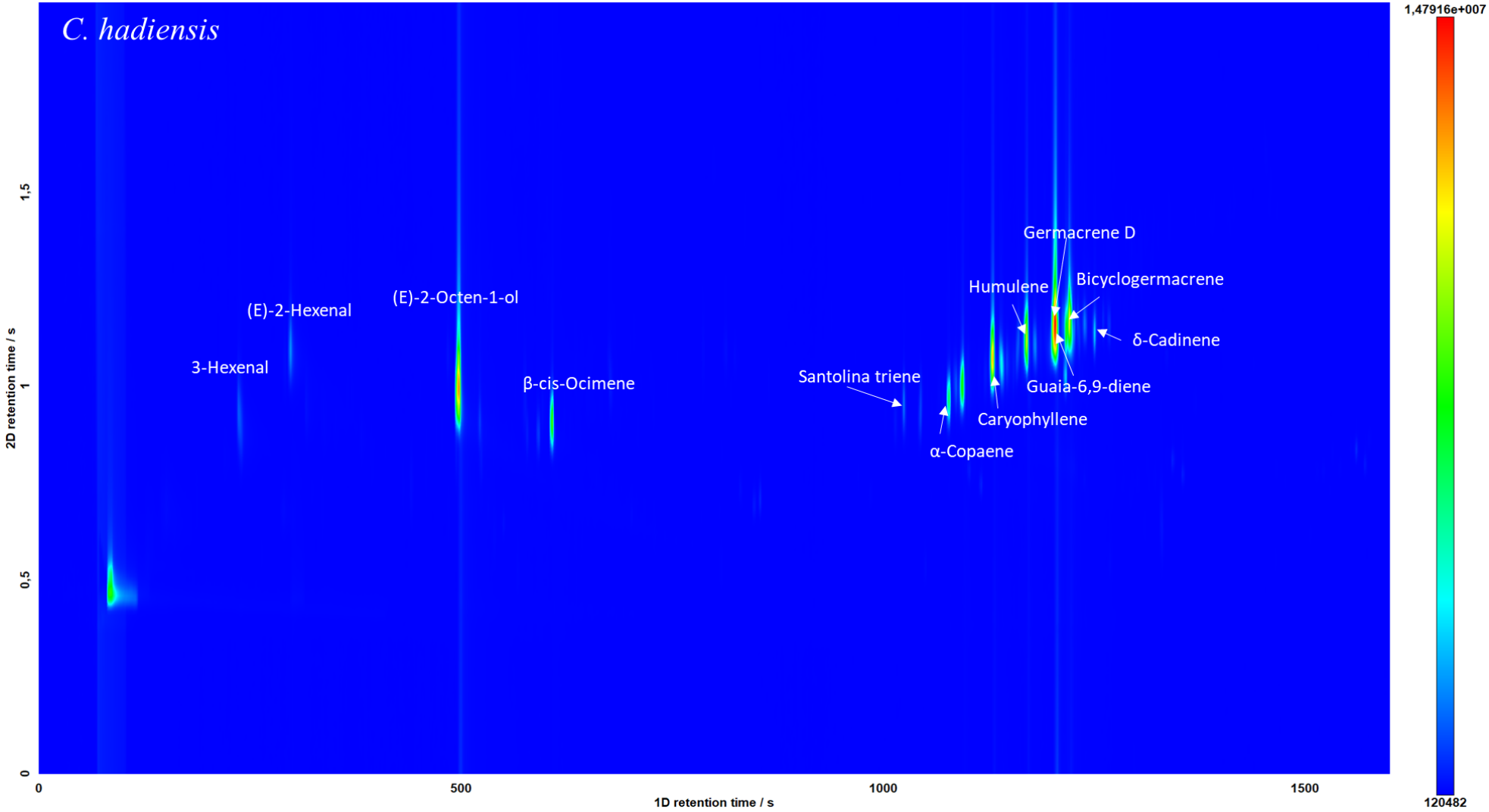


Figure 8: 2D TIC contour plot of replicate extraction 1 of *C. hadiensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

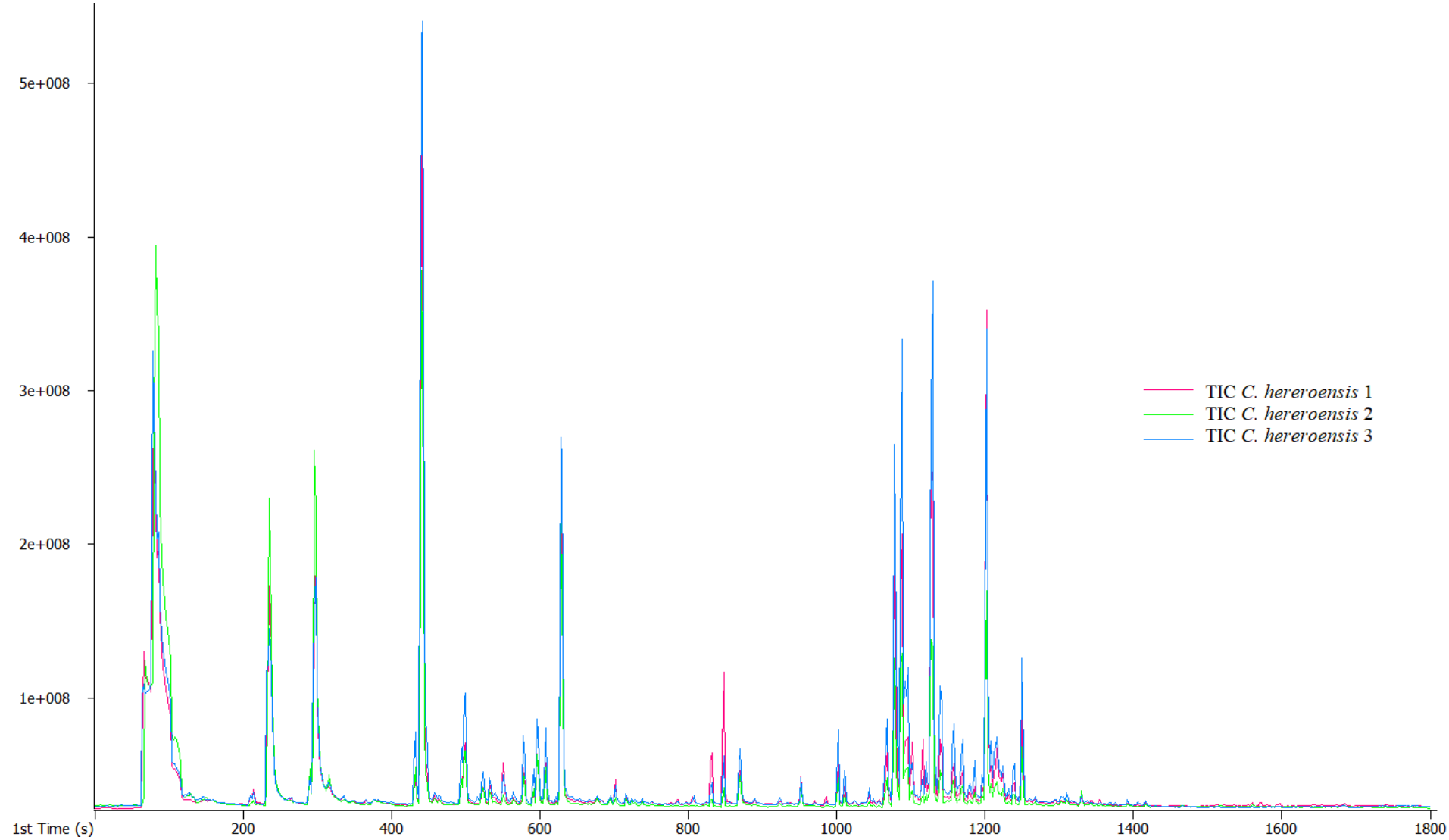


Figure 9: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. hereroensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

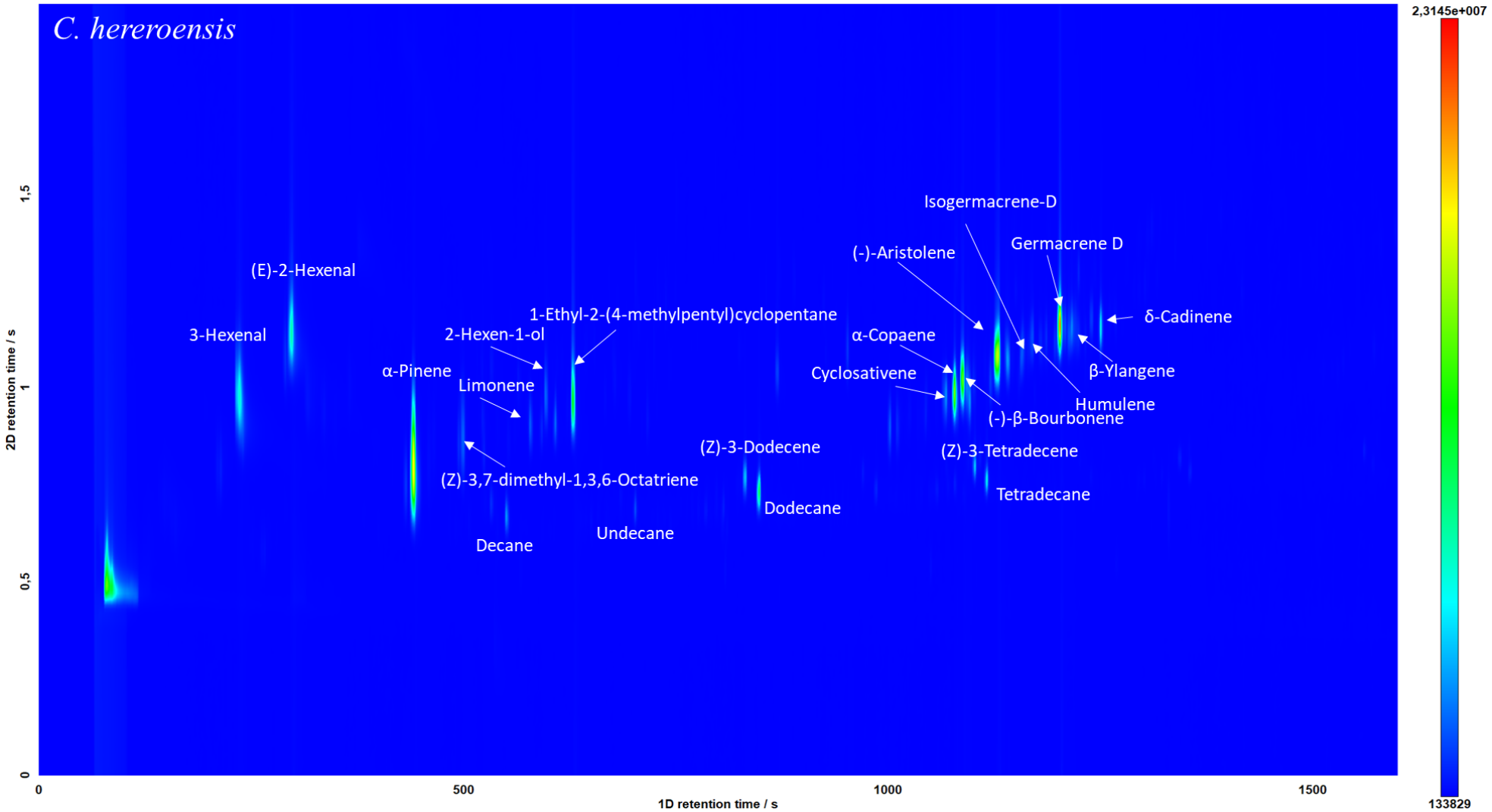


Figure 10: 2D TIC contour plot of replicate extraction 1 of *C. hereroensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

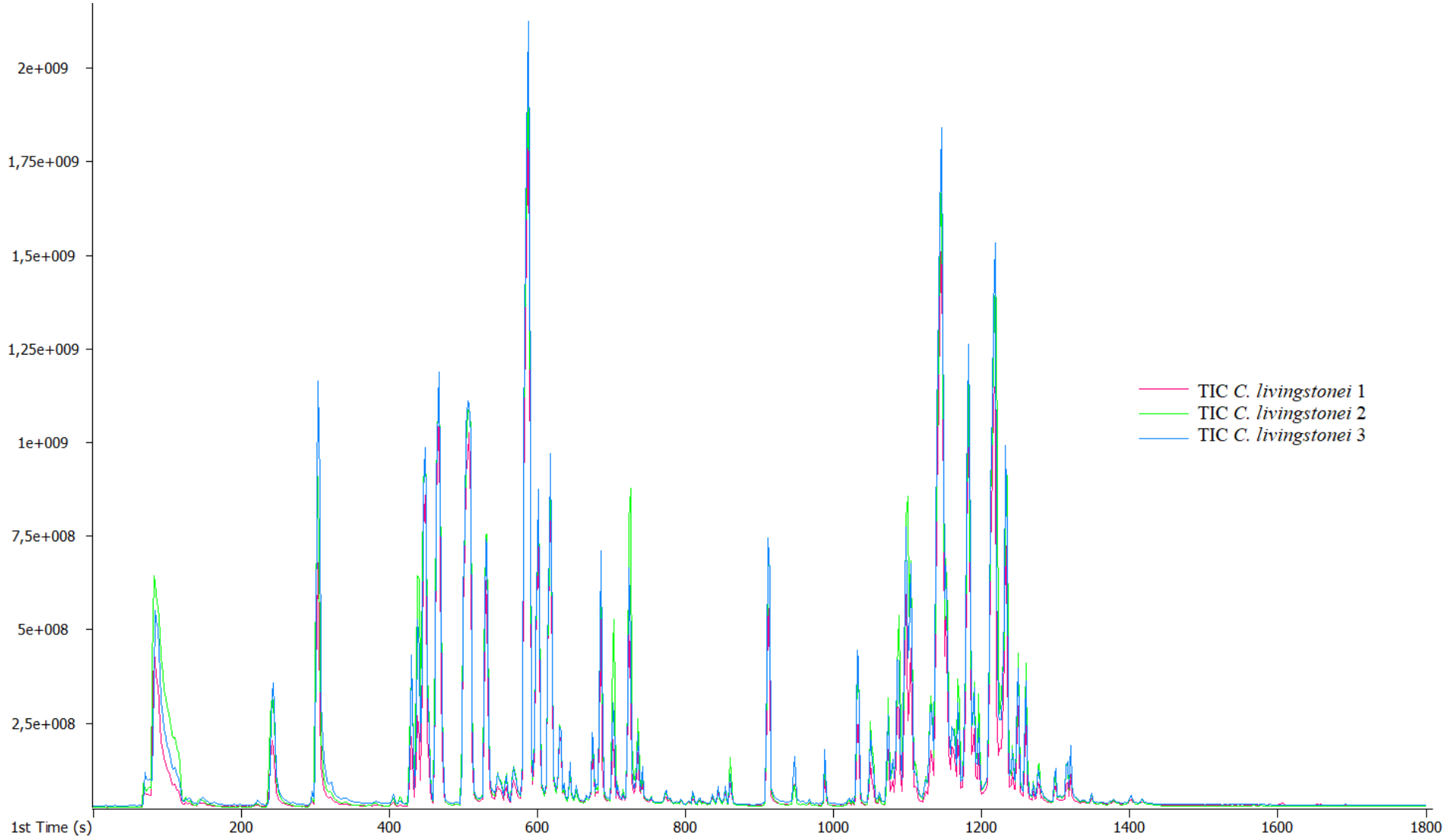


Figure 11: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. livingstonei*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

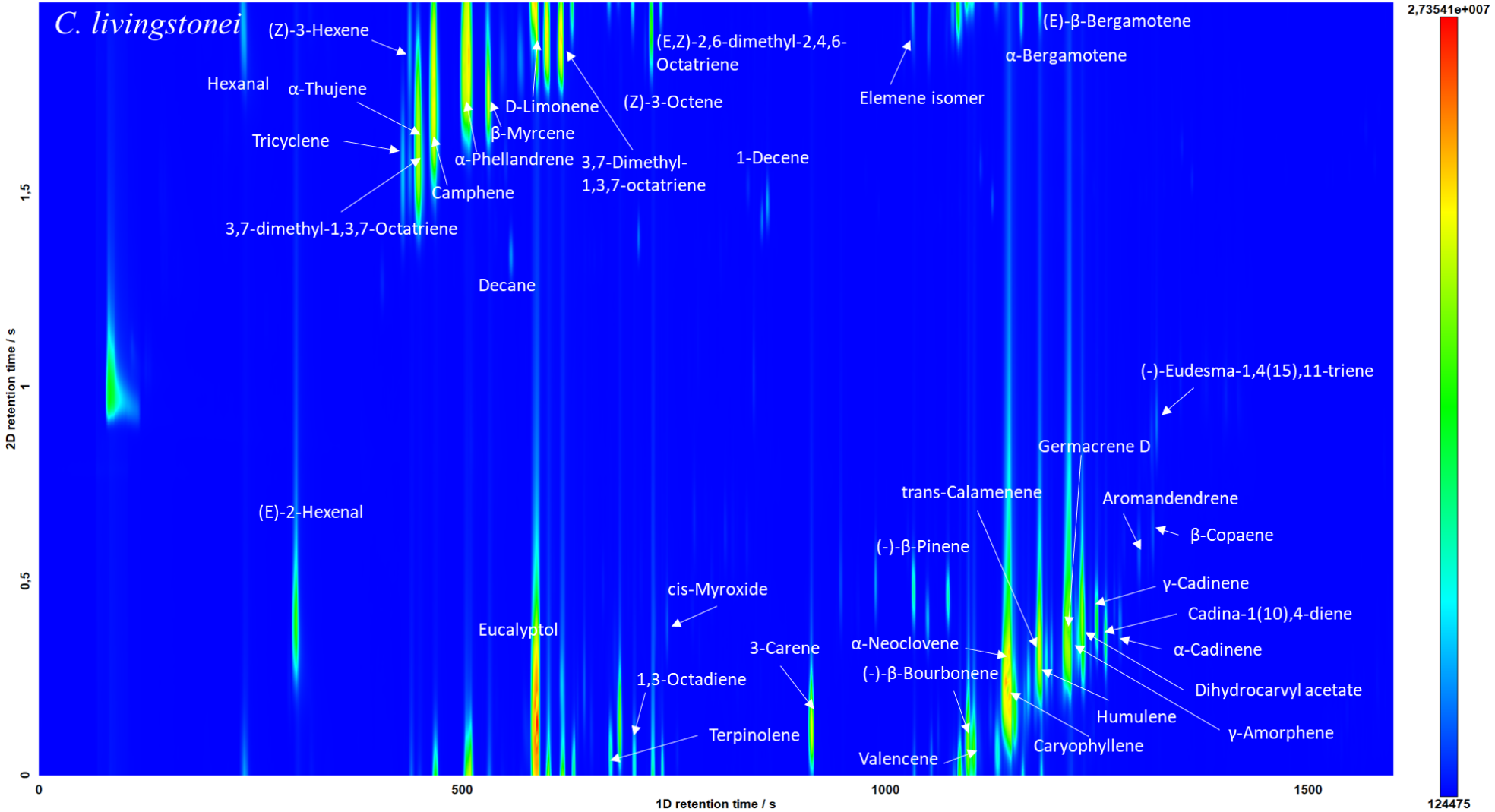


Figure 12: 2D TIC contour plot of replicate extraction 1 of *C. livingstonei*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpe nes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

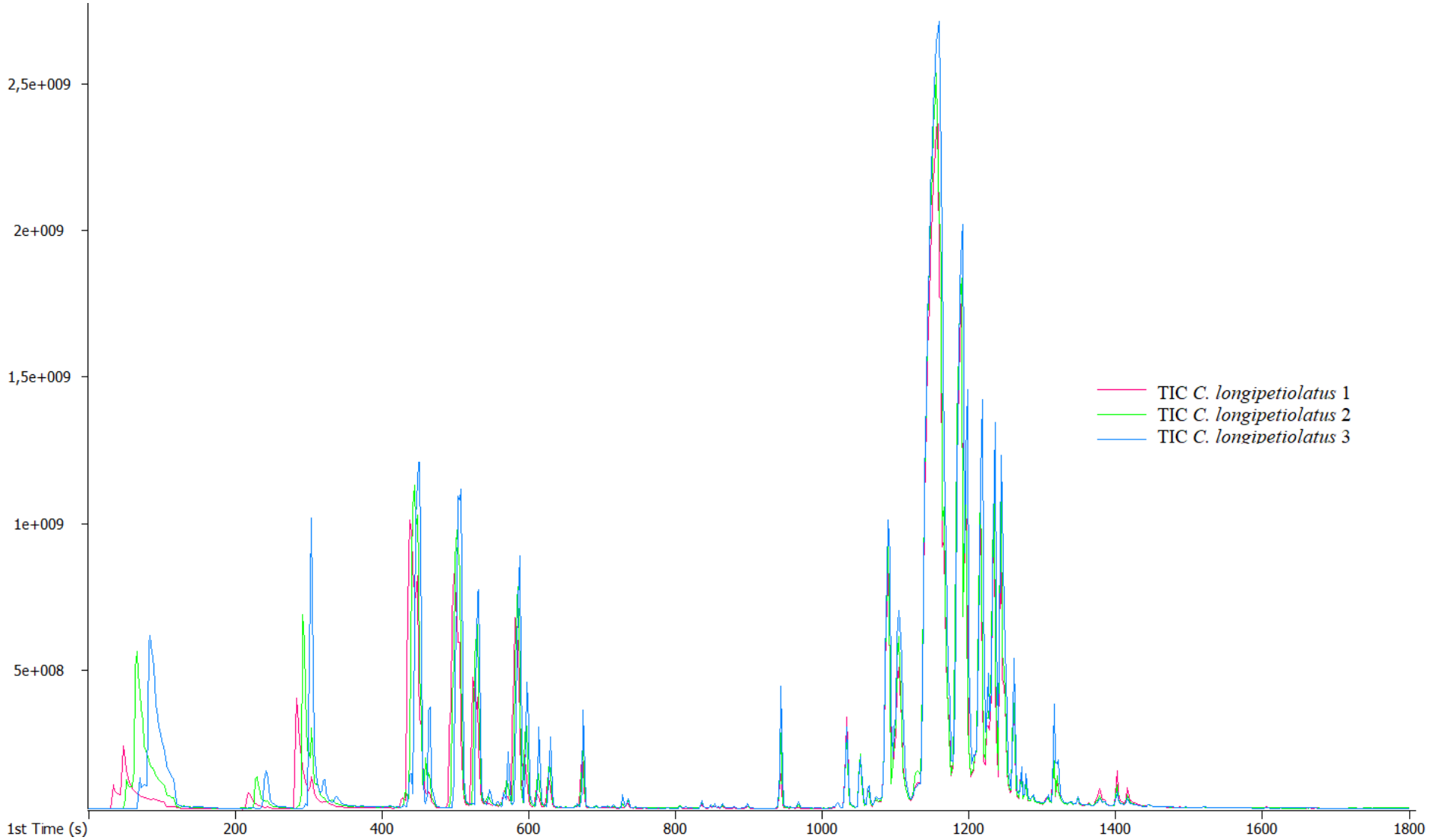


Figure 13: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. longipetiolatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

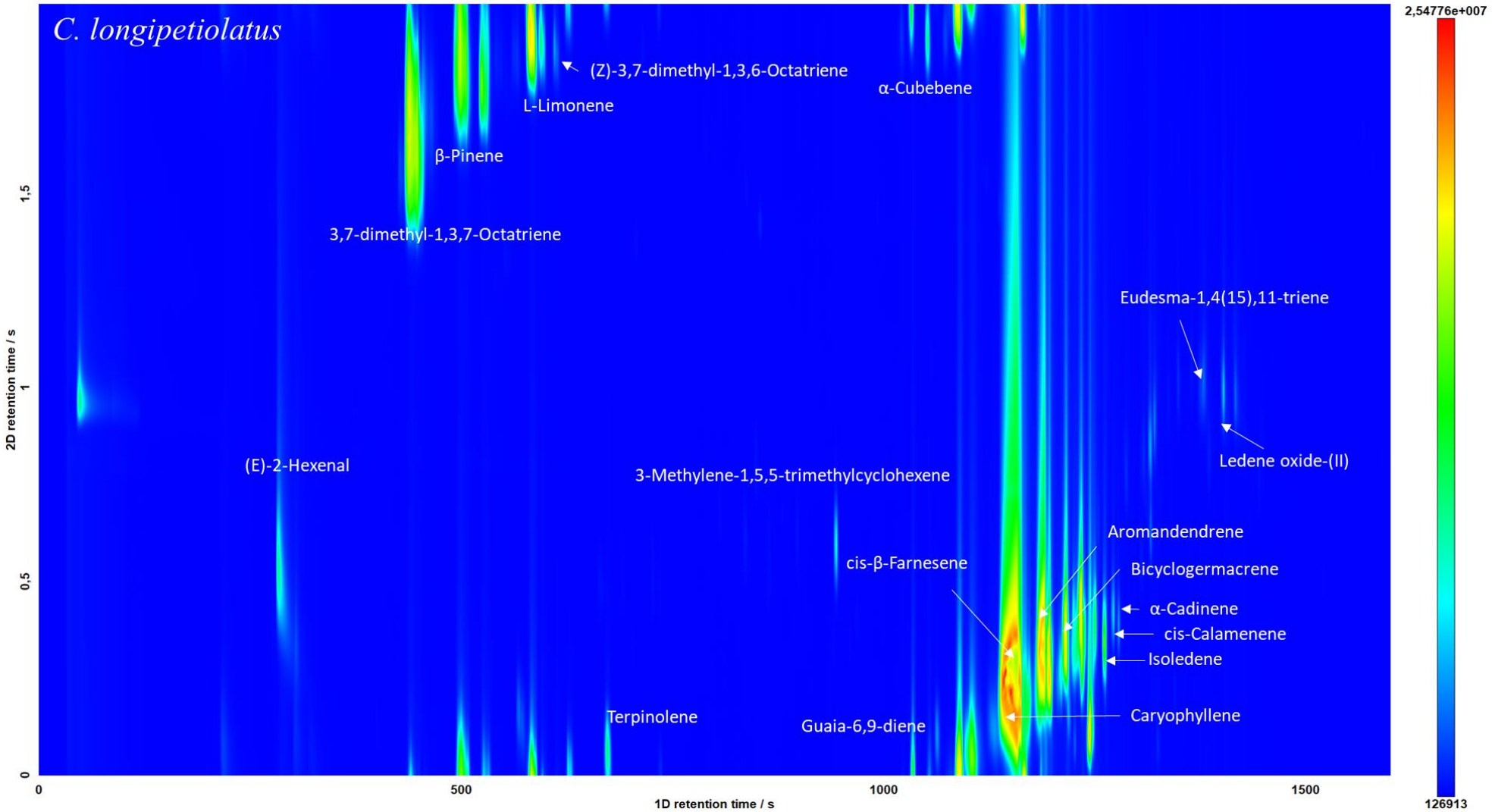


Figure 14: 2D TIC contour plot of replicate extraction 1 of *C. longipetiolatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

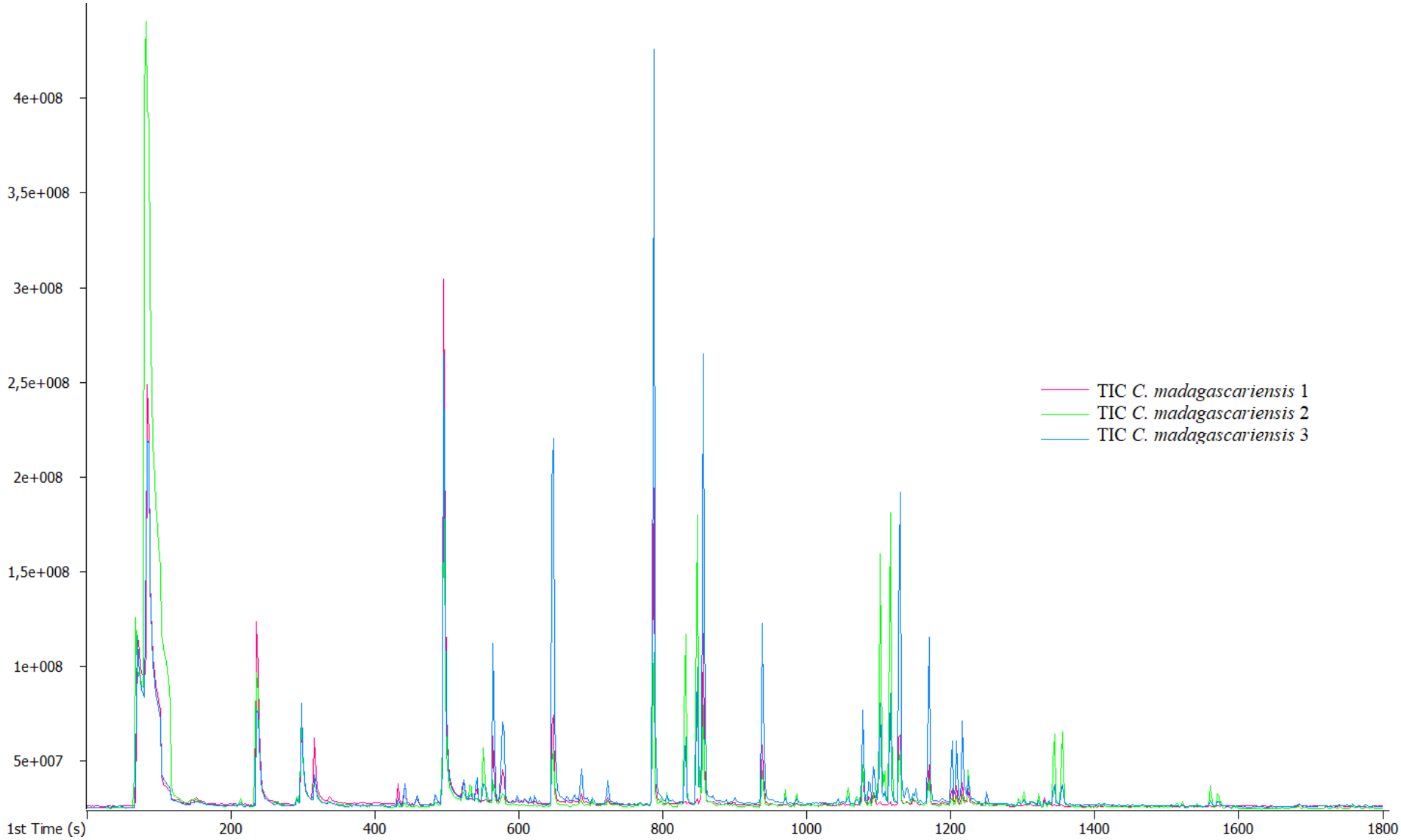


Figure 15: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. madagascariensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

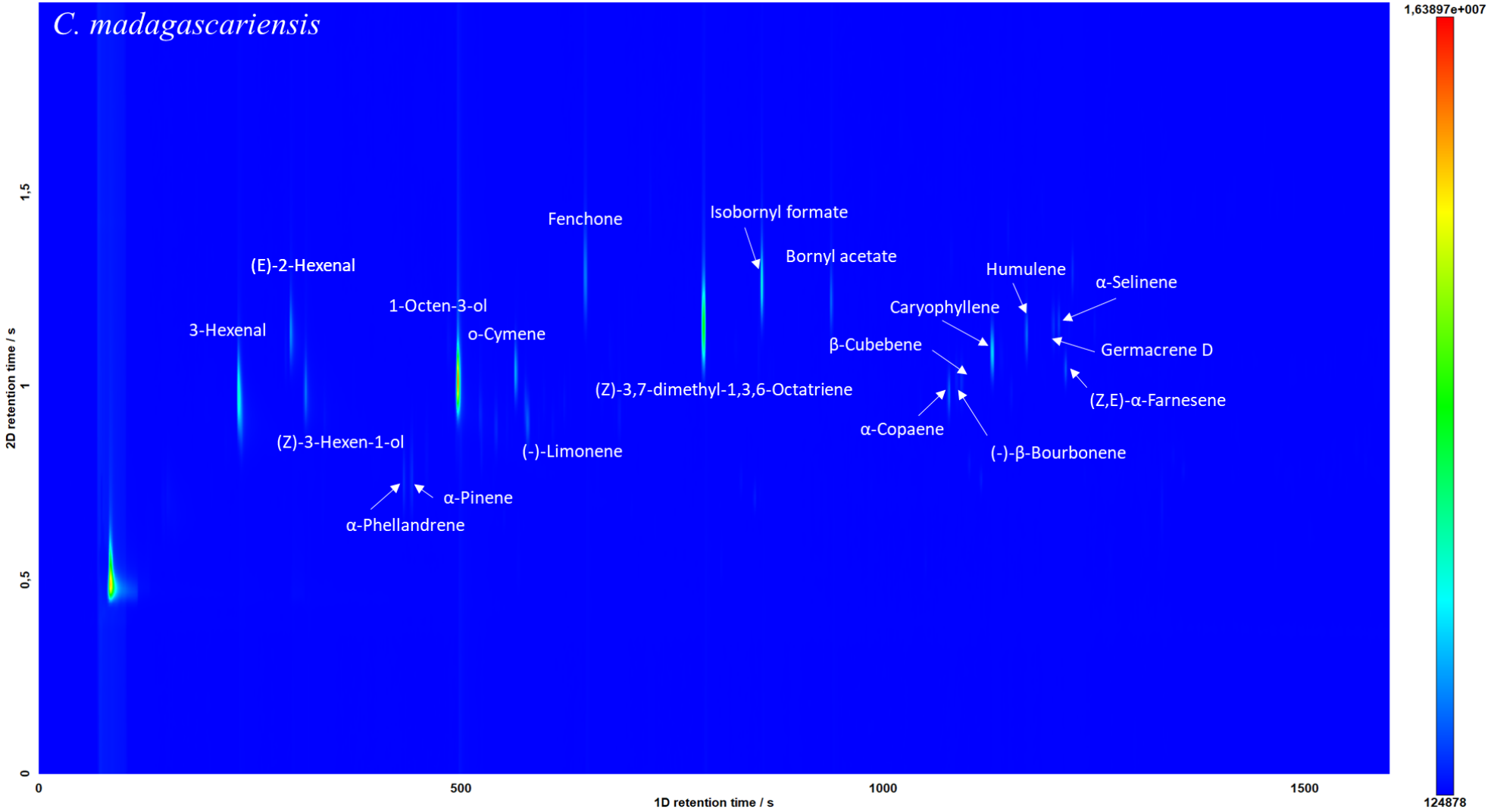


Figure 16: 2D TIC contour plot of replicate extraction 1 of *C. madagascariensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

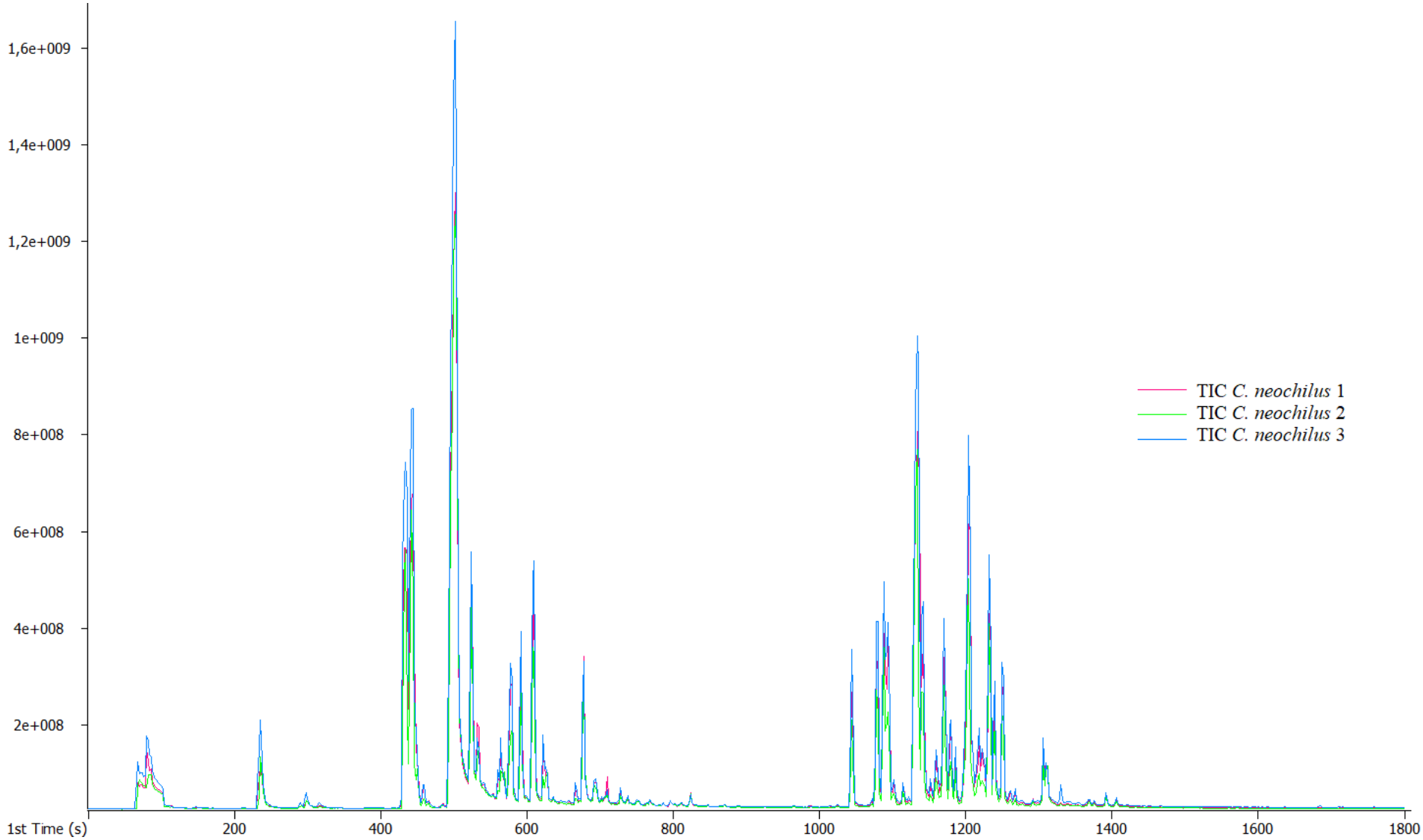


Figure 17: 1D TIC overlay of replicate extractions (n=3) from the leaves of *C. neochilus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

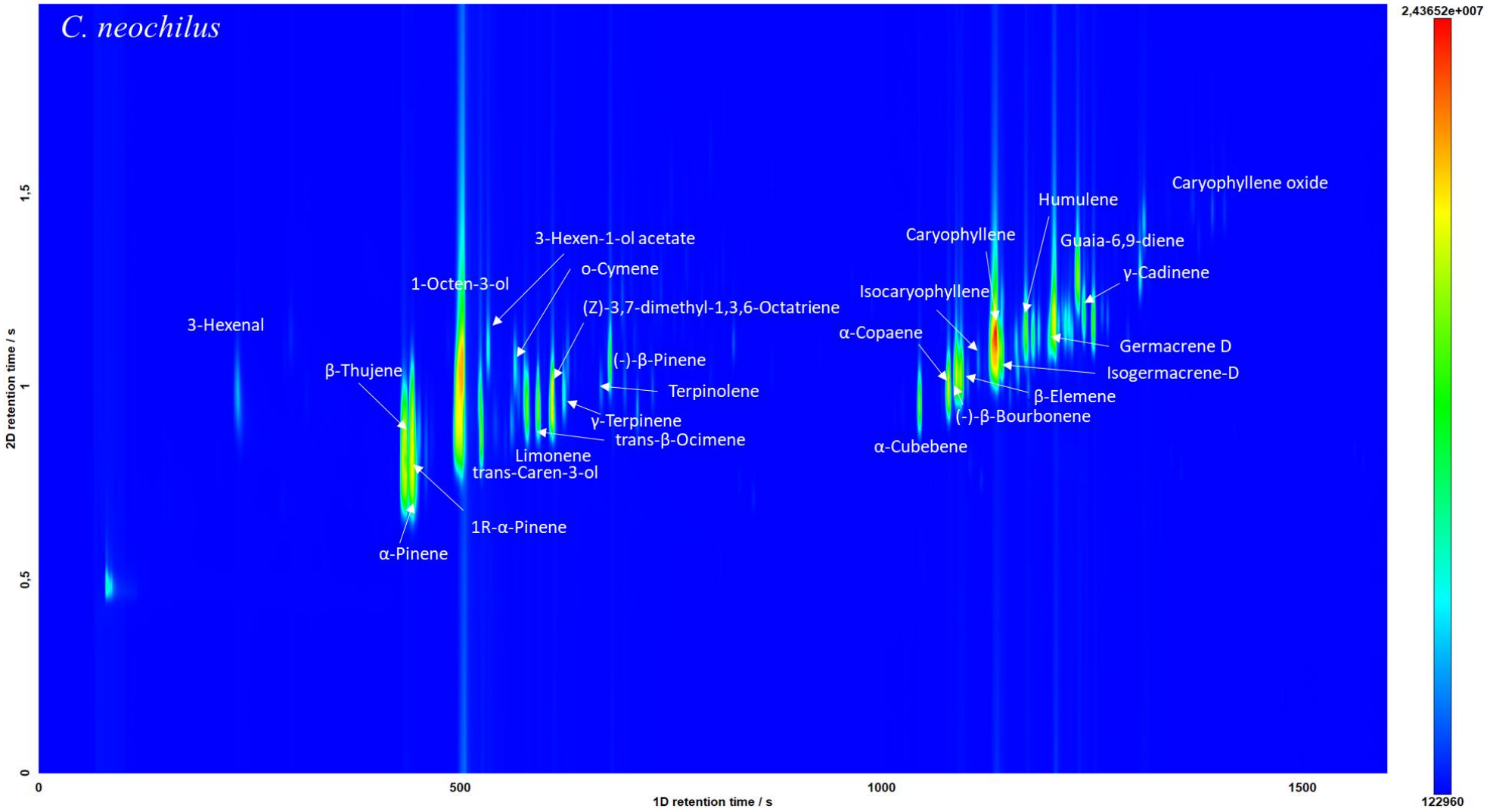


Figure 18: 2D TIC contour plot of replicate extraction 1 of *C. neochilus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

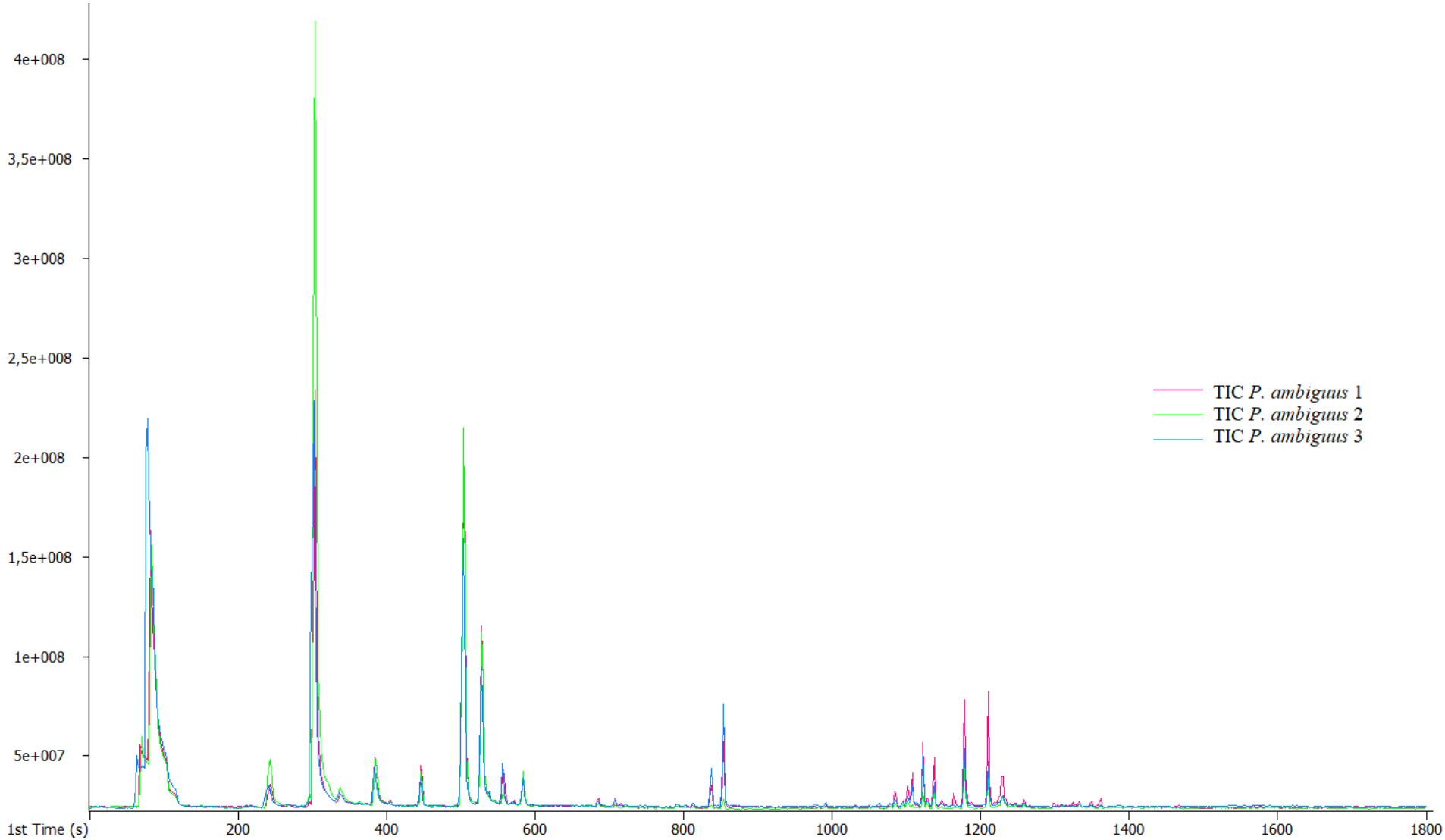


Figure 19: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. ambiguus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

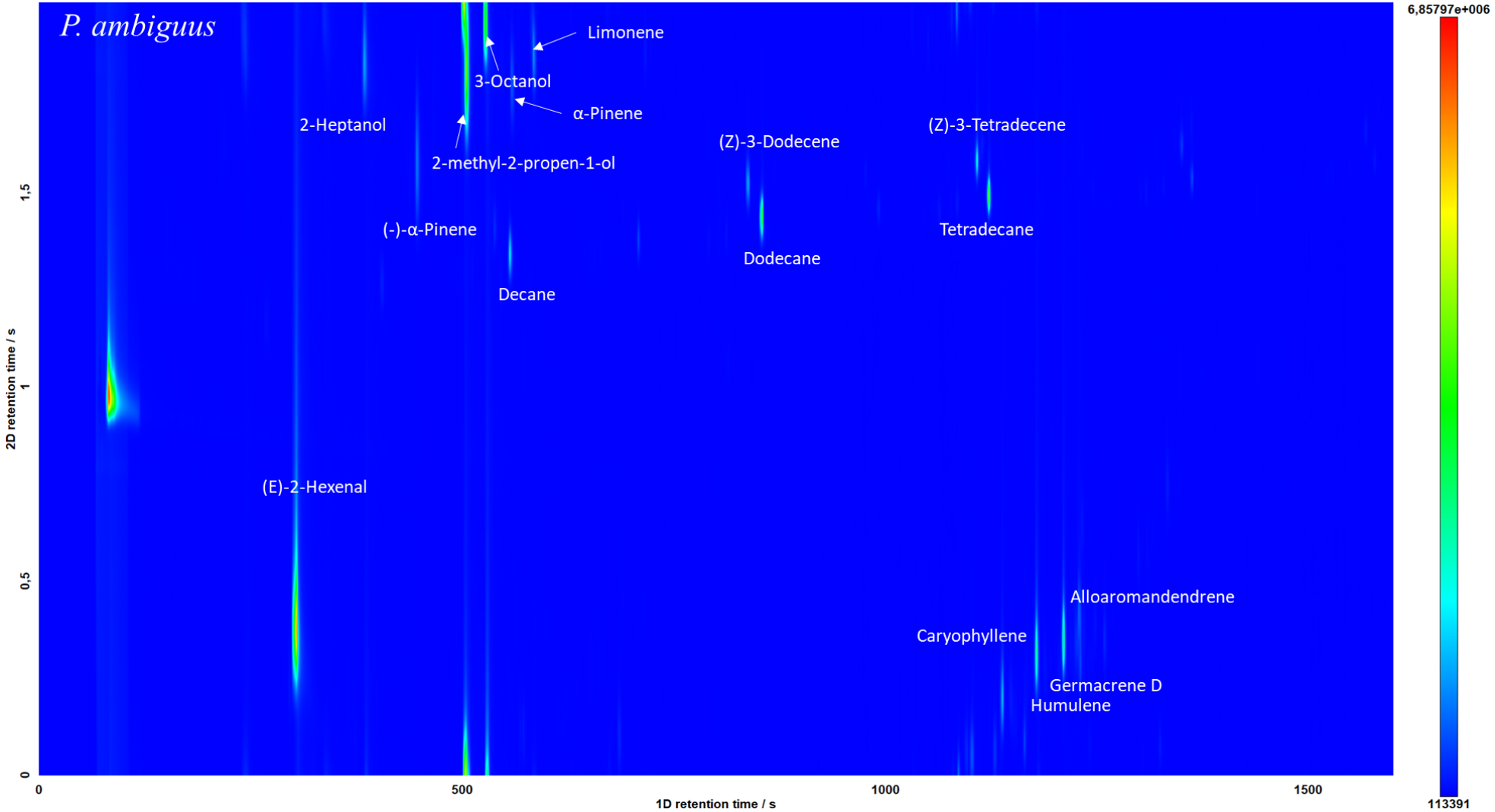


Figure 20: 2D TIC contour plot of replicate extraction 1 of *P. ambiguus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

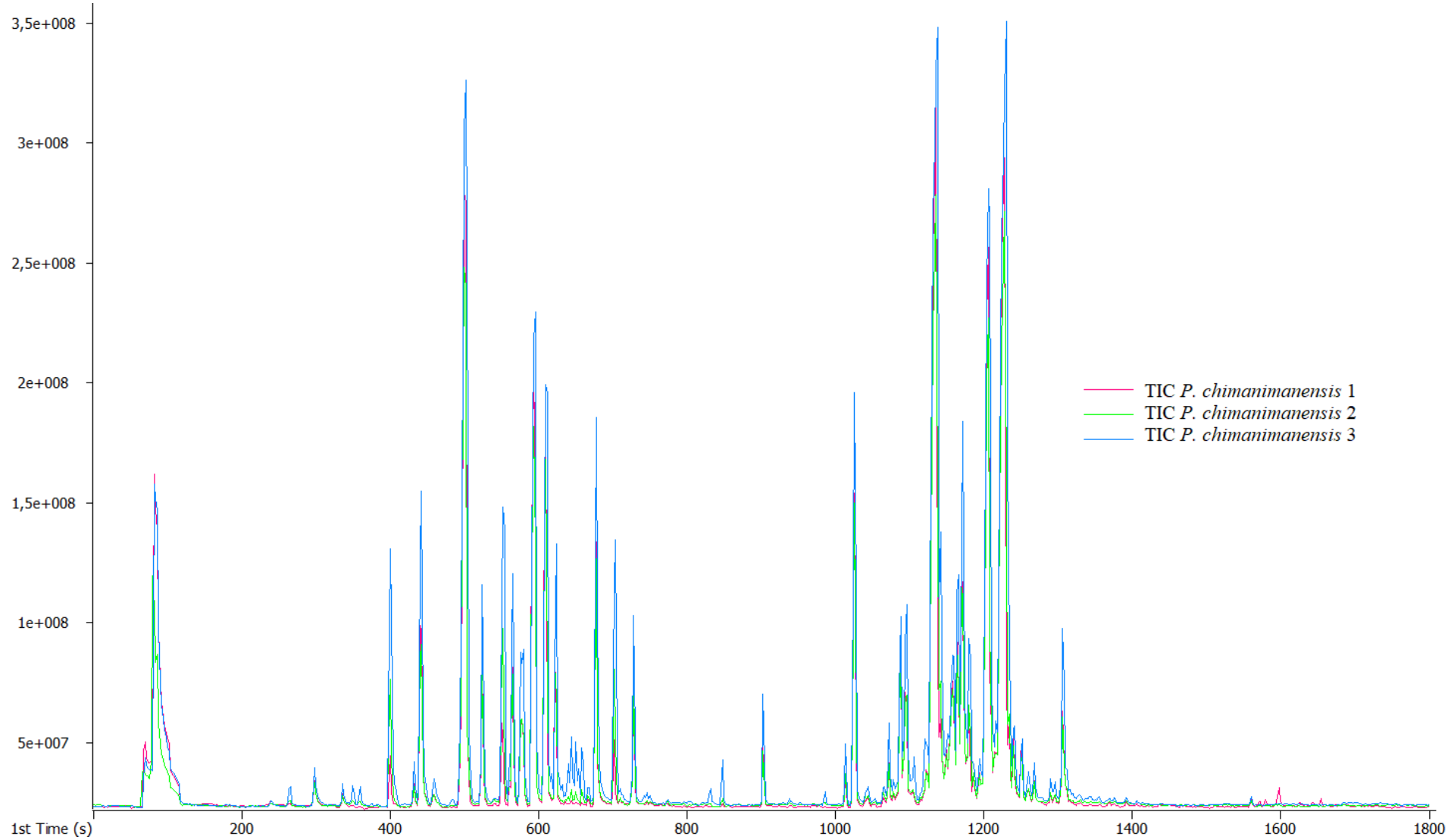


Figure 21: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. chimanimanensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

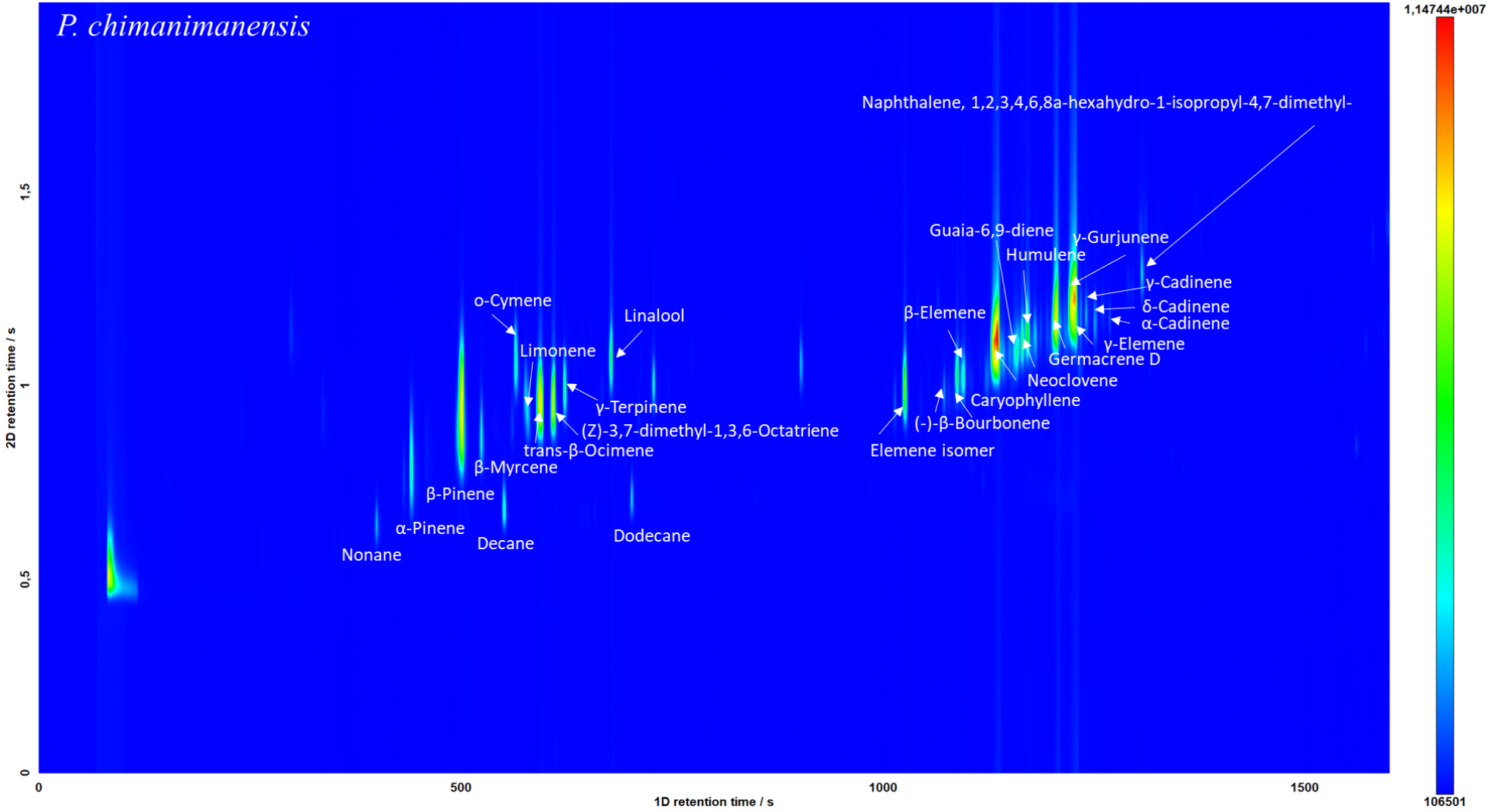


Figure 22: 2D TIC contour plot of replicate extraction 1 of *P. chimanimanensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

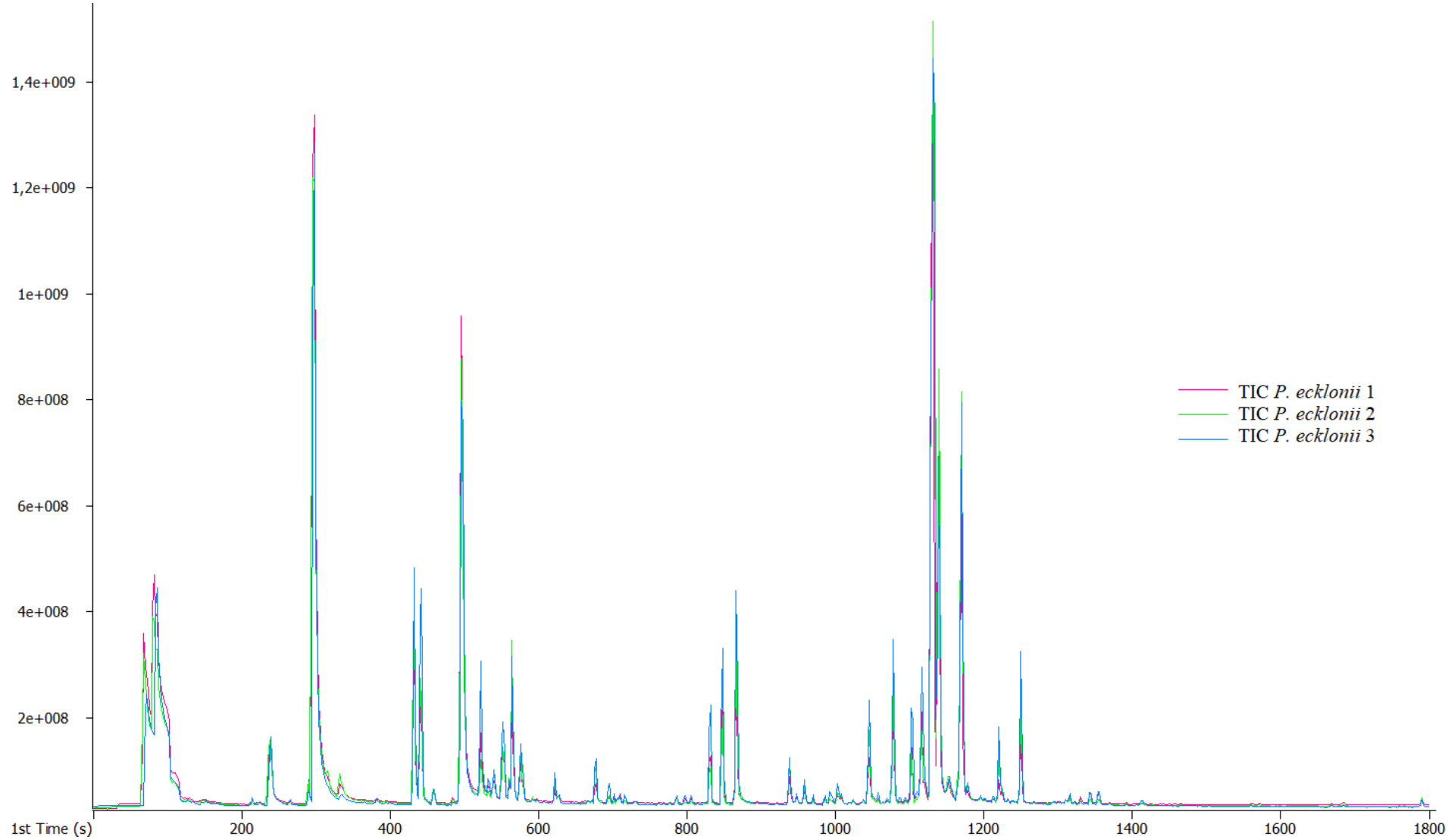


Figure 23: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. ecklonii*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

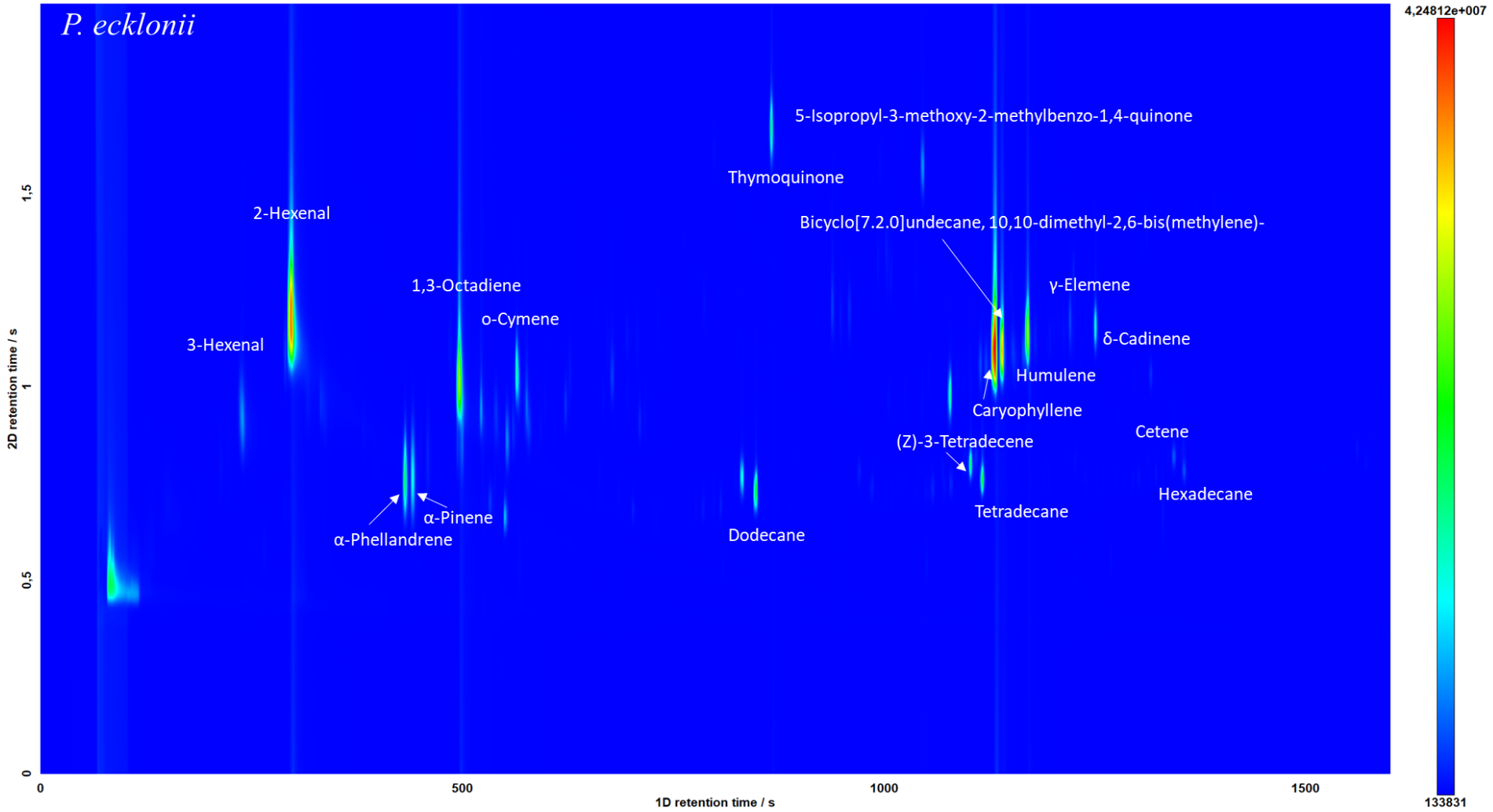


Figure 24: 2D TIC contour plot of replicate extraction 1 of *P. ecklonii*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

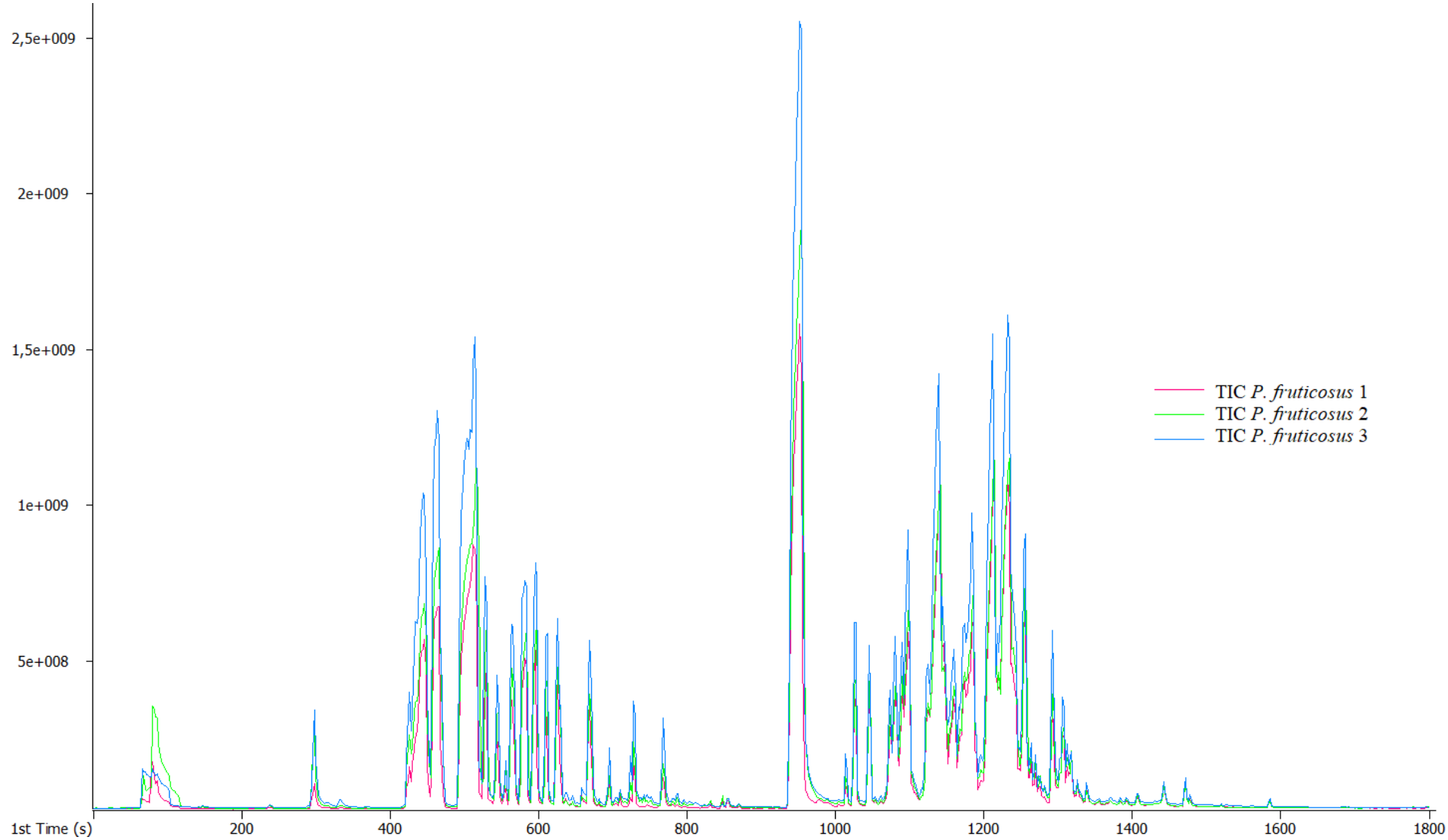


Figure 25: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. fruticosus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

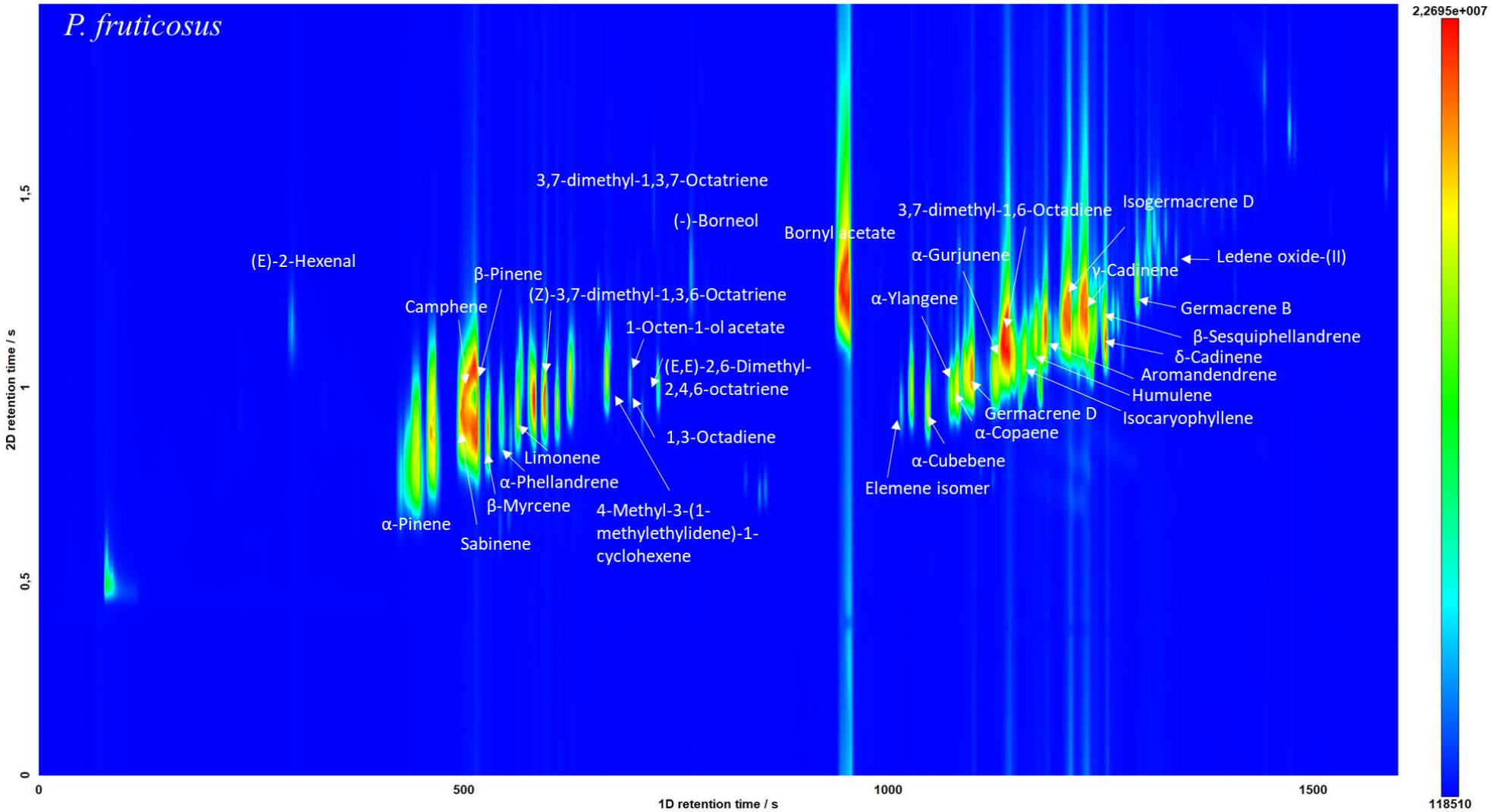


Figure 26: 2D TIC contour plot of replicate extraction 1 of *P. fruticosus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

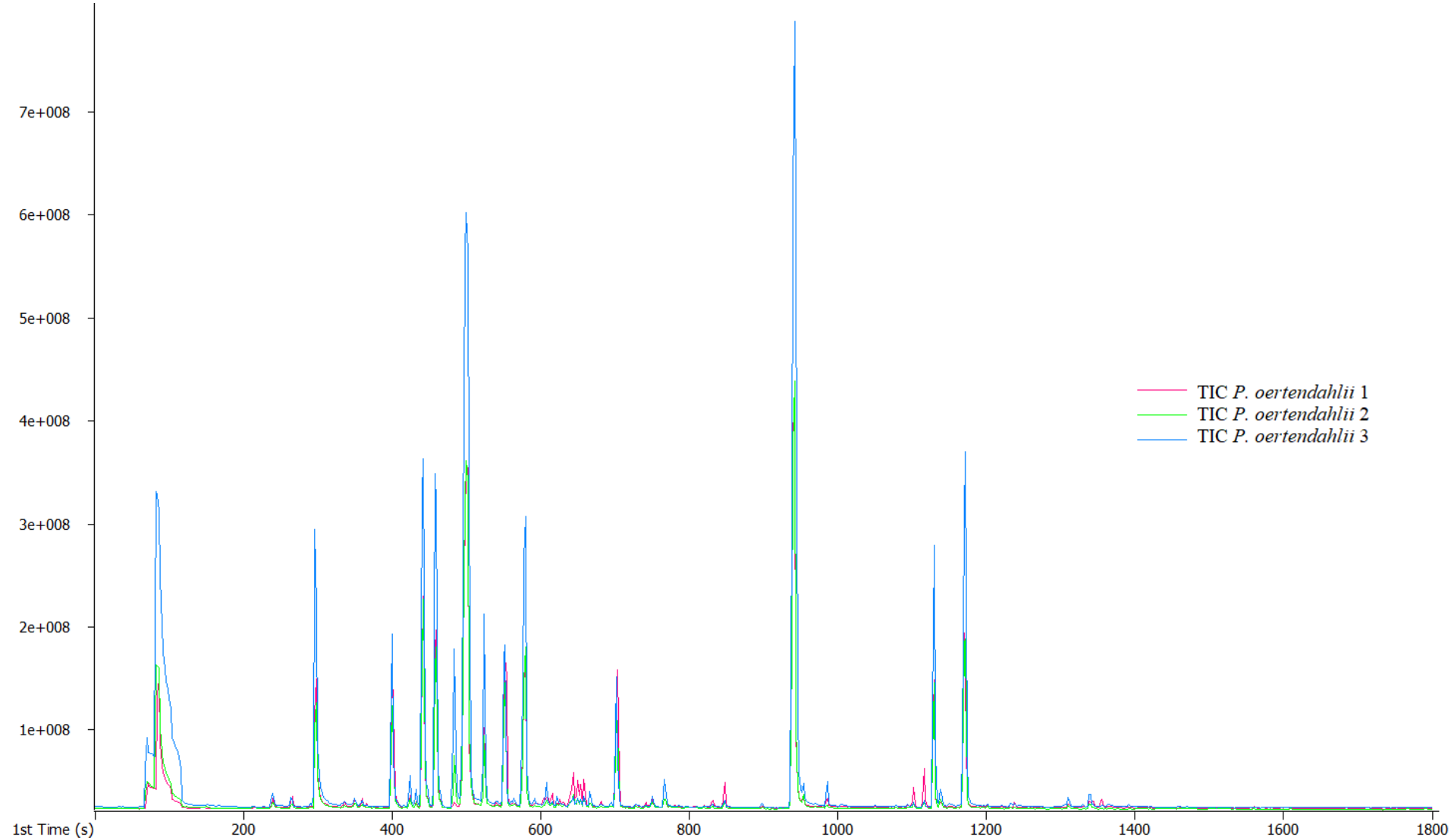


Figure 27: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. oertendahlii*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

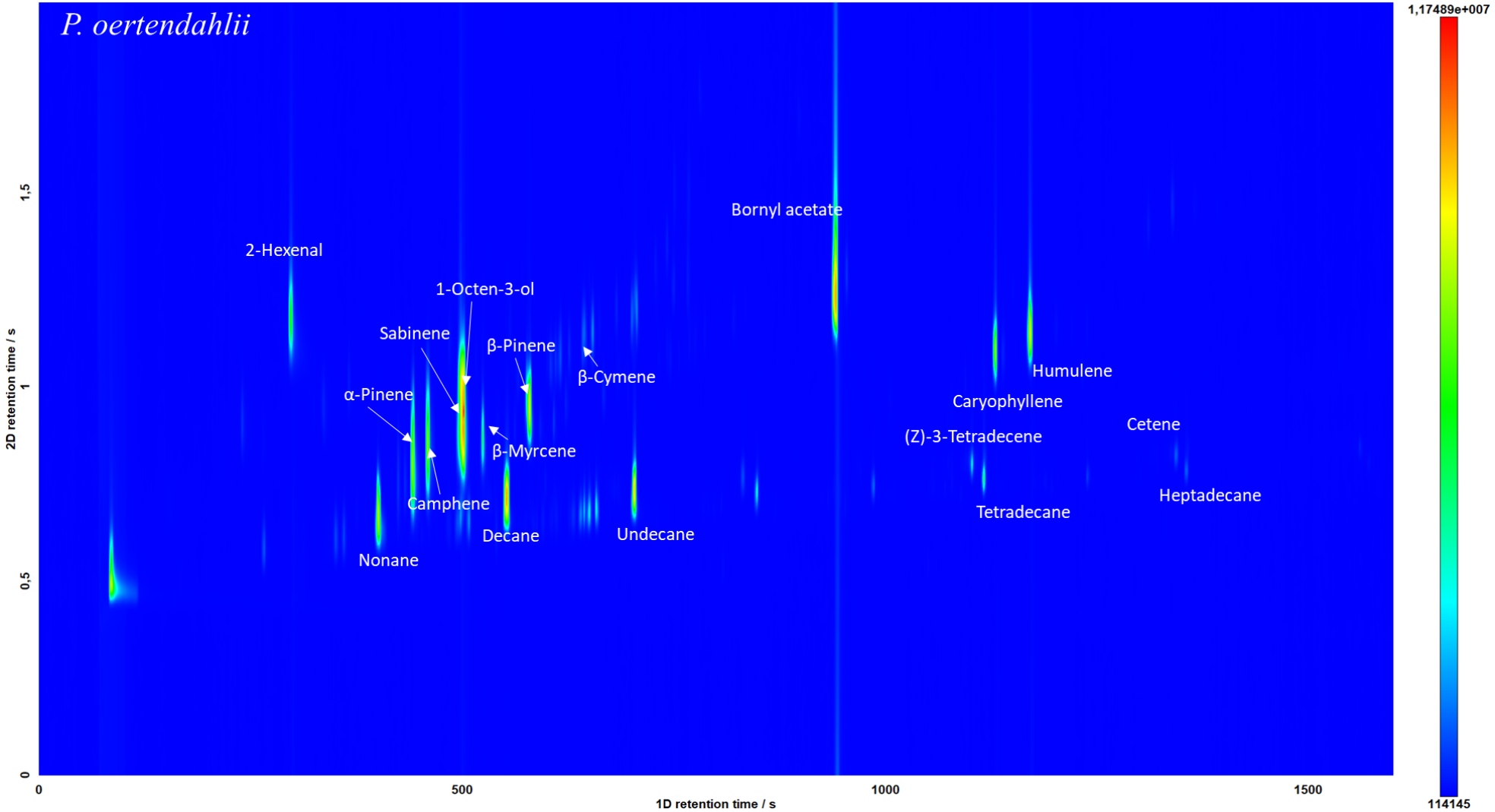


Figure 28: 2D TIC contour plot of replicate extraction 1 of *P. oertendahlii*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: decane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

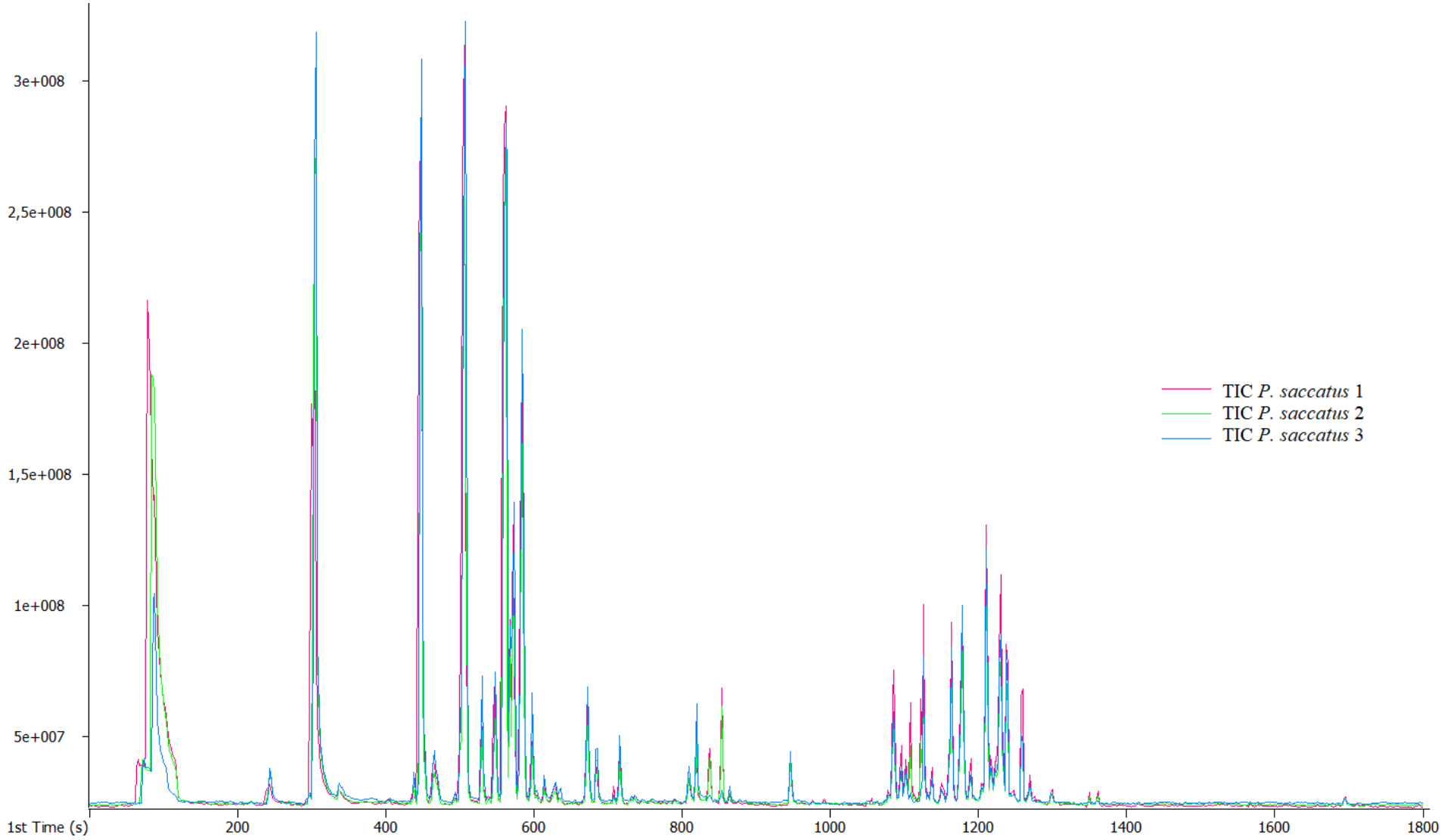


Figure 29: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. saccatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

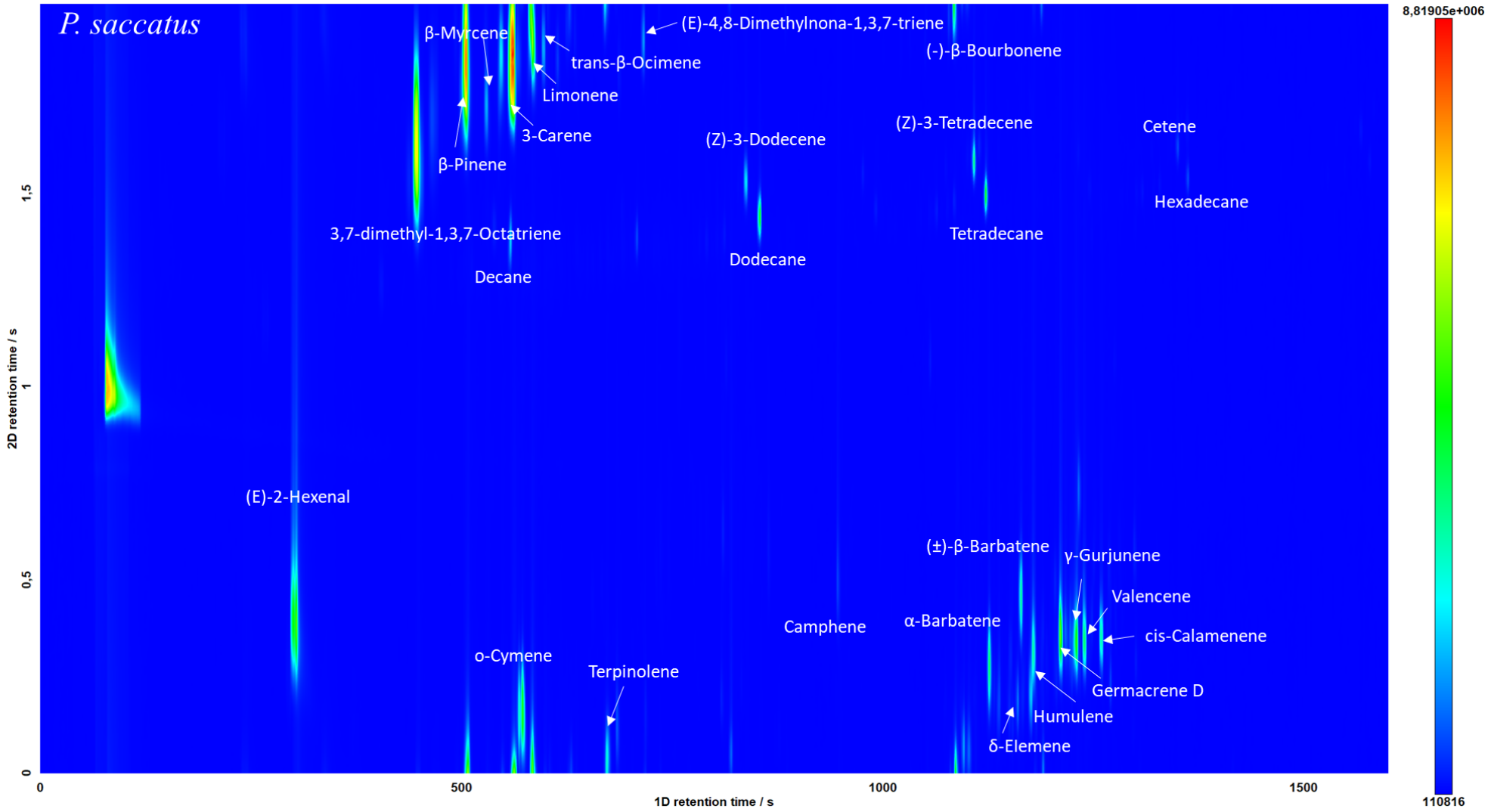


Figure 30: 2D TIC contour plot of replicate extraction 1 of *P. saccatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

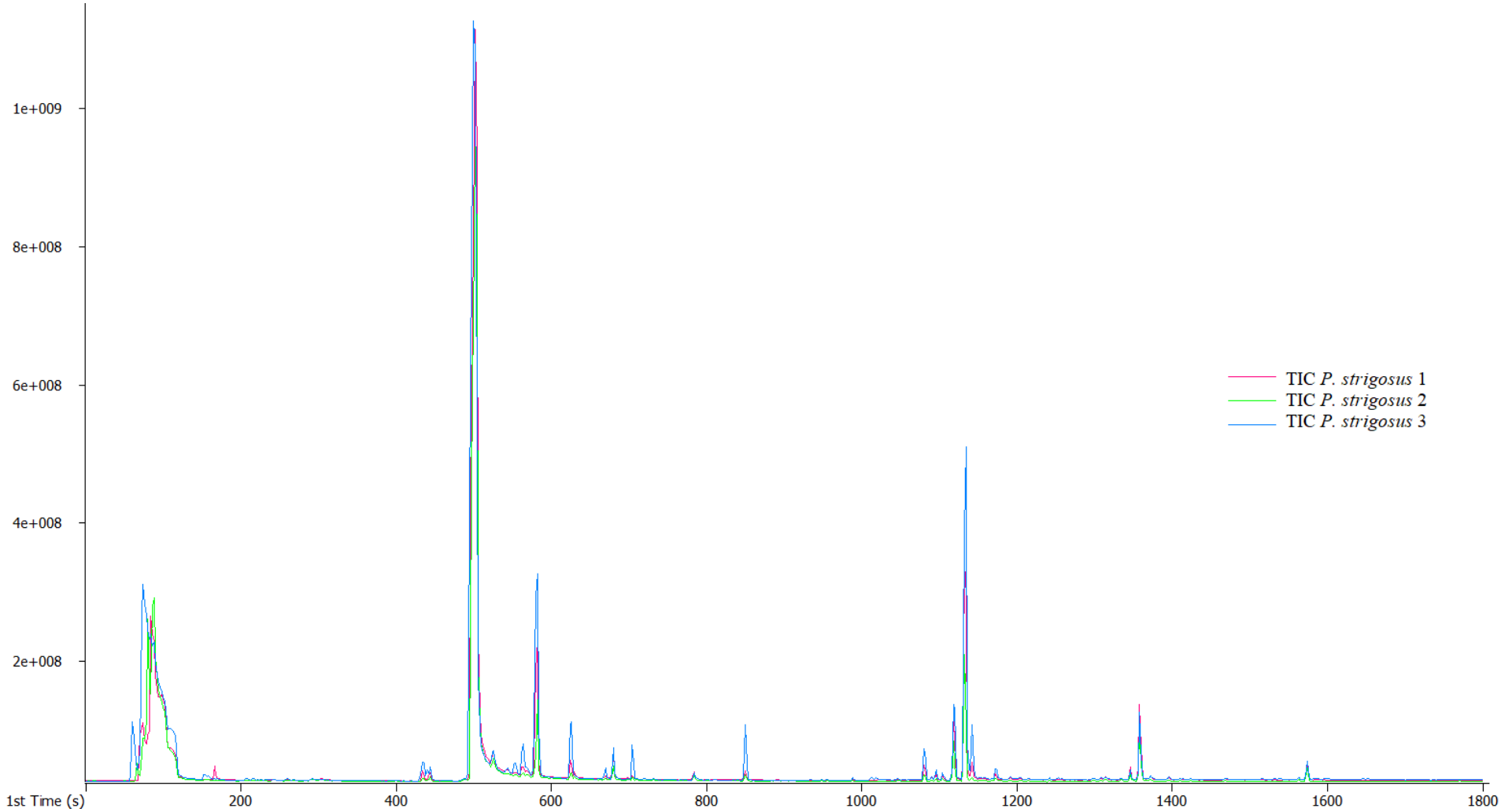


Figure 31: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. strigosus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

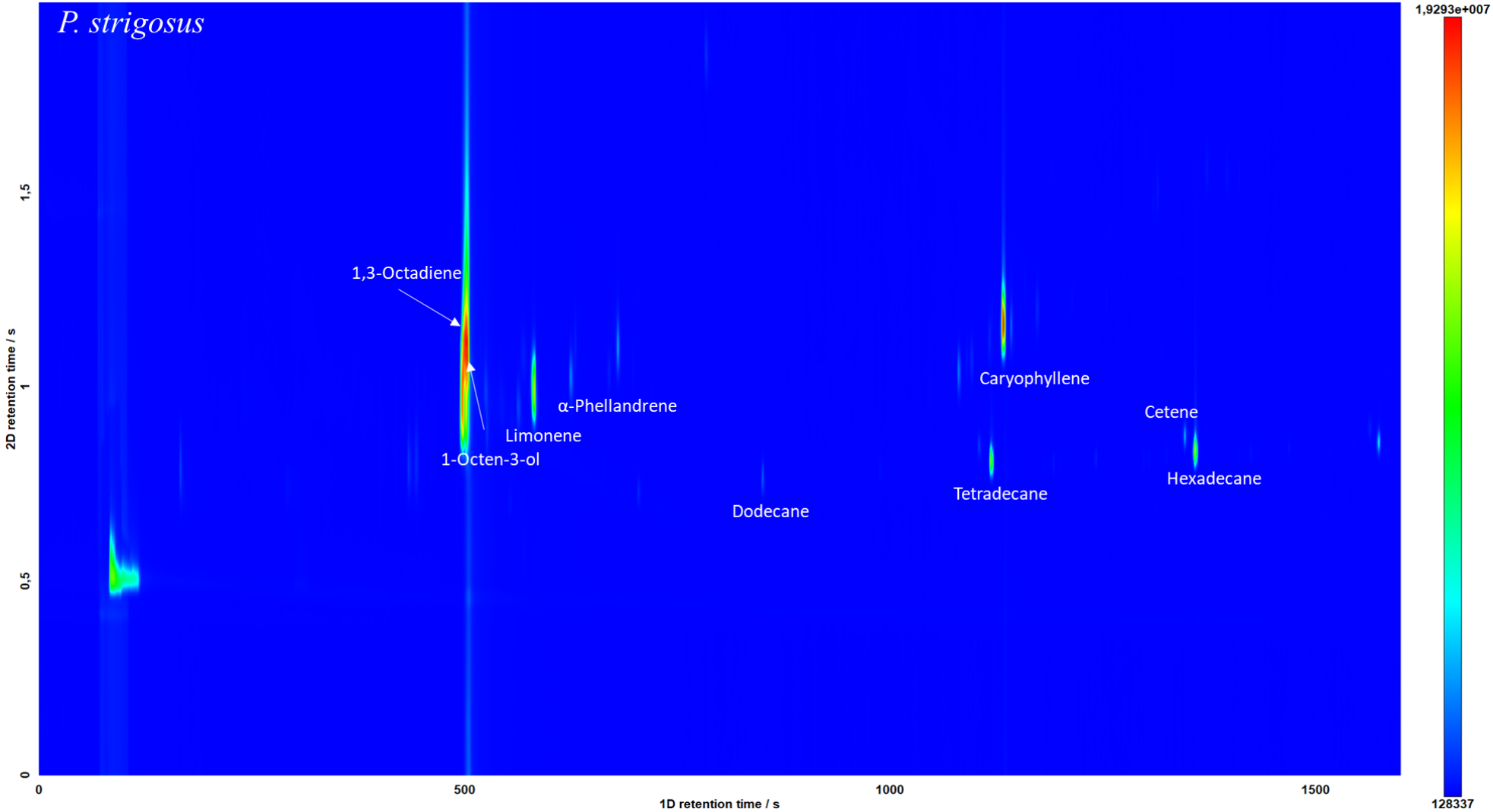


Figure 32: 2D TIC contour plot of replicate extraction 1 of *P. strigosus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

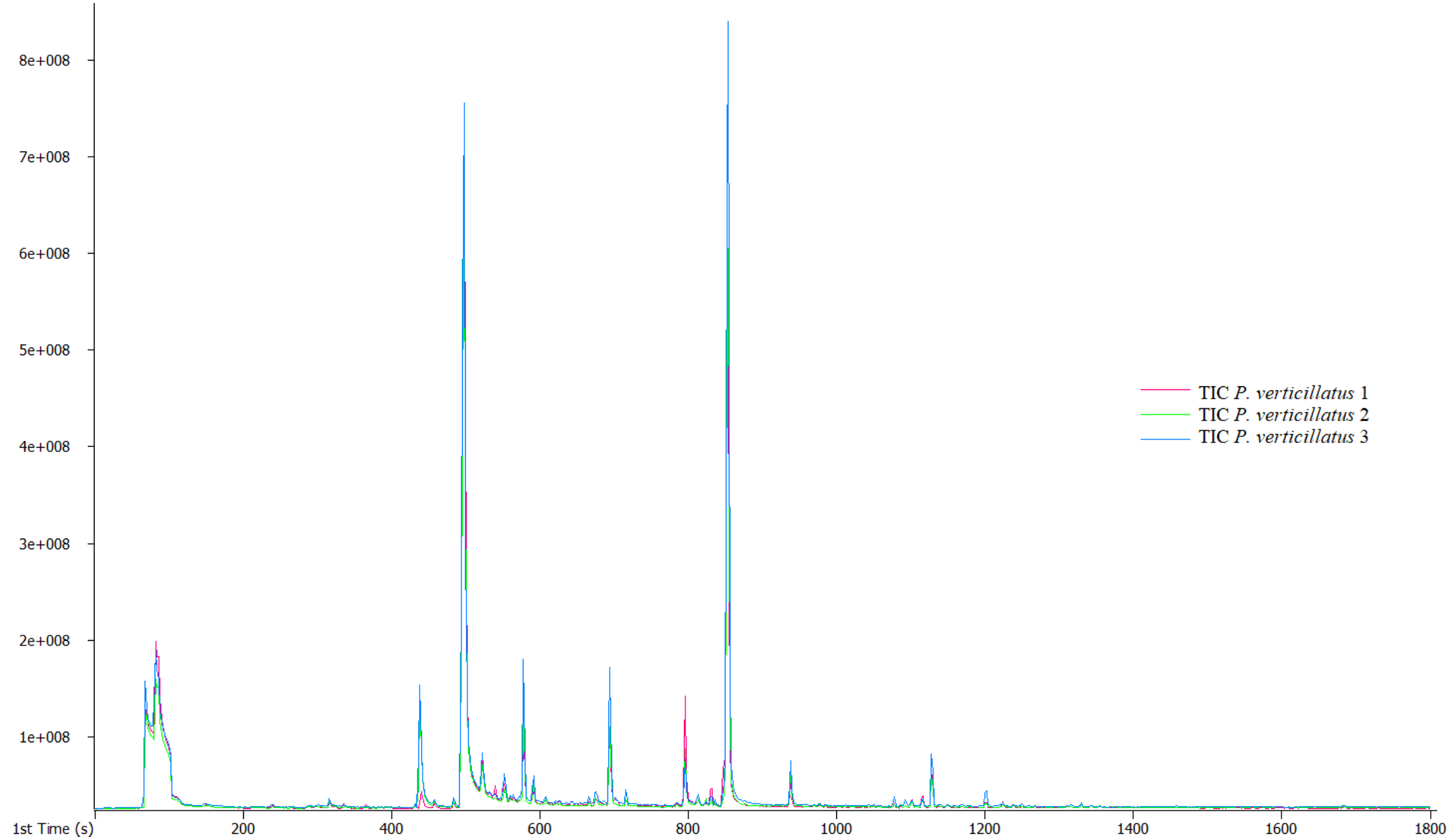


Figure 33: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. verticillatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

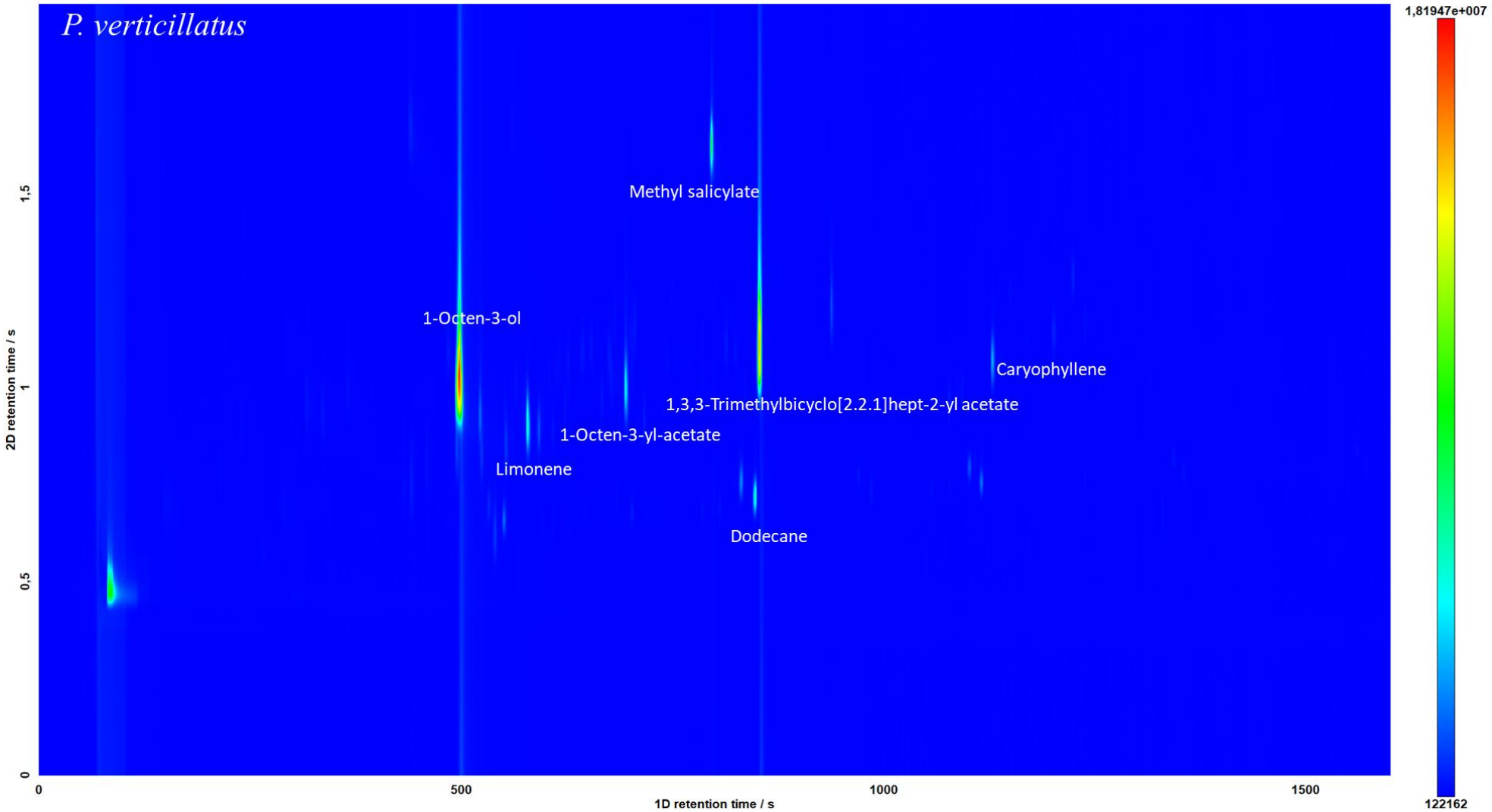


Figure 34: 2D TIC contour plot of replicate extraction 1 of *P. verticillatus*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes). The presence of n-alkanes (e.g.: dodecane) may be due to contamination or persistence from n-alkane (C_6-C_{28}) samples from previous runs.

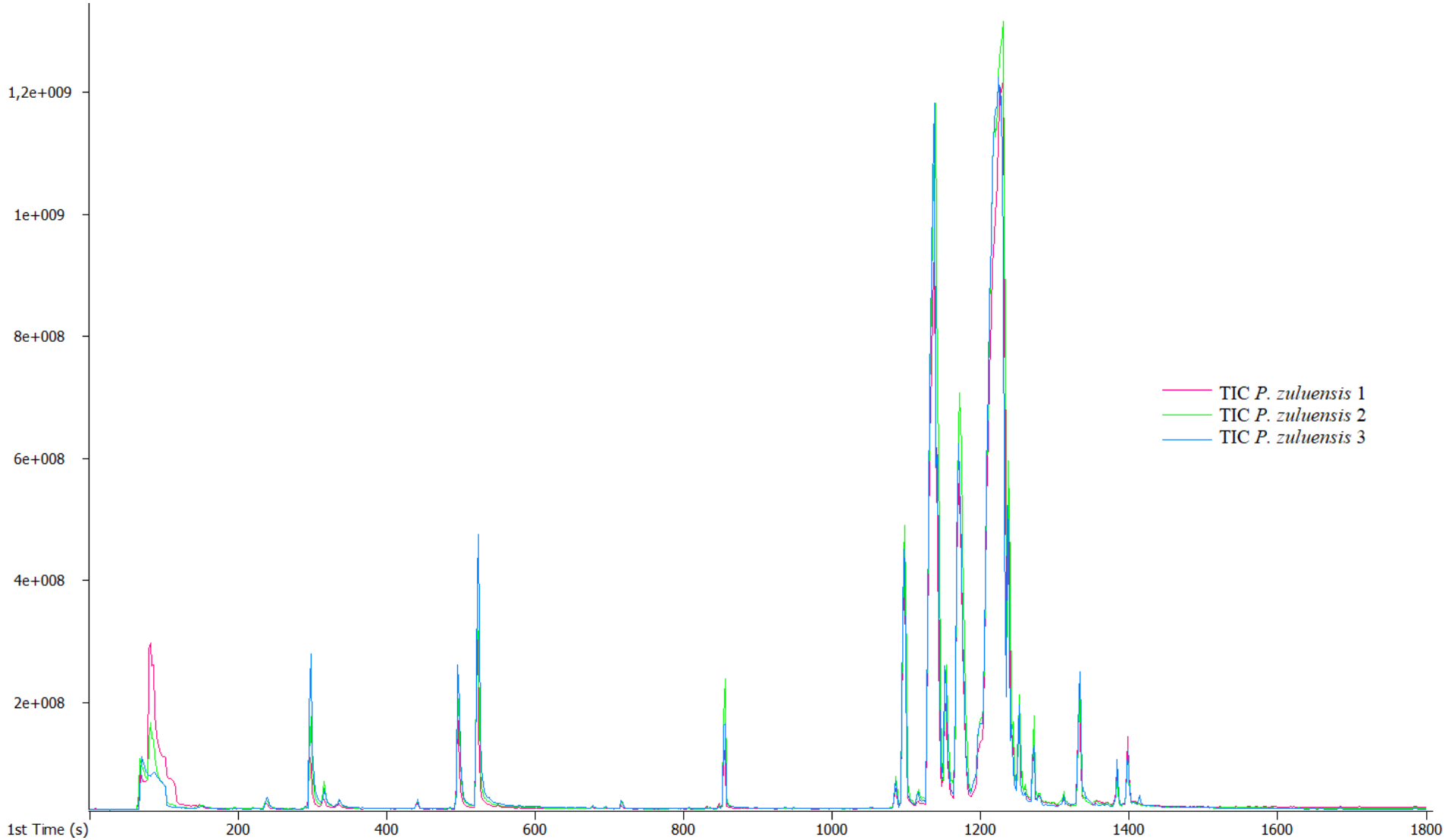


Figure 35: 1D TIC overlay of replicate extractions (n=3) from the leaves of *P. zuluensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

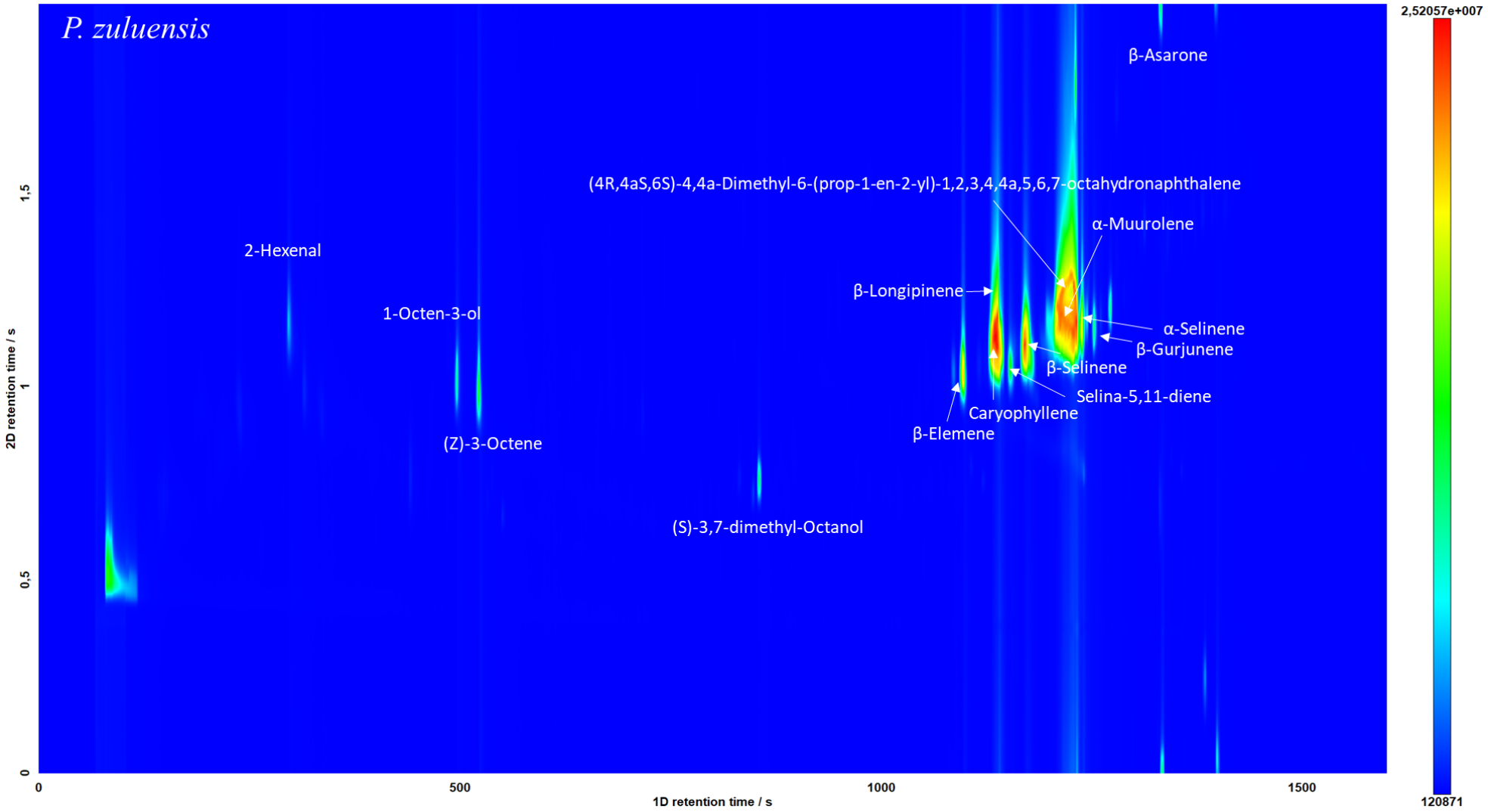


Figure 36: 2D TIC contour plot of replicate extraction 1 of *P. zuluensis*. Peaks of interest fall mostly within the retention time regions of 400-700 s (monoterpenes) and 1040-1300 s (sesquiterpenes).

Figure 37 and Figure 38 are overlaid 1D TIC of single replicates of each species according to genus. Comparison of these suggest comparable peak structure in both terpene regions, however, the complexity of the TIC for *P. fruticosus*, which does not appear to be characteristic of the other *Plectranthus* species, contributes significantly to the peaks present in Figure 38. If *P. fruticosus* is removed from the overlay, as is the case in Figure 39, it is apparent that the *Plectranthus* samples have a sparser variety of peaks, in both the monoterpene and the sesquiterpene regions, compared to those of genus *Coleus* (with the exception of *C. madagascariensis*). As will be seen in subsequent sections, this observation is supported, in the case of the sesquiterpenes, by the results of the preliminary statistical analysis (Chapter 5.3) and the variable selection results of the analysis by machine learning (Chapter 5.7).

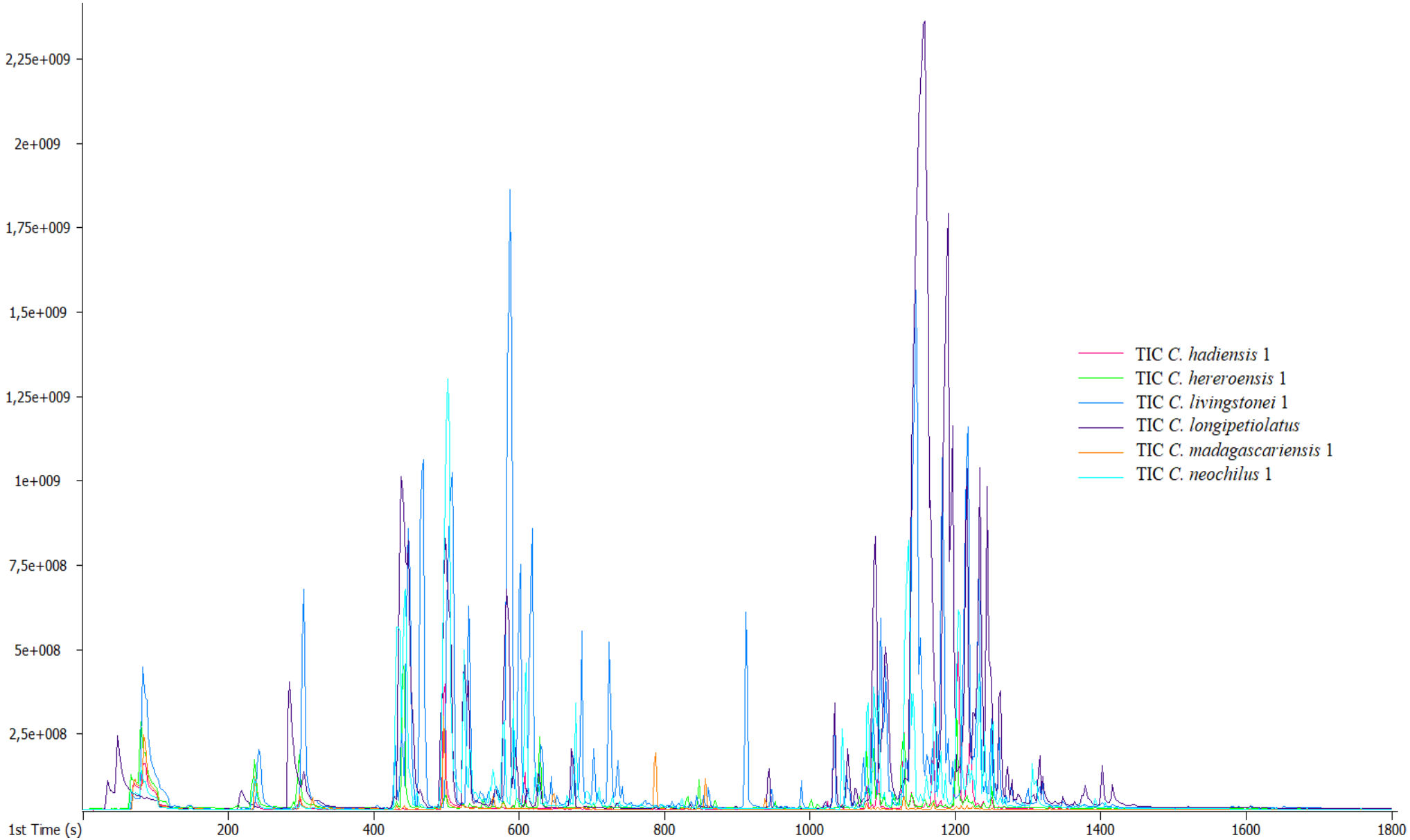


Figure 37: Overlay of the 1D TICs of single replicates of species of genus *Coleus*: *C. hadiensis*, *C. hereroensis*, *C. livingstonei*, *C. longipetiolatus*, *C. madagascariensis*, *C. neochilus*.

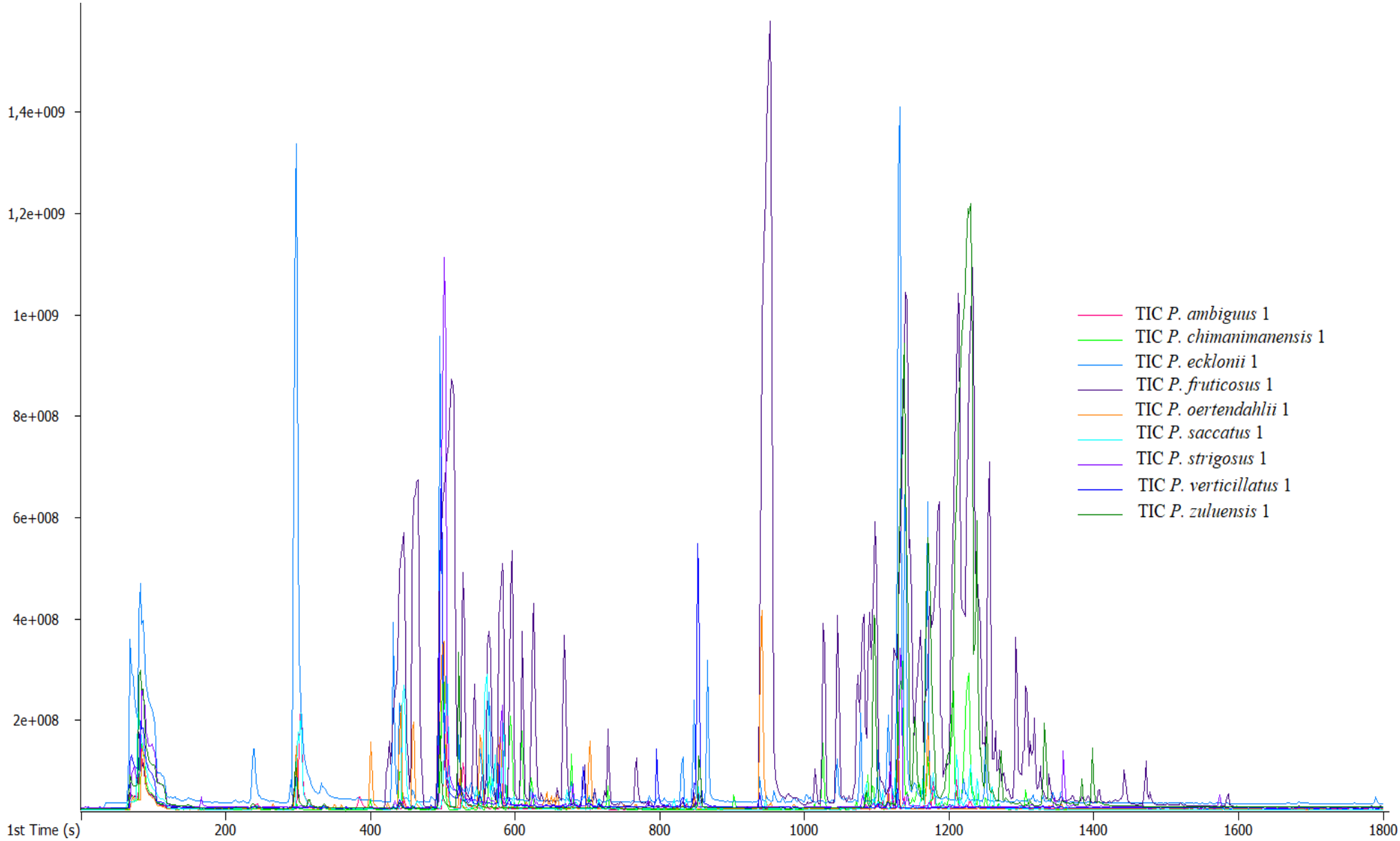


Figure 38: Overlay of the 1D TICs of single replicates of species of genus *Plectranthus*: *P. ambiguus*, *P. chimanimanensis*, *P. ecklonii*, *P. fruticosus*, *P. oertendahlii*, *P. saccatus*, *P. strigosus*, *P. verticillatus*, *P. zuluensis*.

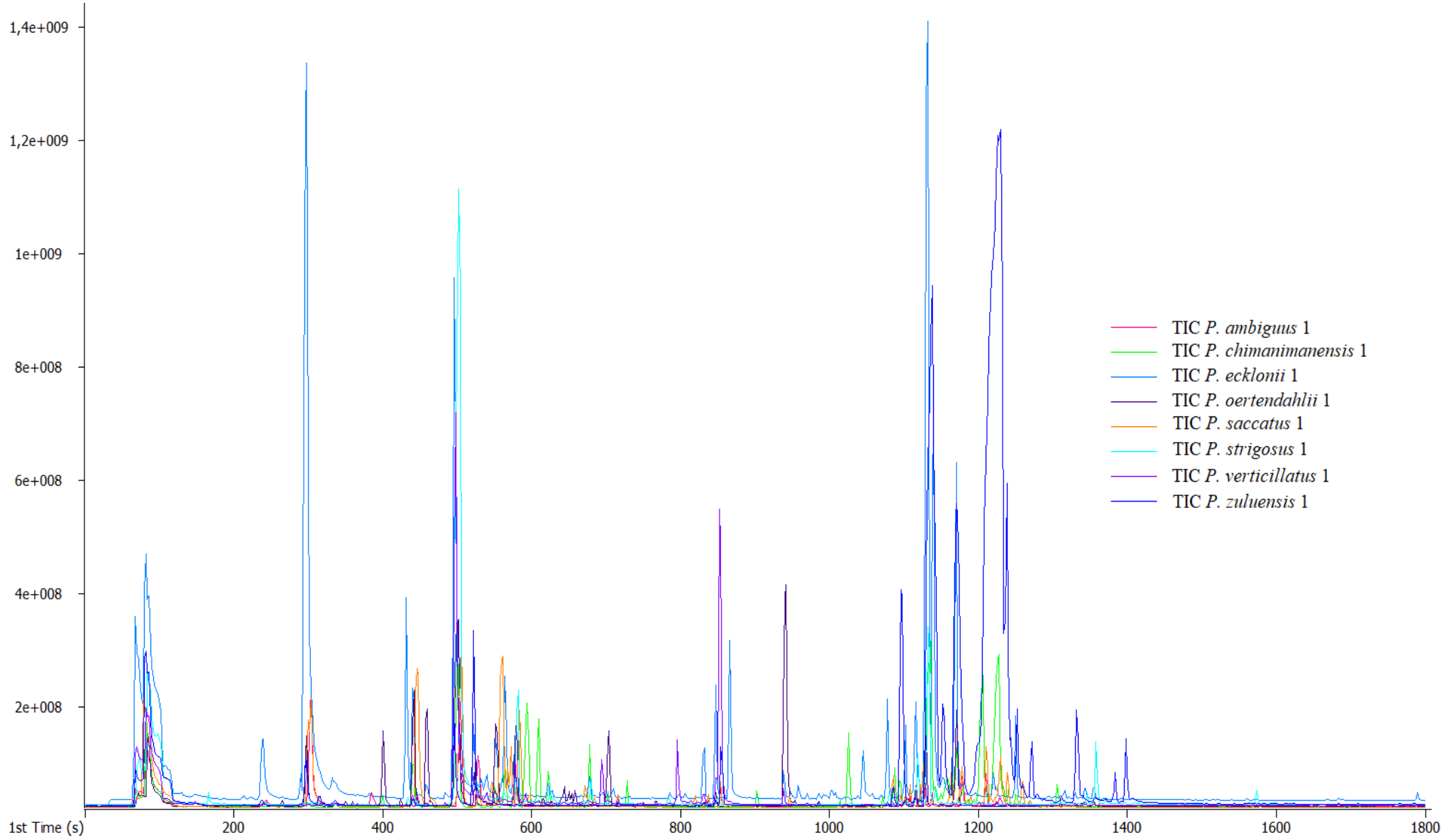


Figure 39: Overlay of the 1D TICs of single replicates of species of genus *Plectranthus* (c.f: Figure 38), with *P. fruticosus* excluded.

5.2) Olfactory descriptors

The leaves of southern African species of *Plectranthus* and *Coleus* have fairly distinct odour profiles, qualitative descriptors of which were derived (N=2 human noses) and reported in Table 1:

Table 1: Olfactory descriptors of the odour profiles of the leaves of the *Plectranthus* and *Coleus* species included in the study.

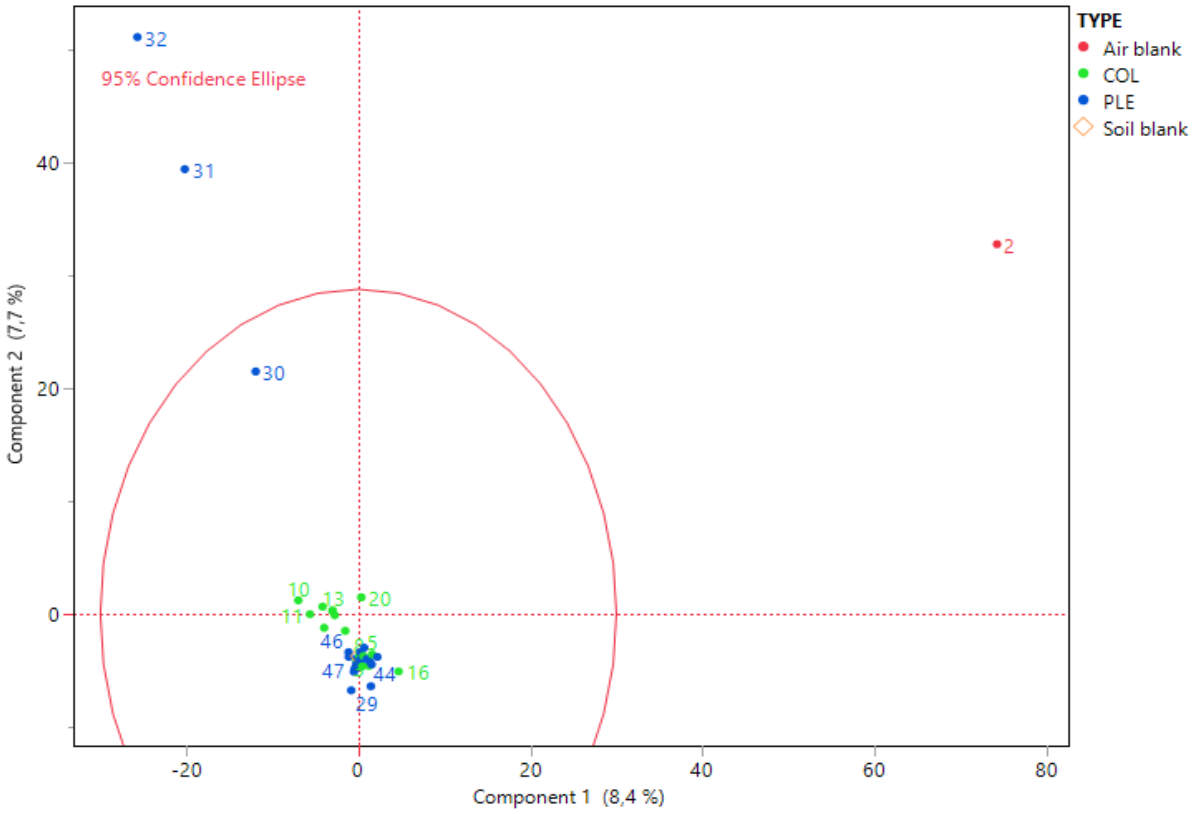
Species	**Olfactory descriptor**
C. hadiensis	Mild herbaceous-aromatic
C. hereroensis	Musty-peppery
C. livingstonei	Moderate grass-like
C. longipetiolatus	Pungent peppery
C. madagascariensis	Moderate sweet-aromatic
C. neochilus	Musty-aromatic
P. ambiguus	Moderate spicy (clove-like)
P. chimanimanensis	Faint herbaceous (rosemary-like)
P. ecklonii	Mild herbaceous (oreganum-like)
P. fruticosus	Strong-lavender
P. oertendahlii	Moderate herbaceous (like unripened berries)
P. strigosus	Moderate herbaceous
P. saccatus	Moderate grape-like
P. verticillatus	Faint neutral
P. zuluensis	Moderate peppery-herbaceous

Species with odours of strong intensity, such as *C. longipetiolatus* (Figure 13 and Figure 14), *C. neochilus* (Figure 17 and Figure 18) and *P. fruticosus* (Figure 25 and Figure 26) have chromatograms with overloaded peaks — for compounds including 1-octen-3-ol; camphene; β-pinene; 3,7-dimethyl-1,6-octadiene; cis-β-farnesene; caryophyllene and bornyl acetate — whereas species with feint to moderate odours, such as *P. ambiguus* (Figure 19 and Figure 20), *P. strigosus* (Figure 31 and Figure 32) and *P. verticillatus* (Figure 33 and Figure 34) have chromatograms of sparser peak distribution and intensity. Species with mild to moderate scents have chromatograms of intermediate complexity and peak intensity, including such species as *C. hadiensis* (Figure 7 and Figure 8), *P. chimanimanensis* (Figure 21 and Figure 22) and *P. ecklonii* (Figure 23 and Figure 24), however, *P. zuluensis* (Figure 35 and Figure 36), which like *C. longipetiolatus* has a "peppery" descriptor, also has overloaded peaks (including cis-β-farnesene and caryophyllene) in the higher retention time range. The sesquiterpene caryophyllene is associated with a "peppery" odour, and derivatives of hexenal are generally associated with "herbaceous" odours [4]. Overall, sesquiterpenes tend to have a "herbal" odour quality [4]. However, a complete olfactory analysis of the molecular basis of these odours is beyond the scope of this study.

5.3) Preliminary statistical analysis: PCA and LDA

As a preliminary assessment of variation in the foliar VOC profiles of members of *Plectranthus* and *Coleus*, principal component analysis (PCA) and linear discriminant analysis (LDA) were performed on the full dataset[3], described by 1794 variables (compounds), prior to processing for supervised learning.

Figure 40 is a PCA score plot for the first and second components, of the blank (air and soil), *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with the 95% confidence ellipse shown, and reveals the clustering of the samples in terms of the predictor loading scores. Figure 41 is an amplified view of the same plot. The loading matrix and the tabulated eigenvalues are provided in Appendix A.1. Figure 42B, a biplot, is an overlay of the loading plot for the predictors with the score plot. PCA decomposes the full dataset into 45 components capturing 99.95% of the variance, the eigenvalues of which are plotted in Figure 42A. The first component, which accounts for the greatest variation, represents only 8.4% of the cumulative total, suggesting that the variation is not strongly influenced by a small number of predictors.



1) Soil blank; 2) air blank; 3-5) *C. hadiensis*; 6-8) *C. hereoensis*; 9-11) *C. livingstonei*; 12-14) *C. longipetiolatus;* 15-17) *C. madagascariensis*; 18-20) *C. neochilus*; 21-23) *P. ambiguus*; 24-26) *P. chimanimanensis*; 27-29) *P. ecklonii*; 30-32) *P. fruticosus*; 33-35) *P. oertendahlii*; 36-38) *P. strigosus*; 39-41) *P. saccatus*; 42-44) *P. verticillatus*; 45-47) *P. zuluensis*.

Figure 40: Score plot for the first two principal components of the blank (air and soil), *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with the 95% confidence ellipse shown.

[3] The full foliar VOC peak area dataset of 1794 variables is available upon request.

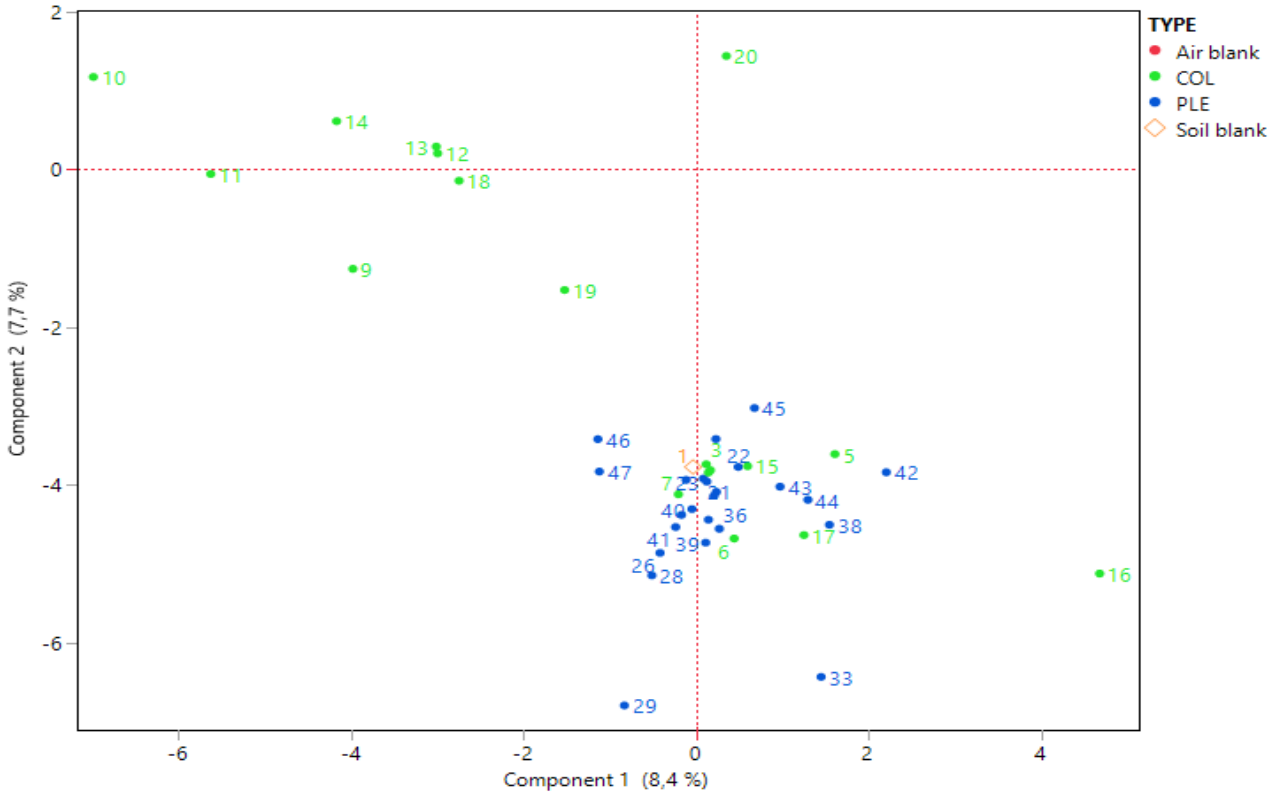


Figure 41: Amplified view of the region around the origin in the score plot in Figure 40.

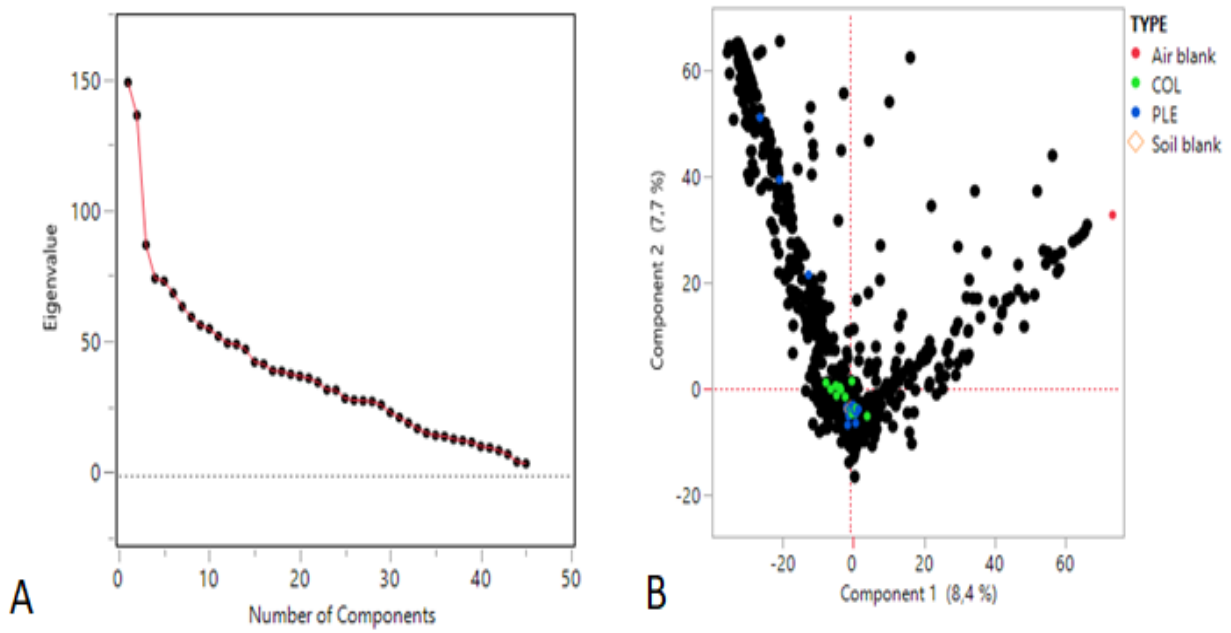
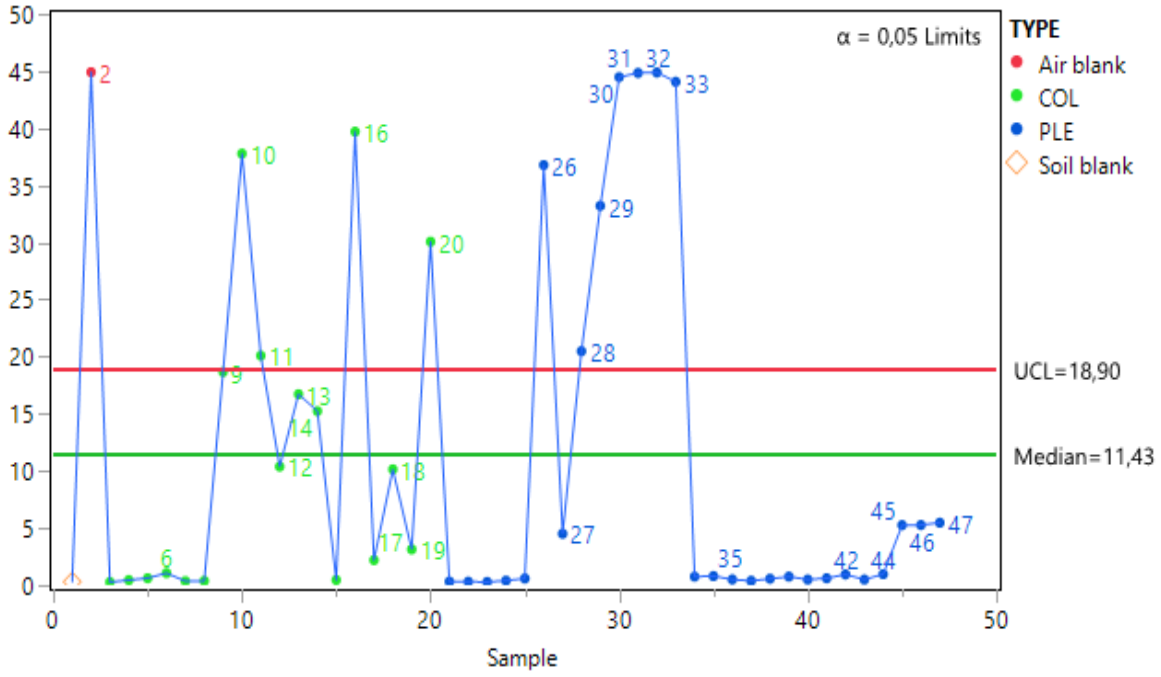


Figure 42: A) Scree plot of the eigenvalues for the 45 principal components; B) PCA biplot, overlaying the score plot in Figure 40 and the loading plot for the predictors (point labels for the latter not provided for the sake of clarity).

The foliar samples, with the exception of the three replicates of *P. fruticosus*, cluster within ±5-7 units of the origin of the first component. The *P. fruticosus* replicates are loosely distributed along both components, and appear to be outliers. A multivariate outlier test using Hotelling's T^2 statistic reveals twelve outliers at a significance threshold of $\alpha = 0.05$, and includes the *P. fruticosus* samples (Figure 43). The variation observed is likely due to

underlying, and potentially informative, biological variation, and not necessarily determinate error. Since a test for a hypothesis of difference was not the purpose of the PCA in this study, and since biological variation is expected, the points identified as outliers were retained.



1) Soil blank; 2) air blank; 3-5) *C. hadiensis*; 6-8) *C. hereoensis*; 9-11) *C. livingstonei*; 12-14) *C. longipetiolatus;* 15-17) *C. madagascariensis*; 18-20) *C. neochilus*; 21-23) *P. ambiguus*; 24-26) *P. chimanimanensis*; 27-29) *P. ecklonii*; 30-32) *P. fruticosus*; 33-35) *P. oertendahlii*; 36-38) *P. strigosus*; 39-41) *P. saccatus*; 42-44) *P. verticillatus*; 45-47) *P. zuluensis.*

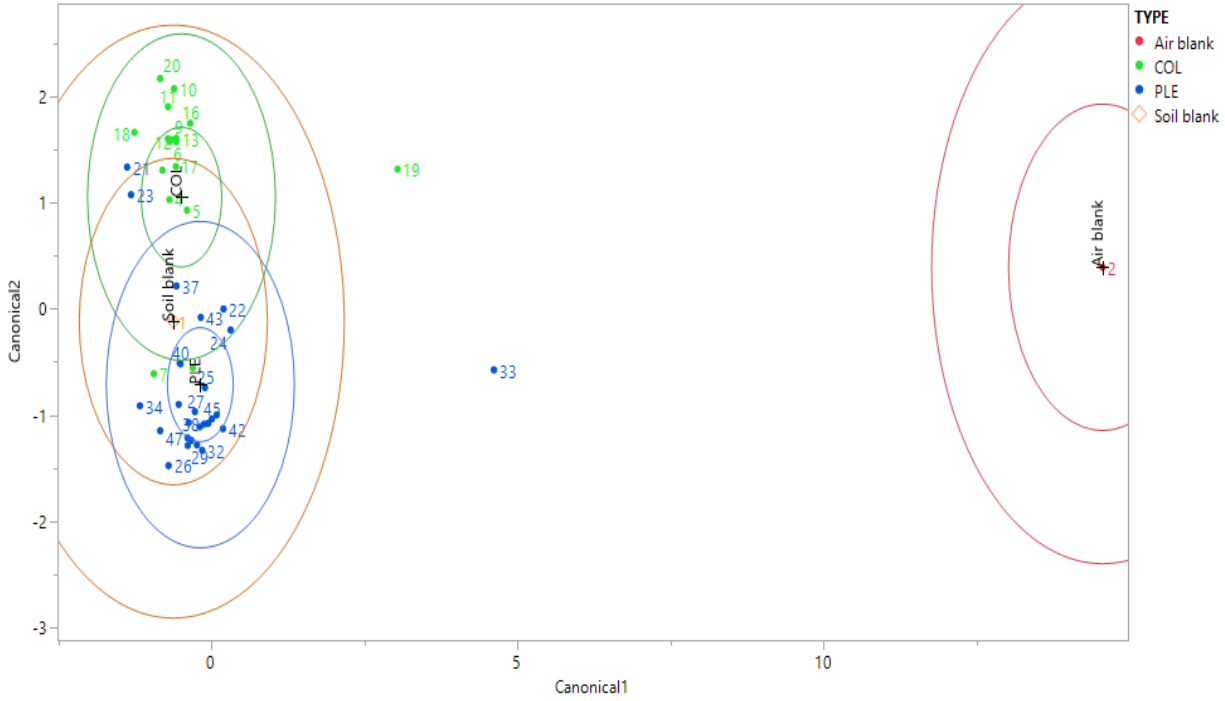
Figure 43: Outlier analysis using Hotelling's T^2 statistic (α=0.05); UCL = upper confidence limit.

The air-blank was included in the preliminary statistical analysis in order to confirm that it could be clearly differentiated from the samples after blank correction (as described in Chapter 4.6.1) for air-borne contaminants. The point for the air-blank falls far to the right of the cluster of samples (Figure 40), at high values of the first component, which indicates that the blank correction procedure was effective. The point corresponding to the soil blank, however, falls within the main cluster of samples. This is expected in light of the fact that soil components tend to accumulate inside leaves. It should be noted that the soil blank itself is corrected by the air blank, which leads to negative peak area values for the former. Since a negative peak area is not meaningful, these values were converted to zero during processing, which resulted in a value of zero for all predictors for the soil sample. For this reason, the point for the soil blank on the score plot falls close to the origin, in the region of low eigenvalues for the first two components. This underlies the importance of the soil blank correction, as the variation observed in the PCA plot is not weighted strongly towards compounds present in both the soil and the leaves.

An amplified view of the region near the origin of the score plot (Figure 41) shows species of both genera to cluster within about two units either side of the origin of the first components,

and within about two units of the negative region of the second component. However, the replicates of *C. neochilus* (18-20), *C. livingstonei* (9-11), and *C. longipetiolatus* (12-14) are distributed along a greater distance, and at greater absolute values, of the first component, and are separated from the cluster of *Plectranthus* samples along both components. This suggests that these samples are characterised by high variance predictors (predictors with high loading scores). This is consistent with the observation made in Chapter 5.1 that these three species are characterised by TICs of greater complexity, and thus greater foliar VOC diversity.

LDA plots the data according to the maximal separability between samples of different groups, and produces analogous results to PCA. Figure 44 is a canonical plot for the supervised clustering of the foliar samples (*Plectranthus* and *Coleus*), the air blank, and the soil blank, and Figure 45 is the 3D projection of the plot. The two genera are clustered distinctly, but with overlap of the normal ellipses around the regions containing 50% of the members of each group respectively. The point for the air blank is well removed from the sample points, and the blank occupies an intermediate position between the genera, similar to what is observed for the principal component score plot (Figure 40). Two points — for *P. oertendahlii* (33) and *C. neochilus* (19) — appear to be outliers, although in the 3D canonical plot (Figure 45) these points still reside in the 95% confidence ellipse for the true group mean.



1) Soil blank; 2) air blank; 3-5) *C. hadiensis*; 6-8) *C. hereoensis*; 9-11) *C. livingstonei*; 12-14) *C. longipetiolatus*; 15-17) *C. madagascariensis*; 18-20) *C. neochilus*; 21-23) *P. ambiguus*; 24-26) *P. chimanimanensis*; 27-29) *P. ecklonii*; 30-32) *P. fruticosus*; 33-35) *P. oertendahlii*; 36-38) *P. strigosus*; 39-41) *P. saccatus*; 42-44) *P. verticillatus*; 45-47) *P. zuluensis*.

Figure 44: Canonical plot for the blanks (air and soil) and the *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with 95% (outer) and 50% (inner) confidence ellipses shown.

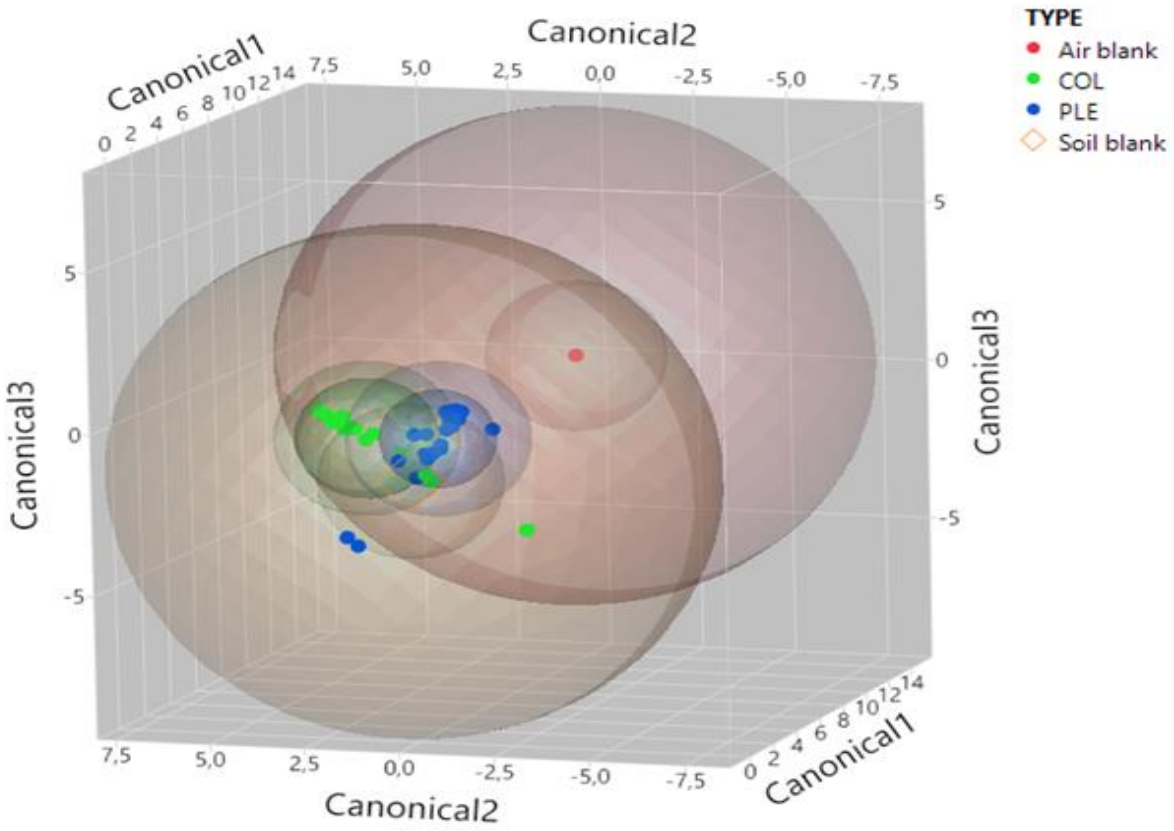


Figure 45: 3D canonical plot for the blanks (air and soil) and the *Coleus* (COL) and *Plectranthus* (PLE) foliar samples, with 95% (outer) and 50% (inner) confidence ellipses shown. Sample labels omitted for clarity.

5.4) Data pre-processing for machine learning

After the removal of peaks arising from inorganic contaminants (Chapter 4.6.1), the full dataset of foliar VOC profiles, consists of 45 observations (replicate samples) described by 1794 predictors, corresponding to 1794 compounds identified by mass-spectral similarity scoring (with a minimum threshold of 75%) using reference spectra from the NIST database. Splitting of the data by a 0.5 ratio at the set seed number (Chapter 4.6.3-4) results in 22 samples in the training set and 23 samples in the testing set. The near-zero variance removal function of the pre-processing step, performed on the 22 samples in the training set, results in the removal of 1160 predictors, and the retention, centring and scaling of 634 predictors. This can be interpreted as a 65% reduction in the dimensionality of the dataset, and emphasises the importance of the pre-processing step in paring the dataset of variables that are uninformative in terms of genus-distinction. A PCA can be included in the pre-processing step, which results in a model consisting of 18 principal components accounting for 95% of the variance, compared to the 45 components for 99.95% of the variance for the PCA performed in the preliminary statistical analysis (Chapter 5.3) on the full dataset. Again, this demonstrates the

utility of the pre-processing step in capturing those predictors that account for any potentially significant variation in the data.

5.5) Training, tuning and model selection

As discussed in Chapter 2.3.2, in order to optimise the accuracy of a predictive model, it is necessary to systematically tune its parameters. Essentially, tuning involves the assessment of the AUC/ROC for a number of models, constrained by different values of their coefficients/parameters, *via* in-training resampling, and the selection of the model with the highest AUC/ROC. Each of the three algorithms used in this study were trained and tuned over a range of parameters. Figure 46, Figure 48 and Figure 49 are plots of the AUC/ROC for different parameters of the elastic-net regression, random forest and support-vector machine. The tabulated data is included in Appendix A.2.

For the elastic-net regression (Figure 46), the tuning parameters are alpha (α), or the mixing percentage, which corresponds to the ratio of the ridge-to-lasso penalty type, and lambda (λ), which corresponds to the magnitude of the penalty. A ridge regression penalty is taken as the square of the coefficients, whereas a lasso penalty is taken as the absolute value of the coefficients (Chapter 2.4.1). Consequently, coefficients in a lasso regression can reach zero and become excluded, producing a sparser model, whereas those of a ridge regression can only tend towards zero, and are thus retained. In this case, the optimal model, with an AUC/ROC of 0.82, is a 95% (α=0.95) lasso-regression elastic-net with a small penalty (λ=0.1112). Notably, the accuracy of models with a high lasso-regression composition (values of alpha greater than 0.5) decreases at higher values of lambda due to models becoming overly sparse from variable reduction. In contrast, the accuracy of models with a greater ridge-type penalty, which retain most or all of the variables, are not as much affected. A pure ridge-regression (α=0) demonstrates constant performance across all values of lambda. This may imply that model accuracy is more sensitive to variable reduction than to the magnitude of the coefficients.

Figure 47 plots the full regularisation (or tuning) path of the elastic-net along the L_1 norm. High values of the L_1 norm equate to low values of lambda. As the penalty of lambda increases, the coefficients are decreased, until they reach zero and are excluded from the model. Model complexity thus decreases with an increase in lambda, until, at sufficiently high penalties, all coefficients are excluded from the model and the intercept-only model (a regression equation consisting of zero predictors) is reached. Overly simplified models show poor predictive accuracy, which is reflected by the drop in the AUC/ROC (Figure 46) for high penalties.

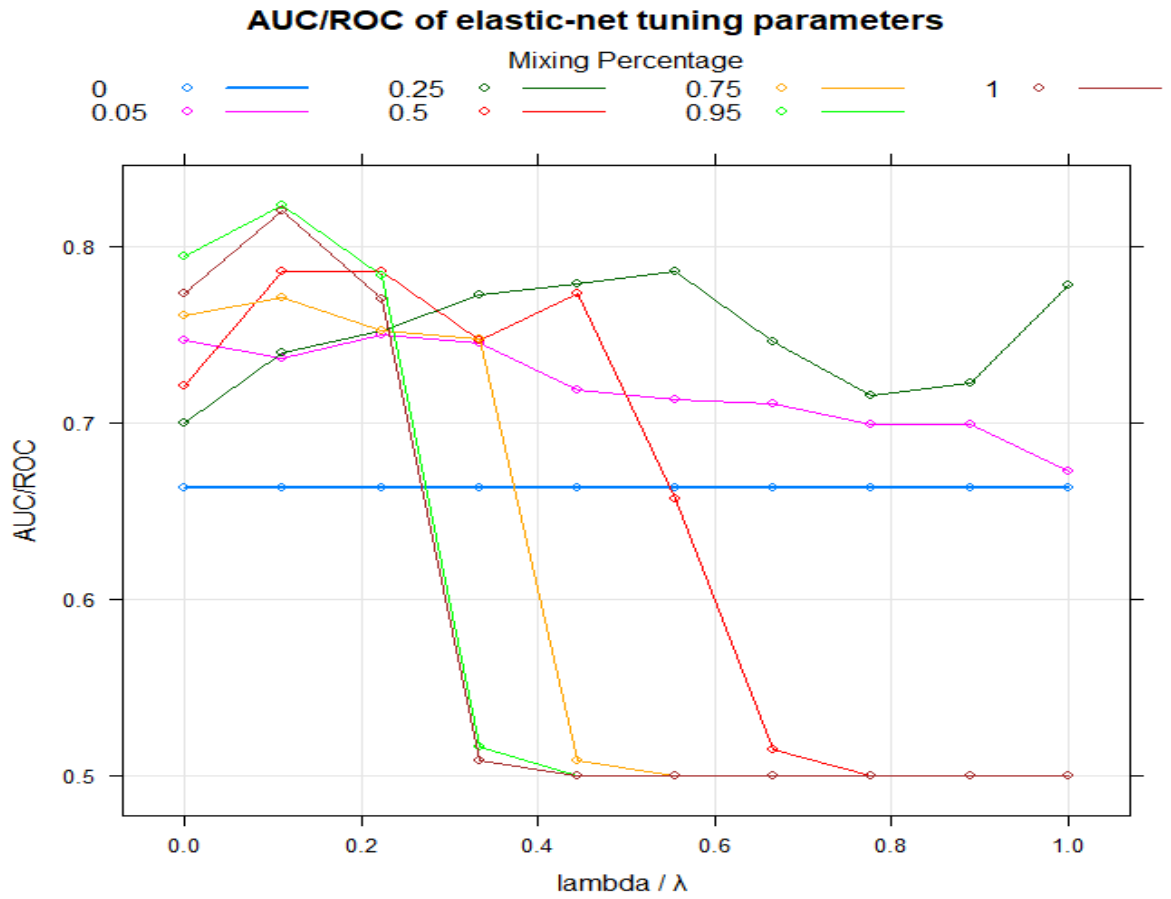


Figure 46: AUC/ROC (by cross-validation) of the elastic-net tuning parameters; mixing percentage = alpha (α).

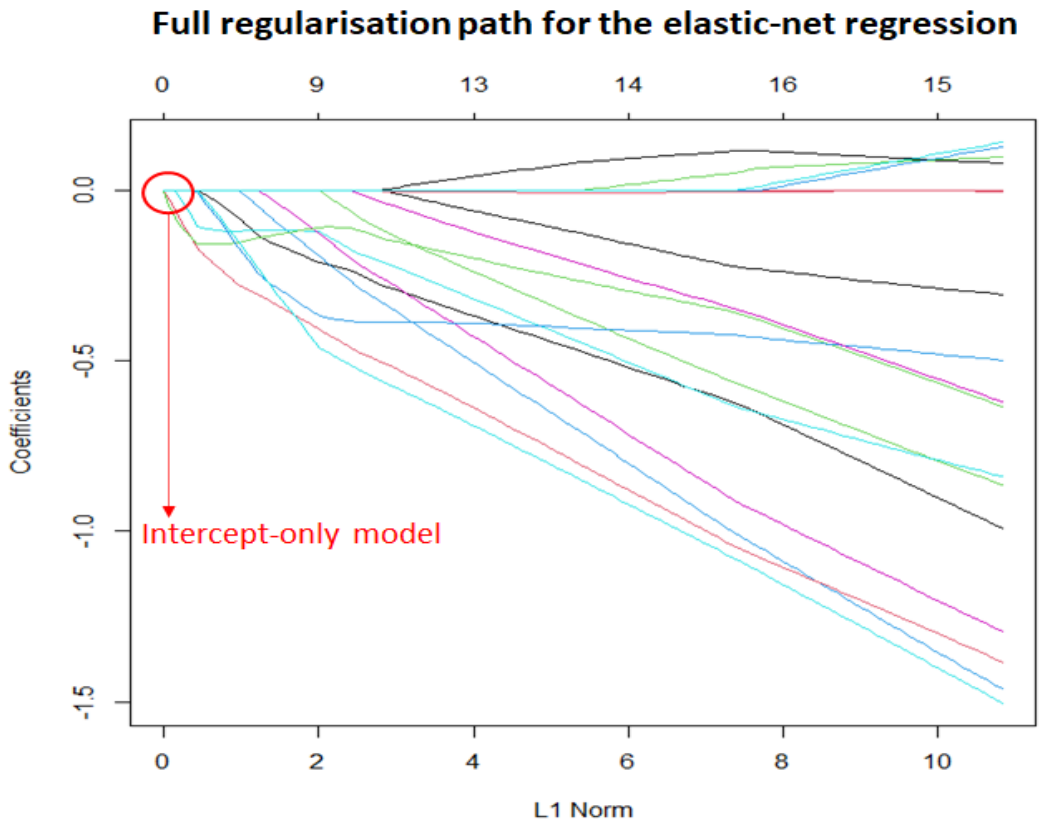


Figure 47: Full regularisation (tuning) path for the elastic-net regression. Model complexity decreases from right to left as the value of lamda (λ) is increased.

The algorithm (ranger) for the random forest specifies three parameters: 1) the splitting rule (gini or extratrees), 2) the minimal number of nodes at each split (minimal node size), and 3) the number of randomly selected predictors for each node (mtry). The results of tuning for different random forests are summarised in Figure 48. The model of optimal AUC/ROC (0.93), by cross-validation, in terms of these specifications, is one split according to the gini rule, with a minimal node size of one, and three randomly selected predictors for each node. Overall, the gini rule produces more stable AUC/ROC values, however all parameters result in high values.

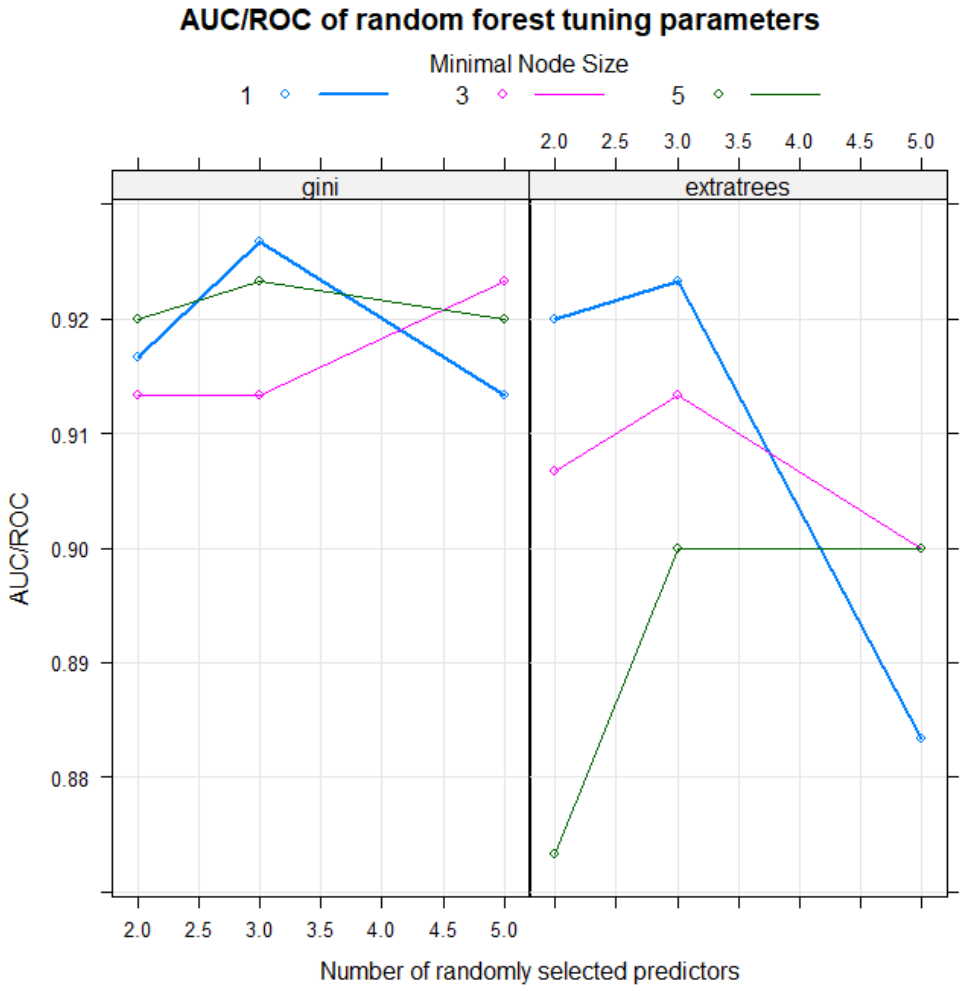


Figure 48: AUC/ROC of the random forest tuning parameters: the splitting rule (gini, left; extratrees, right), the number of randomly selected predictors at each node split, and the minimal node size.

Figure 49 is a plot of the AUC/ROC, by cross-validation, for the support-vector machine with a polynomial kernel function. The models are described by three parameters: the cost (C), the degree of the polynomial function, and the scale of the function (Chapter 2.4.3). For all values of C and each scale, a first-degree polynomial, gives the highest AUC/ROC values. A second-degree polynomial produces similar results at very low values of C. Overall, an increase in C beyond one does not significantly improve, and may even decrease, model performance. The

optimal model, with an AUC/ROC of 0.62, is a first-degree polynomial with a scale of three, and C=0.1 (Figure 49). Thus, a linear kernel function is able to accurately model the data.

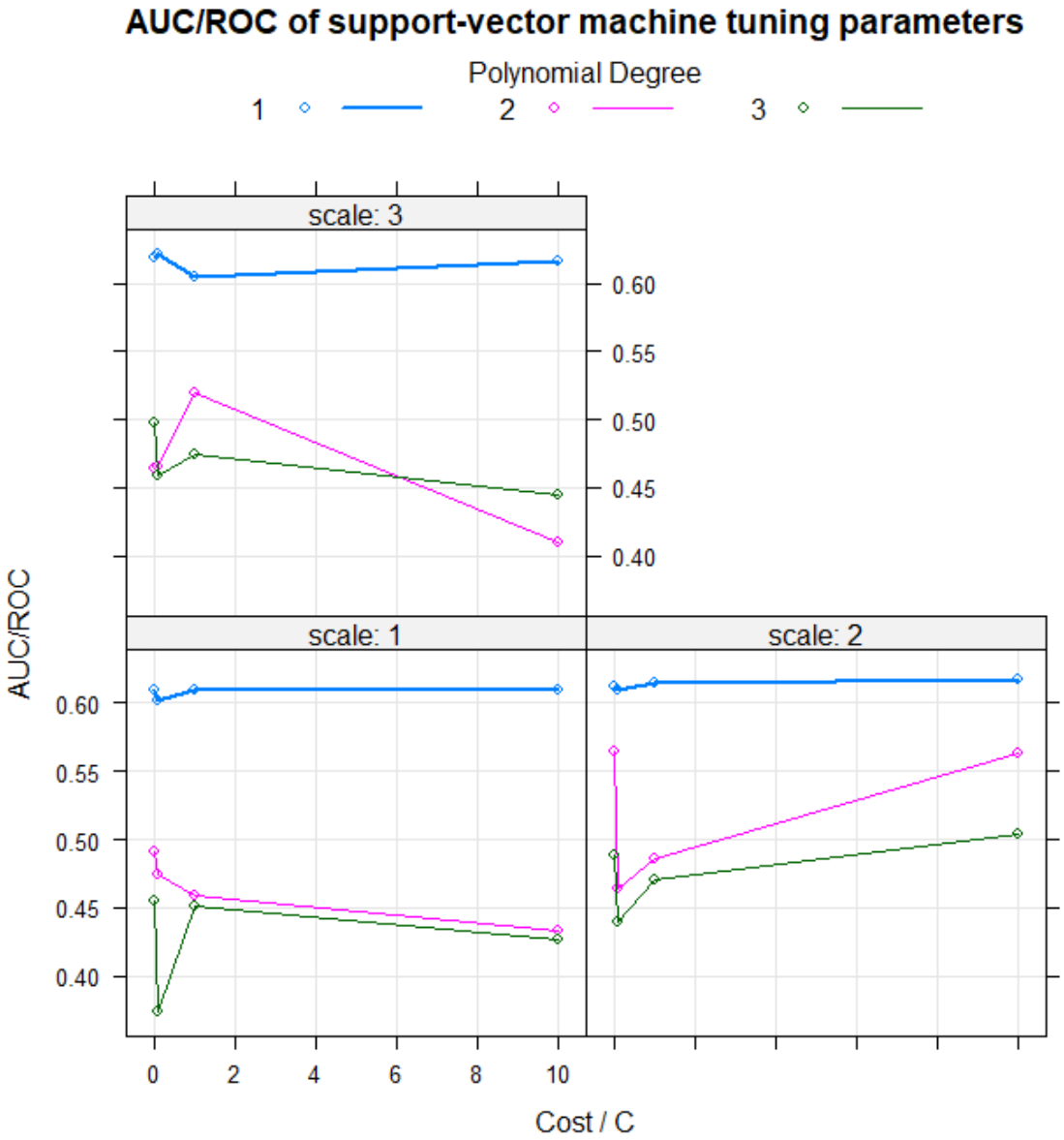


Figure 49: AUC/ROC of the support-vector machine (with a polynomial kernel) tuning parameters: the polynomial degree, the scale and cost (C).

5.6) Testing and prediction

In order to assess the predictive accuracy of a model, it must be applied to unknown observations in a testing set. In other words, it must be able to predict the category of samples that it was not trained on, and that were not used as input in its construction. Prediction involves inputting sample data, and calculating the probability that the given sample falls under a given category, in this case genus *Plectranthus* or *Coleus*. Table 2 lists the probabilities and corresponding predicted categories for each of the 23 samples in the testing set. The predictions of each model are summarised in a confusion matrix (Figure 50) with corresponding

performance statistics. In this case, genus *Coleus* is defined as the *positive* class, and genus *Plectranthus* the *negative* class.

Table 2: Computed probabilities of samples in the testing set belonging to either genus *Coleus* (COL) or *Plectranthus* (PLE) and their predicted genus, for the elastic-net, random forest and support-vector machine models.

	Elastic-net			Random forest			Support-vector machine		
	Genus probability		Predicted genus						
Sample	COL	PLE		COL	PLE		COL	PLE	
P, ambiguus 1	0,1987	0,8013	PLE	0,3270	0,6730	PLE	0,3243	0,6757	PLE
P, ambiguus 2	0,2098	0,7902	PLE	0,3390	0,6610	PLE	0,2126	0,7874	PLE
C, livingstonei 2	0,9222	0,0778	COL	0,6737	0,3263	COL	0,9995	0,0005	COL
C, hereroensis 1	0,6377	0,3623	COL	0,5377	0,4623	COL	1,0000	0,0000	COL
C, longipetiolatus 3	0,3820	0,6180	PLE	0,5846	0,4154	COL	1,0000	0,0000	COL
P, strigosus 2	0,2487	0,7513	PLE	0,3156	0,6844	PLE	0,1900	0,8100	PLE
P, fruticosus 1	0,3200	0,6800	PLE	0,5098	0,4902	COL	1,0000	0,0000	COL
P, saccatus 1	0,1998	0,8002	PLE	0,3470	0,6530	PLE	0,2795	0,7205	PLE
P, chimanimanensis 3	0,1925	0,8075	PLE	0,2530	0,7470	PLE	0,0001	0,9999	PLE
P ecklonii 1	0,6419	0,3581	COL	0,4481	0,5519	PLE	1,0000	0,0000	COL
C, longipetiolatus 2	0,4561	0,5439	PLE	0,5724	0,4276	COL	0,9291	0,0709	COL
C, hadiensis 1	0,6939	0,3061	COL	0,5684	0,4316	COL	0,7513	0,2487	COL
P, saccatus 3	0,2081	0,7919	PLE	0,3245	0,6755	PLE	0,1406	0,8594	PLE
P, ecklonii 2	0,4035	0,5965	PLE	0,4341	0,5659	PLE	0,2961	0,7039	PLE
P, oertendahlii 2	0,1922	0,8078	PLE	0,2000	0,8000	PLE	0,0001	0,9999	PLE
C, madagascariensis 3	0,5864	0,4136	COL	0,5591	0,4409	COL	0,9342	0,0658	COL
P, fruticosus 2	0,4992	0,5008	PLE	0,4595	0,5405	PLE	0,0001	0,9999	PLE
P, verticillatus 2	0,2054	0,7946	PLE	0,3961	0,6039	PLE	0,0943	0,9057	PLE
P, zuluensis 2	0,2078	0,7922	PLE	0,2952	0,7048	PLE	0,1140	0,8860	PLE
P, verticillatus 3	0,1969	0,8031	PLE	0,3528	0,6472	PLE	0,4221	0,5779	PLE
P, ambiguus 3	0,1978	0,8022	PLE	0,3250	0,6750	PLE	0,2592	0,7408	PLE
P, strigosus 1	0,2659	0,7341	PLE	0,3464	0,6536	PLE	0,1476	0,8524	PLE
C, hereroensis 2	0,7670	0,2330	COL	0,4808	0,5192	PLE	1,0000	0,0000	COL

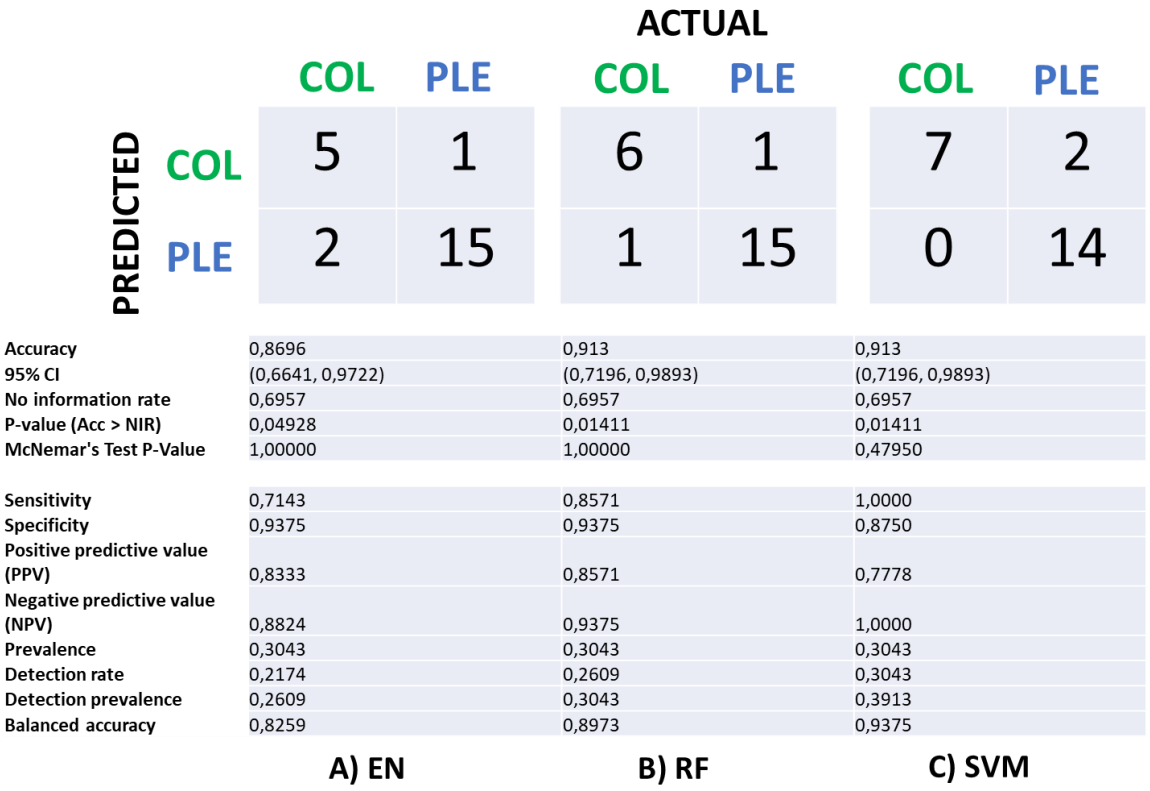


Figure 50: Confusion matrix and associated statistics for the predictions of the tuned models: A) elastic-net (EN), B) random forest (RF), and C) support-vector machine (SVM). COL = *Coleus*; PLE = *Plectranthus*.

The no-information rate (NIR) is equivalent to the percentage of samples in the majority category (in this case *Plectranthus*), or in other words, the percentage of true negatives and false positives in the testing set. The NIR can be interpreted as the predictive accuracy obtained by a null model that classifies every sample as *negative*/*Plectranthus*. For a model to be considered predictive, it must have a higher predictive success rate than is obtained from null prediction, which is to say, the accuracy of the model must be greater than the NIR [5]. For each of the three models, the accuracy (the percentage of observations correctly classified) is high (more than 85%) and greater than the NIR (70%) (Figure 50), and thus the models can be concluded to be predictive. The p-values for the accuracy being greater than the NIR are less than 0.05 for all cases, indicating a fairly low probability of model predictions being due to chance, especially given the limited sample size and imbalance in the proportion of *Plectranthus* samples to *Coleus* samples. However, it should be emphasised that testing was performed on a split of the dataset, and may thus be overfit to this dataset, even if the testing set was not used for model construction. The performance of the model may suffer if applied to a testing set sampled and analysed independently of the training set (for example, using different individual specimens for the training and testing groups, or collecting the samples for the training set one summer, and those for the second set the following summer).

Two key statistics are the sensitivity and the specificity (Figure 50) [6]. The sensitivity is the proportion of true positive cases of those predicted to be positive. Since *Coleus* is defined as positive, the sensitivity measures the ability of the model to correctly classify foliar VOC samples of species from genus *Coleus*. The support-vector machine demonstrates the highest sensitivity (100%), followed by the random forest (86%) and the elastic-net (71%). Two similar statistics are the detection rate, which is the proportion of true positives of all samples in the testing set, and the detection prevalence, the percentage of true and false positives. These values are expected to be lower than the sensitivity, since they are calculated as proportions of the total testing set, but they still indicate accurate predictive performance. The high sensitivity of the three models tested means that they each also have a high positive prediction value (PPV). The PPV accounts for the sensitivity, the prevalence (the proportion of positive/*Coleus* samples) and the false positive rate [6].

The specificity is the proportion of the true negative cases of all the cases predicted to be negative, and in this case measures the ability of the model to correctly classify samples from genus *Plectranthus*. The elastic-net and the random forest have the highest specificity (94%). Since the specificity for each model is high, the negative prediction value (NPV) is significant for each model. The NPV accounts for the specificity, the false negative rate, and the proportion of negative/*Plectranthus* samples in the testing set [6].

5.7) Variable importance ranking

The final step of the machine learning pipeline is to identify key predictors that distinguish the categorical types in question. The algorithms of the caret package have inbuilt functions available for the ranking of variable importance (Chapter 2.3.4). The ranking functions of the elastic-net and the random forest are based on the parameters of the models themselves. For the elastic-net regression, the absolute values of the regression coefficients are used to rank predictors. For the random forest, normalised mean differences in the accuracy of predictors, determined by bootstrapping, are computed. The variable importance function of the support-vector machine is not based on the model itself, but determines the AUC/ROC for each predictor over a range of thresholds [6].

Table 3: List of the top predictor compounds for the elastic-net, random forest and support vector machine models.

Elastic-net		**Random forest**		**Support-vector machine**	
Predictor compound	**Scaled score**				
Cyclohexanone, 2,2,6-trimethyl-	100.00	(E)-4-Oxohex-2-enal	100	Hexanal	100.00
β-Ylangene	90.000057	Furan-2-ethyl	71.42	(E)-4-Oxohex-2-enal	90.65
Hexanal	72.96508	Hexanal	68.29	β-Cubebene	86.92
(E)-4-Oxohex-2-enal	54.65623	β-Ylangene	68.26	β-Ylangene	83.18
Octane, 1,1'-oxybis-	47.25864	α-Cubebene	67.81	β-Copaene	81.31
1,3-Octadiene	41.887	1,3-Octadiene	67.34	Furan-2-ethyl	77.57
1,2,4,4-tetramethylcyclopentene	35.55628	2-Hexenal	64.77	α-Cubebene	75.70
Furan-2-ethyl	22.40628	3-Hexen-1-ol, acetate, (Z)-	62.61	Isogermacrene D	70.09
1-Octene, 6-methyl-	15.73988	α-Cadinene	61.26	trans-β-Ionone	70.09
1-Hexene	1.9706	Isogermacrene D	58.85	1,3-Octadiene	66.36
1H-Pyrazole, 4,5-dihydro-5,5-	0.07446	trans-β-Ionone	57.91	Propanoic acid, 2-methyl-, anhydride	66.36
dimethyl-4-isopropylidene-		γ-Cadinene	56.59	2,4-Hexadienal, (E,E)-	64.49
		α-Selinene	54.97	γ-Cadinene	64.49
		β-Cubebene	54.25	3-Hexen-1-ol, (Z)-	62.62
		(Z,E)-α-Farnesene	53.68	1,2,4,4-tetramethylcyclopentene	61.68
		β-Calacorene	52.43	2(5H)-Furanone, 5-ethyl-	61.68
		Decanal	51.24	β-Calacorene	60.75
		2(5H)-Furanone, 5-ethyl-	49.18	3-Hexen-1-ol, acetate, (Z)-	60.75
		Octane, 1,1'-oxybis-	48.49	α-Cadinene	58.88
		Bornyl acetate	47.42	Octane, 1,1'-oxybis-	57.01

Table 3 lists the twenty top-ranking predictors/compounds for the three models, as well as their scaled scores[4]. A few of the compounds are common to more than one of the output lists in Table 3, although their specific ranks and scores differ. A number of the top variables are molecules belonging to the isomeric class of sesquiterpenes, of molecular formula $C_{15}H_{24}$. These tentatively identified species include β-ylangene, α-cubebene, β-cubebene, β-copaene, α-cadinene, γ-cadinene, isogermacrene D, α-selinene and (Z,E)-α-farnesene. A bicyclic sesquiterpene, β-calacorene ($C_{15}H_{20}$), is also ranked as a top predictor.

[4] The full scaled-score outputs for the pre-processed set of 634 variables, for each algorithm, are available upon request.

The other high-ranking compounds are of diverse organic classes of lower molecular weight than the sesquiterpenes. This includes a group of C_6 compounds, including furan, 2-ethyl, 1-hexene, hexanal, 2-hexenal, (E,E)-2,4-Hexadienal, (Z)-3-hexen-1-ol, and (Z)-3-hexen-1-ol, acetate, likely to be GLVs— VOCs which are released upon rupture of the foliar tissue, or with changes in temperature and light [1-3], and which give rise to the additional peaks observed for the crushed-leaf TIC of *C. neochilus* in Figure 5 (Chapter 5.1). Both hexanal, hexenal- and hexanol-type GLVs have been reported in a study using proton-transfer-reaction time-of-flight mass spectrometry (PTR-TOF-MS) [3]. Also included are C_6-related compounds such as the unsaturated ketoaldehyde, (E)-4-oxohex-2-enal, and cyclohexanone, 2,2,6-trimethyl-, which are also likely to be GLVs.

Other compounds listed as top-predictors include $C_{8\text{-}9}$ species, including a diene (1,3-octadiene), an unsaturated hydrocarbon, 1-octene, 6-methyl-, and an acid anhydride (propanoic acid, 2-methyl-, anhydride) and a cyclic alkene (1,2,4,4-tetramethylcyclopentene). In addition, there is an ionone (trans-β-ionone), an aldehyde (decanal), a long-chain unsaturated hydrocarbon (octane, 1,1'-oxybis-), a terpene derivative (bornyl acetate) and a pyrazole derivative, the only nitrogenous species on the list.

Notably, there are no monoterpenes or monoterpenoids with high variable importance scores, which could mean that although both genera produce a diversity of monoterpenes, there is no genus-differentiation with regards to their production.

The top compound for the elastic-net regression is the cyclic ketone, cyclohexanone, 2,2,6-trimethyl-, for the random forest the unsaturated aldehyde, (E)-4-oxohex-2-enal, and for the support-vector machine the aldehyde hexanal (Table 3). The two latter aldehydes are within the top four ranks for all three models. Overall, the sesquiterpene compounds do not feature on the output list of the elastic-net regression, which moreover, is most different from the other two algorithms in terms of its variable output. However, β-ylangene has the second highest score for the elastic-net (90), and the fourth highest scores for the random forest and support-vector machine (83 and 68, respectively).

The elastic-net regression outputs only 12 top variables, with all other scores being equal to zero. This is an unexpected result considering the small value of lambda (λ=0.1112; Chapter 5.5), and implies that the values of most of the coefficients are low enough that they are nonetheless excluded, resulting in a sparse model. The final model for the glmnet is different from the other two models in this regard, and in addition, has differences in terms of its variable output. For instance, the sesquiterpenes, with the exception of β-ylangene, are not selected as important features, and the highest scoring variable, cyclohexanone, 2,2,6-trimethyl-, is not on the list for the random forest or the elastic-net.

In short, a variety of VOCs are selected as top predictors of genus for *Plectranthus* and *Coleus*. In particular, sesquiterpenes feature as important variables, and are thus concluded to be candidate markers.

5.8) Retention indices of top-ranking compounds

Retention indices (RIs) for selected top-ranking compounds were calculated from a homologous series of *n*-alkanes (C_6-C_{28}), as described in Chapter 4.5, and are reported in Table 4, along with the unique CAS number[5] for the tentative identification and the mass spectral similarity to reference spectra from the NIST database. RIs were calculated from the equation for Kovats retention index, as well as from a least-squares linear regression equation of the linear retention indices of reference *n*-alkanes.

Literature RI values are provided for comparison to the experimental values. In a majority of cases (indicated with a cross[†]), the values are found to be in agreement within ±50 RI units for at least one of the calculations (via the equation for Kovats RI or from the regression equation).

It should be emphasised that for many compounds in the dataset, and in particular for those selected as top features, there are a number of peaks, with different retention times, of the same mass spectral identification or *hit*. This is due in some instances to overloaded peaks, however, visual inspection of the chromatograms revealed certain separated peaks of the same identification. This may mean that the summed peak area values for each compound in a sample, as well as the average peak area for each compound across all samples, are composed of the peak areas of more than one (potentially unknown) chemical species. This suggests a greater variety of sesquiterpenes (which produce similar mass spectral fragmentation patterns) to be present in the sample pool than tentatively identified. The regiochemical and stereochemical variety of sesquiterpene species is attributable to the many possible combinations of functional group arrangements along the characteristic C_{15} skeletal backbone [7]. Although this limits certainty in the identities of the selected putative markers, these findings nevertheless suggest the potential significance of this class of compounds as markers of taxonomic significance for the genera *Plectranthus* and *Coleus*.

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

The retention times for some of the compounds with high variable scores, including furan-2-ethyl, 1-hexene, 2-hexenal and hexanal, fall outside of the retention time range of the reference n-alkane series, and thus are reported to have RI values <800.

Table 4: Retention indices (RI) of selected top predictor compounds of genus *Plectranthus* and *Coleus*. 2D RI values in red are for samples run after maintenance of the secondary GC column (c.f.: Chapter 4.5). R^2 Kovats linear plot = 0.9965.

Tentative identification	CAS number	Molecular formula	Chemical class	MW (g/mol)	1D time (min)	2D time (s)	1D RI Kovats (nonpolar)	1D RI Regression (nonpolar)	RI (lit.) nonpolar (NIST)	MS similarity match
α-Cubebene	17699-14-8	C15H24	Sesquiterpene	204	1100	0,98 0,95	1294	1408†	1366; 1351	753-916
β-Cubebene	13744-15-5	C15H24	Sesquiterpene	204	1154	1,09 0,45	1415†	1463	1384; 1381	755-893
β-Ylangene	20479-06-5	C15H24	Sesquiterpene	204	1200	0,96 0,955	1435†	1484	1425; 1418	759-895
					1430	1,49	1633†	1658	1598	777-903
β-Copaene	374189 *NIST	C15H24	Sesquiterpene	204	1219	1,29 0,625	1443	1498		
α-Cadinene	24406-05-1	C15H24	Sesquiterpene	204	1273	1,17 0,39	1465	1539†	1522; 1534	794-922
γ-Cadinene	39029-41-9	C15H24	Sesquiterpene	204	1246	1,18 0,39	1454	1519†	1505, 1507	931-941
α-Selinene	473-13-2	C15H24	Sesquiterpene	204	1216	0,365 1,15	1442	1496†	1523; 1500	754-930
Isogermacrene D	317819-80-0	C15H24	Sesquiterpene	204	1222	1,2 0,34	1444†	1500	1431; 1442	752-916
(Z,E)-α-Farnesene	26560-14-5	C15H24	Sesquiterpene	204	1212	0,8 0,98	1440†	1493†	1486; 1477	763-929
β-Calacorene	50277-34-4	C15H20	Bicyclic sesquiterpene	204	1276	0,905	1466	1541†	1548; 1543	812-893
trans-β-Ionone	79-77-6	C13H20O	Ionone	192	1194	1,385	1432†	1479†	1463; 1462	762-859
Bornyl acetate	76-49-3	C12H20O2	Terpene derivative	196	948	1,21 0,52	1237†	1292†	1269; 1270	785-917
2(5H)-Furanone, 5-ethyl-	2407-43-4	C6H8O2	Unsaturated lactone	112	490	0,815	883	1019†	984; 963	753-877
2,4-Hexadienal, (E,E)-	142-83-6	C6H8O	Unsaturated aldehyde	96	388	0,82 1,33	850†	870†	877; 877	770-932
3-Hexen-1-ol, (Z)-	928-96-1	C6H12O	Unsaturated alcohol	100	318	0,09 0,97	824†	818†	872; 838	753-952
3-Hexen-1-ol, acetate, (Z)-	3681-71-8	C8H14O2	Unsaturated ester	142	534	1,06 0,2	895	981†	987; 981	816-954
1,2,4,4-tetramethylcyclopentene	65378-76-9	C9H16	Cyclic alkene	124	340	0,695 1,42	832†	834†	857,6; 856,5	794-872
Cyclohexanone, 2,2,6-trimethyl-	2408-37-9	C9H16	Cyclic ketone	140	567	1,19	900	998†	1013; 1008	827-857
l-4-Oxohex-2-enal	374042 *NIST	C6H8O2	Unsaturated ketoaldehyde	112	540	0,31 0,92	896	985†	958; 950	751-880
Octane, 1,1'-oxybis-	629-82-3	C16H34O	Ether	242	1420	0,9 1,765	1628†	1650†	1657; 1660	869-927

Tentative identification	CAS number	Molecular formula	Chemical class	MW (g/mol)	1D time (min)	2D time (s)	1D RI Kovats (nonpolar)	1D RI Regression (nonpolar)	RI (lit.) nonpolar (NIST)	MS similarity match
Decanal	112-31-2	C10H20O	Aldehyde	156	826	1,1 1,1	1093	1202†	1183, 1184	824-832

†RI value falls within ±50 units of one and/or two of the literature values.

5.9) Relative abundance of top-ranking compounds

The species-wise normalised peak area values for the top compounds are plotted as a heat map in Figure 51, representing their relative abundance across the two genera. Most, including the sesquiterpenes and the C_6-C_9 molecules, are seen to be more abundantly distributed within the *Coleus* clade, with some exceptions. For example, sesquiterpenes such as α-cubebene, γ-cadinene, isogermacrene D, and the bicyclic sesquiterpene, β-calacorene, are also present in the *P. fruticosus*. This in line with the complexity observed in the sesquiterpene region of the TIC of *P. fruticosus* (Figure 25 and Figure 26), and is also consistent with the outlying dispersion of the points for *P. fruticosus* on the PCA score plot (Figure 40). In addition, the sesquiterpenes α-selinene and (Z,E)-α-farnesene, (Z)-3-hexen-1-ol as well as trans-β-ionone, are noticeably present in *P. zuluensis*, despite their overall distribution in genus *Coleus*. The compounds hexanal, 2-hexenal, 1,3-octadiene, 2-ethyl furan and 1,1'-oxybis-octane occur significantly in *P. ecklonii*, as well as across a number of *Coleus* species. In fact, hexanal and 2-ethyl furan are almost completely absent in all other *Plectranthus* samples.

In summary, the top-ranking compounds are more extensively distributed within genus *Coleus*, and are thus candidate markers of this clade.

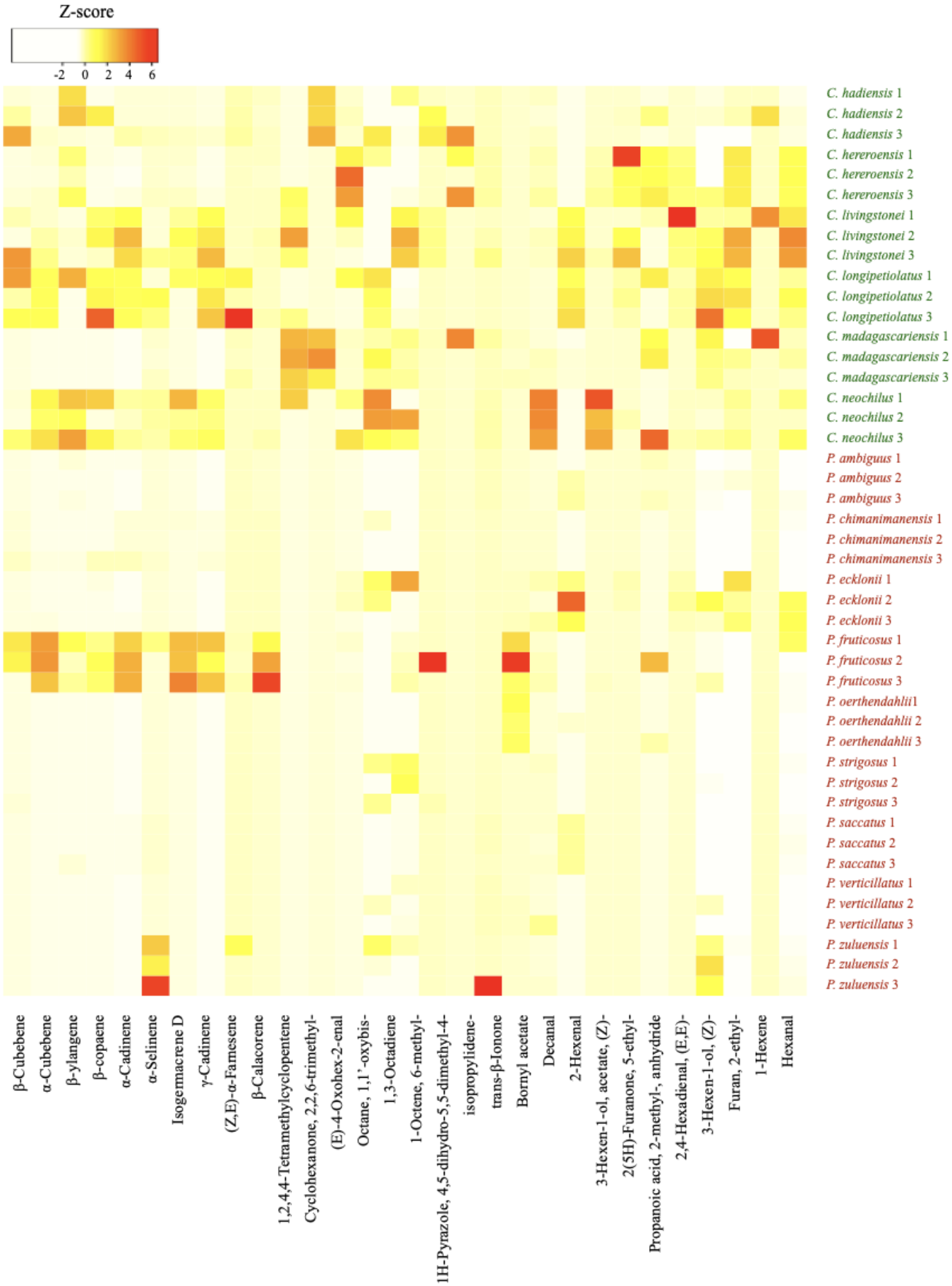


Figure 51: Heatmap of species-wise relative abundance of the compounds ranked as top variables by machine learning (c.f.: Table 3). Green = *Coleus*; red = *Plectranthus.* The colour intensity scale depicts the relative abundance Z-score.

5.10) Limitations of the study and future considerations

There are several limitations to the study. The first pertains to the moderate sample population size (N=45), which limits the reliability of the results[6]. The splitting of the dataset into training and testing sets requires a large overall sample size. Data splitting in this case can result in two or three of the replicates for one species ending up in only the training or testing sets, which can affect which variables are included in the final model, as well as the accuracy of the predictions during testing. Ideally, the sample size should be large enough that splitting will always result in a representative proportion of all included species in both splits. Fluctuations in predictive accuracy are exaggerated with smaller sample sizes, since changes in the random training-test splits can alter the composition of the splits, which in turn can affect the results.

Though the samples included in this study are broadly representative of the southern African species of both genera, population-level diversity is not captured, since only single individuals/specimens for each species were included (as triplicates) in the study.

A second limitation (discussed in Chapter 5.6) is the possibility of the models being overfit to the data. Overfitting is a general problem associated with the bias-variance trade-off (Chapter 2.2). In this case, even though the models were trained and tested on split sets, the models may still be overfit to the training data in the sense that the data from both sets were collected from the same specimens in the same round of data collection. In a future study, sample data could be collected from different specimens independently for the training and testing sets.

Another aspect to consider is the effect of changes in temperature and season on the emission of foliar VOCs, which has been reported for the sesquiterpenes [8]. All plant specimens included in the study were planted and maintained under the same ambient conditions, and sampling and analysis were conducted during a single summer period. Nevertheless, unaccounted-for variational factors could influence the results.

The last limitation considered here is the lack of certainty in the tentative identifications of the predictor compounds. As discussed previously (Chapter 5.8), there are several calculated RIs for those compounds selected as putative markers. These additional peaks, particularly in the case of the sesquiterpenes, likely correspond to different, potentially unidentified, isomeric species. Thus, the average peak areas for certain predictors may be composed of the summed

[6] Note, however, that this is a preliminary investigation, and thus a moderate sample population is suited to the scope of the study.

areas of peaks from different chemical species. This limitation could be overcome to a large extent through the use of certified reference standards, particularly of the sesquiterpene class. Calculation of the Ris for sesquiterpene standards, and examination of their experimental mass spectra, could aid in the identification of sample components, and in discriminating peaks of potentially unreported molecules.

Despite the above limitations, if there is greater chemical variety in the VOC profiles than is captured by the dataset, this is nevertheless consistent with the finding that genus *Coleus* appears to be distinguished by a more diverse expression of sesquiterpene and sesquiterpenoids than *Plectranthus*, with the exception of some species, such as *P. fruticosus*. Considering the variety of sesquiterpenes observed in this study, the phytochemical properties of genus *Coleus* may be worth investigating in future studies.

5.11) Conclusion

The complexity of foliar VOC profiles, obtained by GC×GC-TOFMS, poses a substantial challenge to comprehensive analysis. The full set of foliar VOCs from the leaves of the species of southern African *Plectranthus* and *Coleus*, after air- and soil-blank correction and removal of known contaminant peaks, consists of a high dimensional set of 1794 tentative compound identifications. The removal of near-zero variance predictors prior to machine learning reduces the dimensionality of the data set by 64.66%, and this pared-down dataset of 634 variables can be used as input to construct machine learning models.

The124hromatogramms of both species are characterised by peak clusters in two main regions of the separation space (at lower and higher 1D retention times, respectively) which correspond to the isomeric groups of the monoterpenes and sesquiterpenes. Species with comparatively complex chromatograms, such as *C. livingstonei*, *C. longipetiolatus*, *C. neochilus* and *P. fruticosus*, show a higher peak density, particularly in the region of the sesquiterpenes. Chromatograms from crushed-leaf samples show additional peaks to chromatograms from whole-leaf samples, which are likely to correspond to GLVs released upon damage of the foliar tissue.

PCA and LDA were performed on the full dataset as a preliminary assessment of the variation in the data, with both showing separable but overlapping clustering with respect to the two genera. Genus *Coleus* is characterised by variables of high loading scores for the first component of the PCA, suggesting the presence of compounds associated with this genus.

Three machine learning algorithms (an elastic-net regression, a random forest and a support-vector machine) were used to construct models of the training data and to make predictions on the testing data. Optimised models were selected by cross-validation. All three of the models tested show a high degree of accuracy (up to 90%) in the prediction of genus, with a sensitivity (for genus *Coleus*) of up to 100%.

The top-ranking variables listed by each of the models include C_6-C_{15} compounds of a variety of chemical classes. One key group are the sesquiterpenes (including β-ylangene, α-cubebene, β-copaene, α-cadinene and isogermacrene D), which are observed to be more abundantly and widely distributed overall (but not without exception) across those *Coleus* specimens included in the sample population, and are thus potential markers of genus *Coleus*. Another group of VOCs with high predictor scores are C_6 unsaturated compounds, likely to be GLVs, which are also observed to have a greater distribution in general across the *Coleus* samples.

A range of retention indices are found for the putative markers, which in the case of the sesquiterpenes suggests a greater isomeric variety than is captured in this analysis, as well as a rich phytochemistry for both genera. The use of analytical standards in a future study could aid in peak identification. Larger, independently sampled, training and testing sample populations would be required to corroborate these results, especially on a larger and more species-diverse population scale.

References

[1] Loreto, F., Barta, C., Brilli, F., Nogues, I. 2006. *On the induction of volatile organic compound emissions by plants as consequence of wounding or fluctuations of light and temperature*. Plant Cell Environ., 29(9): 1820-1828. https://doi.org/10.1111/j.1365-3040.2006.01561.x.

[2] Tholl, D., Boland, W., Hansel, A., Loreto, F., Röse. U.S.R., Schnitzler, J-P. 2006. *Practical approaches to plant volatile analysis*. Plant J., 45(4): 540-560. https://doi.org/10.1111/j.1365-313X.2005.02612.x.

[3] Brilli, F., Ruuskanen, T.M., Schnitzhofer, R., Müller, M., Breitenlechner, M., Bittner, V., Wohlfahrt, G., Loreto, F., Hansel, A. 2011. *Detection of plant volatiles after leaf wounding and darkening by proton transfer reaction "time-of-flight" mass spectrometry (PTR-TOF)*. PloS ONE, 6(5): e20419. https://doi.org/10.1371/journal.pone.0020419.

[4] TGSC information system. Accessed 11/05/2022. http://thegoodscentscompany.com.

[5] Kuhn, M. 2008. *Building predictive models in R using the caret package*. J. Stat. Softw., 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

[6] Kuhn, M. 2019. *The caret Package*. Available from: http://topepo.github.io/caret/index.html [Accessed: 23/11/2021].

[7] Chappell, J., Coates, R.M. 2010. 1.16-*Sesquiterpenes*. In: Hung-Wen, B.L., Mander, L. (eds.). Comprehensive natural products II, Chemistry and Biology (1). Elsevier, 2010: 609-641. https://doi.org/10.1016/0168-1176(95)04294-U.

[8] Duhl, T.R., Helmig, D., Guenther, A. 2008. *Sesquiterpene emissions from vegetation: a review*. Biogeosciences, 5(6): 761-777. https://doi.org/10.1016/0168-1176(95)04294-U.

Identifying cutaneous VOCs as predictive markers of malaria-status using comprehensive GC×GC-TOFMS and machine learning

Chapter 6: Cutaneous VOCs as potential markers of malaria-infection

Chapter 6A: Summary and background

6A.1) Summary

Cutaneous volatile organic compounds (VOCs) emitted from human skin may function as kairomones to hematophagous insects such as mosquitoes of the genus *Anopheles*, the vectors of the malaria-causing *Plasmodium* parasite. Malaria has been found to alter normal cutaneous VOC profiles [1-4], suggesting their potential application as markers of *Plasmodium* infection. The purpose of this study is to identify potential predictive markers of malaria-infection status in the form of cutaneous volatile organic compounds, from the epidermis of suspected patients, using GC×GC-TOFMS and machine learning.

In-house developed PDMS sampling loops, which have previously been demonstrated to be suitable for the purposes of biomarker identification [5-7] were used to extract cutaneous VOCs from the epidermis of patients visiting government clinics in the Vhembe district of Limpopo Province, South Africa, for diagnosis and possible treatment of malaria. The extracted compounds were analysed with GC×GC-TOFMS, and the total ion chromatographic data was used to construct VOC profiles of the participants. Machine learning was used to model the data, to make predictions on the malaria status of samples taken from an individual, and to tentatively identify potential putative markers of malaria-infection.

6A.2) Background: malaria diagnostics

Malaria is a potentially fatal disease caused by infection with eukaryotic parasites of the genus *Plasmodium*. Though there are over 100 known *Plasmodium* species, only four are known to infect humans [8]. The World Health Organisation (WHO) estimated a global total of 229 million cases of the disease in 2019, of which 92% occurred in Africa. Mortalities were estimated at 409 000 worldwide, with Africa accounting for 51% of the total [9]. On the continent, a majority of cases are caused by *P. falciparum*, which is able to rapidly proliferate in the blood, and can cause severe anaemia and blockage of small blood vessels [8].

6A.2.1) The malaria parasite life-cycle: blood forms and gametocytes

The lifecycle of the malaria parasite is complex, and its numerous stages require both a mosquito host and a human host. The stages that occur in human hepatic cells and erythrocytes are haploid and asexual, whereas the diploid zygote occurs in the mosquito. A human host may carry not only the asexual, or blood-form parasites, but also sexual gametocytes [8]. Microscopic examination of a thin film peripheral blood smear from a malaria-infected patient can distinguish asexual-stage parasite forms from sexual-stage gametocytes on the basis of visible morphological differences. The sexual stage is of important epidemiological and clinical significance since the continued infection of more human hosts relies on the transmission of the gametocytes to the mosquito. The accurate diagnosis, followed by appropriate and prompt treatment, of the gametocyte phase of malaria infection is thus essential to controlling the spread of the disease.

6A.2.2) Malaria diagnostics

Diagnosis of malaria using analytical techniques generally involves microscope examination of a blood smear or detection of a biochemical marker of infection. Currently, the two most routine procedures employed to diagnose malaria are Giemsa microscopy and the Rapid Diagnostic Test (RDT).

Giemsa, Wright and Field microscopy are histopathological staining techniques that permit parasite quantification and species identification from a peripheral blood smear [10, 11]. Microscopy is inexpensive, and currently remains the gold standard for clinical drug and vaccine trials, epidemiological studies, and assessing diagnostic techniques. However, use of microscopy in a routine diagnostic environment is hampered by numerous problems and disadvantages. Proficient and correct microscope analysis depends on a number of conditions, including technical training and competence; appropriate and correct slide preparation; routine instrument maintenance; resource quality and availability; and adequate quality assurance and control. One or more of these conditions are often not met in economically disadvantaged regions lacking infrastructure and skilled technicians. Diagnostic errors such as false-positives and -negatives, species misidentification and misestimation of parasite density are more frequent in these regions. In such cases, limits of detection may be higher than 50-100

parasites μL^{-1}, which increases the chance of false negatives [10-12], particularly at low parasite density [13].

The rapid diagnostic test (RDT) is an immunochromatographic assay that involves parasite antigen recognition by labelled monoclonal antibodies. A typical RDT kit consists of a strip of nitrocellulose imprinted with immobile antibody [14]. The three most commonly used antigens for RDT are histidine-rich protein 2 (HRP-2), parasite lactate dehydrogenase (pLDH), and plasmodial aldolase. HRP-2 is specific to *P. falciparum* and is used in more than 90% of RDTs, whereas pLDH and aldolase, which are metabolites of the parasite glycolytic pathway, can be used to detect more than one parasite species. Dual HRP-2/pLDH kits are capable of differentiating between *P. falciparum* and non-*P. falciparum* antigen [10, 11, 14].

Though microscopy remains the gold standard, the RDT has a number of advantages over it [10, 11, 15]: it is simple, requires no special expertise, is not as resource-intensive, and diagnoses can be performed in remote locations lacking infrastructure and resources. It is also more applicable in zones of outbreak or occupational risk, or where microscope diagnostics are not available. The test can be species-specific or -nonspecific, depending on the antibody or combination thereof employed, which lessens the risk of species misidentification. Greater than 95% sensitivity is attainable for most RDTs specific to *P. falciparum*. Sensitivity can be as high as 100% for a parasite density of 500 parasites μL^{-1} but decreases with lower density [10]. Nevertheless, the RDT has its disadvantages. Their conditional performance range encompasses ambient temperatures of 4-30°C and humidity levels less than 70%, conditions which do not prevail in the warm and humid tropics. Additionally, the dependence of the sensitivity on parasite density increases the chances of a false negative at lower parasite densities [10].

The polymerase chain reaction (PCR) Is the most sensitive and specific of the diagnostic techniques, capable of low parasite density detection, and can be used for confirmation in cases where microscopy yields indefinite results [10, 16, 17]. However, despite its robustness and reliability, PCR has, for two main reasons, limited applicability in routine clinical work. Firstly, the method requires costly equipment, reagents and expertise, and is thus impractical in most endemic regions where adequate resources are lacking. Secondly, results are ordinarily only available days or weeks after sampling due to the fact that the sample has first to be stored and transported to the laboratory before the assay can be performed.

For a method to be suitable for routine clinical diagnosis in regions that lack funds, infrastructure and resources, it must be not only sufficiently reliable, in terms of sensitivity and specificity, but also feasible. Though microscopy and the RDT remain the primary and most

widely used tests, the quality of the results they give depend on a number of contingencies which are not ordinarily in place in such areas. PCR has demonstrated superior sensitivity, specificity and reliability compared to microscopy and the RDT, but its routine clinical application is precluded by a high resource demand. This study intends to address the need of a robust and practicable diagnostic technique by investigating the potential of human skin VOCs as markers of malarial parasitaemia.

References

[1] Lacroix, R., Mukabana W.R., Gouagna, L.C., Koella, J.C. 2005. *Malaria infection increases attractiveness of humans to mosquitoes*. PloS Biol, 3(9): e298. https://doi.org/10.1371/journal.pbio.0030298.

[2] Busula A.O., Bousema, T., Mweresa, C.K., Masiga, D., Logan, J.G., Sauerwein, R.W., Verhulst, N.O., Takken, W., de Boer, J.G. 2017. *Gametocytemia and attractiveness of Plasmodium falciparum-infected Kenyan children to Anopheles gambiae mosquitoes*. The Journal of Infectious Diseases, 216(3): 291-295. https://doi.org/10.1093/infdis/jix214

[3] Kelly, M., Su, C-Y., Schaber, C., Crowley, J.R., Hsu, F-F., Carlson, J.R., Odom, A.R. 2015. *Malaria parasites produce volatile mosquito attractants*. mBio, 6(2): e00235-15. https://doi.org/10.1128/mBio.00235-15.

[4] De Moraes, C.M., Stanczyk, N.M., Betz, H.S., Pulido, H., Sim, D.G., Read, A.F., Mescher, M.C. 2014. *Malaria-induced changes in host odors enhance mosquito attraction*. PNAS, 111(30): 11079-11084. https://doi.org/10.1073/pnas.1405617111.

[5] Roodt, A. P., Naudé, Y., Rohwer, E. 2018. *Human skin volatiles: passive sampling and GC×GC-ToFMS analysis as a tool to investigate the skin microbiome and interactions with anthropophilic mosquito disease vectors.* Journal of Chromatography B, 1097-1098: 83-89. https://doi.org/10.1016/j.jchromb.2018.09.002.

[6] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Non-invasive sorptive extraction for the separation of human skin surface chemicals using comprehensive gas chromatography coupled to time-of-flight mass spectrometry: a mosquito-host biting site investigation*. J. Sep. Sci., 43(22): 4202-4215. https://doi.org/10.1002/jssc.202000522.

[7] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Chemical profiling of the human skin surface for malaria vector control via a non-invasive sorptive sampler with GC×GC-*

TOFMS. Anal. Bioanal. Chem., 412(23): 5759-5777. https://doi.org/10.1007/s00216-020-02799-y.

[8] Centres for Disease Control and Prevention, 2018. *Malaria*. https://www.cdc.gov/malaria/about/biology/#tabs-1-6. Accessed 02/05/2019.

[9] WHO. 2020. *World Malaria Report 2020*. https://apps.who.int/iris/rest/bitstreams/1321872/retrieve.

[10] Wongsrichanalai C., Barcus, M.J., Muth, S., Sutamihardja, A., Wernsdorfer, W.H. 2007. *A review of malaria diagnostic tools: microscopy and rapid diagnostic test (RDT)*. Am. J. Trop. Med. Hyg., 77(6): 119-127. https://www.ncbi.nlm.nih.gov/books/NBK1695/.

[11] Tangpukdee, N., Duangdee, C., Wilairatana, P., Krudsood, S. 2009. *Malaria diagnosis: a brief review*. Korean J Parasitol, 47(2): 93-102. https://doi.org/10.3347%2Fkjp.2009.47.2.93.

[12] Ohrt, C., Purnomo, M., Sutamihardja, A., Tang, D., Kain, K.C. 2002. *Impact of microscopy error on estimates of protective efficacy in malaria-prevention trials*. The Journal of Infectious Diseases, 186: 540-546. https://doi.org/10.1086/341938.

[13] McKenzie, F.E., Sirichaisinthop, J., Miller, R.S., Gasser R.A., Wongsrichanalai, C. 2003. *Dependence of malaria detection and species diagnosis by microscopy on parasite density*. Am. J. Trop. Med. Hyg., 69(4): 372-376. https://doi.org/10.4269/ajtmh.2003.69.372.

[14] WHO. 2015. *How malaria RDTs work*. https://www.who.int/malaria/areas/diagnosis/rapid-diagnostic-tests/about-rdt/en/. Accessed 11/02/2019.

[15] Mouatcho, J., Goldring, J.P.D. 2013. *Malaria rapid diagnostic tests: challenges and prospects*. Journal of Medical Microbiology, 62: 1491-1505. https://doi.org/10.1099/jmm.0.052506-0.

[16] Johnston, S.P., Pieniazek, N.J., Xayavong, M.V., Slemenda, S.B., Wilkins, P.P., da Silva, A.J. 2006. *PCR as a confirmatory technique for laboratory diagnosis of malaria*. Journal of Clinical Microbiology, 44(3): 1087-1089. https://doi.org/10.1128/jcm.44.3.1087-1089.2006.

[17] Hänscheid, T., Grobusch, M.P. 2002. *How useful is PCR in the diagnosis of malaria?* Trends in Parasitology, 18(9): 395-398. https://doi.org/10.1016/s1471-4922(02)02348-6.

Chapter 6B: Cutaneous VOCs and their relation to vector attraction and malaria

6B.1) VOCs as diagnostic markers of disease

As discussed in Chapter 1.1.2 and Chapter 1.1.3, biogenic VOCs are produced as secondary metabolites by living organisms, and as such constitute a portion of the metabolome. Such volatiles can provide important information on underlying biochemical processes, and alterations in normal patterns of metabolite production can be indicative of an altered biological state, such as pathology and disease. Certain diseases have long been associated with characteristic, often unpleasant, odours, and the olfaction of patients and bodily fluids has long been utilised by physicians and healthcare workers to aid in the diagnosis of disease [1-3]. Modern analytical techniques have led to the discovery of a large number of volatile biomarkers of different diseases and disorders, including infectious diseases (pneumonia, tuberculosis, cholera, smallpox, diptheria, typhoid and yellow fever), hepatic, gastrointestinal, and cardiovascular diseases, cancer, and genetically inherited metabolic disorders, such as diabetes and uraemia [1-5].

Pathology-induced alterations of VOC profiles can result in the presence of previously absent compounds or the absence of previously present compounds, or it can result in changes in the ratios, or relative abundances, of normally occurring molecules [1]. VOCs associated with pathology may originate from several sources: 1) from endogenous processes initiated in response to the pathological state, such as oxidative stress and inflammation; 2) from endogenous processes altered under the influence of the pathogen; 3) from exogenous processes of the pathogen itself [1-3].

VOCs are present in most biological matrices: blood, saliva, sputum, urine, breath and tissue surfaces, such as the epidermis of the skin [1, 2]. The virtual ubiquity of VOCs in the organism, and the fact that they occur in both fluids and gaseous mixtures that are excreted or emitted from the body into the environment, means that they can be sampled in a non-invasive fashion. This ease of sampling is well suited, not only to biomarker investigations, but to diagnostic applications in the clinic or in the field. Breath analysis, in particular, has shown much promise in the search for volatile biomarkers of disease [1-8].

6B.2) The origin and variety of cutaneous VOCs

Human skin is a site of continuous emission of cutaneous volatile organic compounds. These cutaneous VOCs are the metabolic products of endosymbiont microbiota, derived from the precursors in the secretions of the apocrine glands interspersed across the surface area of the skin [9, 10].

Cutaneous VOCs are the chemical basis of perceived body odour, the latter is largely determined by the resident microbiome. For example, VOCs associated with axillary malodour have been linked to the presence of four main bacterial groups (staphylococci, micrococci, propionibacteria and coryneforms), as well as yeast of the genus *Malassezia* [11]. Examples of identified VOC odorants produced by axilla isolates of *Corynebacteria* sp. Include the carboxylic acids 3-methyl-2-hexenoic acid (3MH2) and 3-hydroxy-3-methylhexanoic acid (HMHA), which are the products of glutamine conjugate precursors converted by the aminoacylase enzyme, N_{α}-acyl-glutamine aminoacylase (N-AGA) [12, 13]. Staphylococci and propionibacteria ferment lactic acid and glycerol to malodorous short-chain (C_2-C_3) volatile fatty acids (VFAs), as well as ethanoic and propionic acid [14].

Much variety, within and between individuals, is observed in the human skin microbiome. Variety within the individual refers to differences in the microbiome from one region of the skin to another. For example, on a single individual, over 150 bacterial species can be found on the palms of the hands [15], and the microbial species population composition between an individual's dominant and non-dominant hands can differ by up to 83% [16]. This so-called topographical variety is like due to the dependence of the microbiome composition on the local microenvironment of the region of the skin in question [17]. Between individuals, the species composition on the palm of the hand has been found to vary as much as 87%, with females showing larger average species diversity than males [16]. Inter-individual variation in the microbiome is distinct enough for the purposes of individual profile matching. For example, using high-throughput pyrosequencing, objects that have been handled by a subject can be linked with high certainty to that subject on the basis of the microbial composition recovered from the object, even up to two weeks after it has been handled [18].

These findings, though narrowly representative, demonstrate the inherent variability of microbiomic profiles in the human population, and suggest the fact that such variability is reflected in the volatilome.

Indeed, within-individual variation, in terms of the types and quantities of VOCs present, has been observed for different regions of the skin [10]; and reproducible differences in the VOC profiles, obtained from axillary sweat, between females and males, and between individuals, have been found and linked to 44 individual-specific, and 12 sex-specific, compounds, using GC-MS pattern recognition techniques [19]. Such inter-individual variation has permitted up to 99% distinguishability between individuals using Spearman rank correlation comparison [20, 21]. In addition, putative age-related differences have been observed for a few compounds, including sulphur VOCs, in individuals aged 41-79 [10].

Given that cutaneous VOCs emitted at any given moment are likely to originate from diverse microbial metabolic pathways, the volatilomic profiles of individuals or populations are likely to be complex and variable. In addition, such volatilomic profiles are likely to be dependent not only on genetic and metabolic factors (of both the host organism and its microbiome) but on contingent environmental factors which may, directly or indirectly, affect gene expression, or which may influence VOC production (such as diet, health status, stress and the use of medication). Though the microbial origin of cutaneous VOCs is not disputed, the many factors that influence their production could problematise attempts at correlating VOC profiles to corresponding microbiomes [22]. In addition, observed VOC profile correlations for a particular designation, such as age, could be due to factors not directly related to that designation, but to one or more other common factors [10].

6B.3) Cutaneous VOCs as kairomones of hematophagous Anopheline vectors of malaria

The vector of *Plasmodium* parasites, the causative agents of human malaria, are female hematophagous mosquitoes of the genus *Anopheles*. In order for the mosquito to feed and transmit the parasite, it must first locate and select a vertebrate host, and land on a suitable area of host skin. This is achieved through a complex process that involves the multimodal integration of different physical cues and chemical signals [23, 24]. In the generalised model of the host-targeting of *Aedes aegypti*, the female is alerted to the presence of a potential host by the detection of local above-ambient fluctuations in carbon dioxide levels arising from exhalation, and flies upwind of the carbon dioxide plume by optomotor anemotaxis, aided by visual stimuli [23]. Within a close radius of the host, humid thermal currents, arising from the epidermal surface of the host, and emissions of kairomones, in the form of cutaneous VOCs,

influence the behaviour of the mosquito, and function not only to orientate it into landing, but also appear to determine the relative attractiveness of the host to the mosquito, and thus whether it will inevitably select its host or not [24].

The importance of skin volatile compounds in host mosquito-attraction has been demonstrated in dual-port olfactometer assays [25-28]. Cutaneous residuum remaining on glass beads after handling by human hands has been found to attract *A. aegypti* females [25] and *A. gambiae* has been found to be attracted to sweat collected from subjects performing physical exercise in warm and humid conditions. Notably, freshly collected sweat, which is odourless, has been found to be less attractive to *A. gambiae* than incubated sweat [26], demonstrating the importance of microbial metabolism during incubation in the formation of volatile attractants.

Differences in the relative attractiveness of different individuals to mosquitoes (commonly recognised by inhabitants of mosquito-endemic regions) has been experimentally demonstrated in field-trials, and have been shown furthermore to be consistent over time [29, 30]. Such variation in host attractiveness is likely based on between-individual variability in cutaneous VOC profiles (Chapter 6B.2). Studies aimed at identifying putative kairomones generally involve the identification of compounds that elicit behavioural responses in behavioural dual-choice olfactometer or wind tunnel assays [31-44], or elicit neural receptor electrophysiological response in electroanntenographic (EAG) [41, 43-45] experiments or single-sensillum recordings (SSR) [46, 47]. EAG and SSR can identify potential kairomones sensed by mosquitoes, but cannot determine their behavioural effects. To investigate the behavioural effect of a kairomone, EAG data is used in conjunction with behavioural assays [31-41, 45]. In many cases, gas chromatography-mass spectrometry (GC-MS), or gas chromatography coupled to electroantennography (GC-EAG), is used in tandem with electrophysiological techniques for the identification of electrophysiologically-active compounds [41-45, 47].

L-lactic acid and ammonia were two of the first cutaneous volatile compounds identified as kairomones used by hematophagous insects in host location and selection [31-34]. A correlation between a wide range of concentrations of L-lactic acid and increased attractiveness to *A. aegypti* has been observed [31, 32], and ammonia is active in *A. gambiae* at fractional millimolar concentrations [33]. Both compounds occur in human sweat; the pH of incubated sweat samples increases, indicating bacterial production of ammonia, and freshly collected sweat, which does not show high mosquito attraction, contains ammonia levels below the attraction threshold [33]. However, the activity of L-lactic acid and ammonia is synergistic, and depends on their concentration relative to the concentration of other components present in a mixture [34-37], as well as the species of mosquito in question [33]. Notably, a large

number of compounds structurally related to L-lactic acid, present on human skin in sweat, affect the electrophysiology and/or behaviour of anthropophilic Anopheline vectors of malaria, either by themselves, but mostly in a mixture with ammonia and/or L-lactic acid, and/or other carboxylic acids. Examples include, but are not limited to: C_2-C_5 α-substituted aliphatic carboxylic acids, and esters of α-hydroxy-β-phenyl aliphatic acids [38, 39]; short to medium chain (C_1-C_{14}) saturated and unsaturated carboxylic acids [40, 41, 45]; and medium to long chain unsaturated alcohols such as 1-octen-3-ol [45, 46] 4-hexen-1-ol and 1-hepten-3-ol [42].

Mixtures of kairomone VOCs can also result in a masking effect, whereby the apparent attractive or repellent effects of a compound are negated in the presence of other components. In this regard, two human-specific unsaturated carboxylic acids that occur prevalently in axillary sweat— (E/Z)-3-methyl-2-hexenoic acid (3M2H) and 7-octenoic acid —show dose-dependent selective EAG-activity in *A. gambiae* at near-microgram thresholds [47]. Of these compounds, only 7-octenoic acid is more attractive to *A. gambiae* females compared to a blend of carbon dioxide and "human odour", whereas a mixture of 3M2H and 7-octenoic acid appears to mask the attractive effect of carbon dioxide and "human odour" [47]. Thus, despite the fact that antennal response demonstrates the bioactivity of these VOCs, their influence on the behavioural response is not necessarily straightforward, and may depend on the presence of other compounds. This is also seen for human-derived VOCs found to decrease host attractiveness to mosquitoes of *A. aegypti* [48]. A variety of compounds are found in significantly greater quantities on the hands of individuals less attractive to *A. aegypti*: benzaldehyde, 6-methyl-5-hepten-2-one (6M52H), octanal, nonanal, naphthalene, decanal and geranylacetone. Of these, five compounds 6M52H, octanal, nonanal, decanal and geranylacetone, reduce upwind anemotaxis and show a lower relative attractiveness in behavioural assays when added to a "standard hand" representative of attractive human odour [48].

Much about the role of different VOCs in mosquito attraction and behaviour remains to be elucidated, but the semiochemical activity of a suspected kairomone seems to depend on the presence of other compounds, and that the relative attractiveness and repellence of an individual is determined not by the presence or absence of any one particular kairomone, but on the relative abundances of numerous VOCs present together.

6B.4) Malaria-induced changes in cutaneous VOC profiles

Research in the past two decades indicates that *Plasmodium* infection may enhance the attractiveness of malarial hosts to naïve candidate Anopheline vectors. The effect has been observed for *Plasmodium* cultures in dual-choice olfactometric assays [49], and in avian (for *P. relictum* and *C. pipiens*) [50, 51], rodent (for *P. chabaudii*) [52, 53] and human (for *P. falciparum* and *P. vivax*) [54, 55] models, although two studies have observed the opposite effect [56, 57]. In most cases, infection-associated enhanced vector attraction has been observed only for hosts carrying microscopic-level gametocytes in the blood [54, 55, 58, 59], however it has also been observed for individuals harbouring the asexual erythrocytic stage of the parasite [59]. The effect is lost after treatment and clearance of the parasite [53-55, 58, 59] and the attractiveness of treated subjects who have previously harboured gametocytes appears to drop below that of healthy subjects and treated erythrocyte-stage subjects, which may be due to the non-viability of the erythrocytes of hosts rendered anaemic by infection [54, 59].

The volatile compounds giving rise to enhanced vector attraction associated with *Plasmodium* infection could be produced directly by the parasite, or they could arise from alterations in normal microbial populations [60]. Another possibility is that the altered VOC emissions are associated with a general pathophysiological state induced by disease conditions, such as oxidative stress [55].

The chemical profiles of *Plasmodium*-infected humans exhibiting enhanced attractiveness to Anopheline vectors have only recently been investigated. A group of EAG-active aliphatic aldehydes present in significantly greater quantities in the samples of infected individuals have been identified, including heptanal, octanal, I-2-octenal, nonanal and I-2-decanal [55]. Of these, heptanal increases the attractiveness of parasite-free body odour, but not of a standard synthetic blend, whereas a mixture of aldehydes increases the attractiveness of a synthetic blend [55]. More recently, analysis of skin VOC data using genetic algorithms have identified specific VOCs markers that distinguish malaria status from disease statuses of similar symptoms and pathologies [61]. These putative markers include aldehydes such as hexanal and octanal, the substituted ketone 4-hydroxy-4-methylpentan-2-one, two alkanes (octanal and decanal) and *o*-xylene [61].

6B.5) The sampling and analysis of cutaneous VOCs

Figure 52 summarises different approaches to the sampling of cutaneous VOCs from the epidermis. Sample analytes can either be extracted, or pre-concentrated, prior to analysis, or they can be introduced directly from the skin into an analytical instrument, without prior extraction, for real-time and online analysis. The former approach is more common, and is suited to both targeted and non-targeted analysis using GC-MS [9, 10, 12, 15, 20-22, 42-45, 48, 55, 62-75] and GC×GC-TOFMS [76-78]. Direct-introduction techniques do not include a step for chromatographic separation, and are suited to targeted analysis using chemical ionisation methods [79-83].

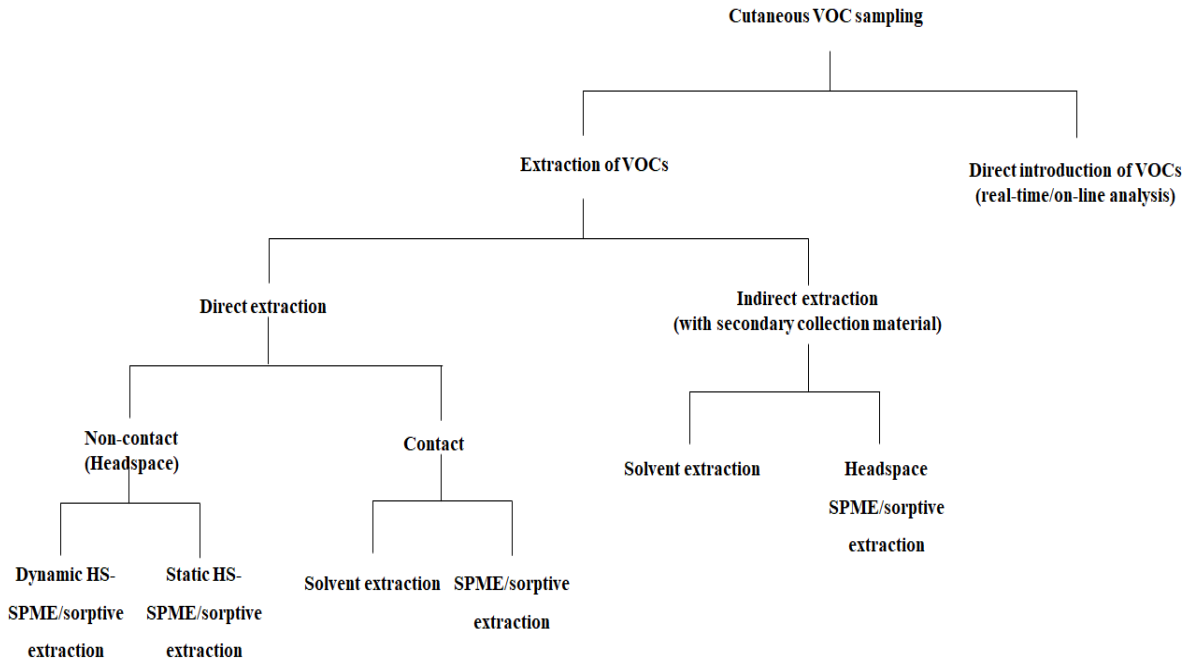


Figure 52: Classification scheme (derived by author) of methods for the sampling of cutaneous VOCs.

6B.5.1) Extraction prior to analysis

Cutaneous VOCs can be obtained either from skin (or sweat) of a subject (direct sampling), or from a material that has been in contact with the skin of the subject (indirect sampling) (Figure 52). The method of extraction can be either solvent-based, but is more prevalently done using solid-phase microextraction (SPME), or sorptive extraction (Chapter 1.2). Solvent can also be used for direct sampling, but requires that the solvent come into contact with the skin [62].

Sorptive samplers are typically composed of nonpolar and porous organic polymers, such as polydimethylsiloxane (PDMS), alone or in combination with Carboxen (CAR) and/or divinylbenzene (DVB) [10, 22, 45, 62-69, 76-78]. The advantage of using polymeric sorbents is that they exclude higher molecular weight species, so that there are fewer potential matrix interferents present in the sample. For this reason, sorptive extraction is a more selective technique for the sampling of VOCs. In addition, the preparation and clean-up stages involved in solvent-based methods are obviated.

The sampler may take any form suitable to being brought either into contact with (and adhering to) a region of the epidermis (membrane/film, stir bar, loop) or being suspended in the headspace above the epidermal surface [10, 15, 22, 63-65, 76-78]. The headspace conditions can be either static (under equilibrium conditions) or dynamic (under gas-flow conditions) [10, 45, 64-66]. Headspace sampling can be achieved by enclosing a portion of the body to be sampled (for example, the foot), or even the whole body (except for the head and neck) inside a nalophan bag, through which filtered air is drawn towards a sorbent housed in a tube [48, 55, 64].

In indirect sampling, cutaneous VOCs are first collected onto a secondary collection material that has been in contact with the epidermis (either via direct attachment or handling). Such a material may be a natural or synthetic cellulose-based textile such as cotton, challis, polyester, rayon or wool [12, 20, 21, 67-74]; or it may be material with an adsorbent surface, such as a glass bead [42-44, 75] or a steel razor blade [64]. Alternatively, compounds are collected into a nalophan bag [21]. The analytes are then extracted from the secondary material using solvent extraction [12, 64, 71-75] or headspace sorptive extraction [12, 20, 21, 68-72]. In cases where glass beads are used to collect VOCs via handling, the beads can be placed inside a GC injection port [42-44] for direct desorption. The disadvantages of indirect sampling are that the extra step of analyte transferal from the secondary material to the sorbent increases the chances of contamination, and that the material may overly retain a portion of analytes during transferal [9]. Indirect sampling has potential applicability to forensic work, where volatile analytes cannot be obtained from a subject directly, but must be collected from objects and materials that the subject has touched or worn [20, 21, 69].

6B.5.2) Real-time and online analysis of cutaneous VOCs

In direct introduction techniques, cutaneous analytes are not extracted prior to analysis, nor chromatographically separated, but instead are directly introduced into a mass spectrometer for online real-time analysis [79-83]. Examples include chemical ionisation methods such as secondary-electrospray ionisation (SESI-MS) [80, 81], selected-ion flow-tube (SIFT-MS) [79], proton-transfer-reaction (PTR-MS) and selective-reagent-ionisation mass spectrometry (SRI-MS) [83]. A technique closely related to MS, ion mobility spectrometry (IMS), has also been used for the analysis of VOCs in "near *real time*", but involves hyphenation to a multi-capillary column [82].

In real-time applications, the inlet leading to the ion source is connected by a tube to the epidermal headspace, either of the entire body at rest in a conditioned chamber [83], or a particular region of the body, such as the hand, enclosed in a nalophan bag [79]. Alternatively, a part of the body, such as the hand, can be held at the entrance of the inlet [80-82]. If a portion of the body is enclosed in a bag, the headspace can be allowed to equilibrate for a few minutes before introduction into the inlet [79]. If the whole-body headspace, or the ambient air around the skin, is to be sampled, air Is continuously pumped into the inlet [80-83].

Since real-time methods do not require a sampling step prior to analysis, they are quick and simple to perform. In addition, limits of detection and quantification as low as a few fractional parts per trillion are possible [82, 83]. However, these methods are limited to targeted analysis of known VOCs, and thus not suitable for nontargeted analysis of complex samples.

References

[1] Shirasu, M., Touhara, K. 2011. *The scent of disease: volatile organic compounds of the human body related to disease and disorder*. J. Biochem., 150(3): 257-266. https://doi.org/10.1093/jb/mvr090.

[2] Buljubasic, F., Buchbauer, G. 2014. *The scent of human diseases: a review on specific volatile organic compounds as diagnostic markers*. Flavour Frag. J., 30(1): 5-25. https://doi.org/10.1002/ffj.3219.

[3] Sethi, S., Nanda, R., Chakraborty, T. 2013. *Clinical applications of volatile organic compound analysis for detecting infectious diseases*. Clin. Microbiol. Rev., 26(3): 462-475. https://doi.org/10.1128/cmr.00020-13.

[4] Schmidt, K., Podmore, I. 2015. *Current challenges in volatile organic compound analysis as potential biomarkers of cancer*. J. Biomark., 2015: 981458. https://doi.org/10.1155%2F2015%2F981458.

[5] Probert, C.S.J., Ahmed, I., Khalid, T., Johnson, E., Smith, S., Ratcliffe, N. 2009. *Volatile organic compounds as diagnostic biomarkers in gastrointestinal and liver diseases*. J. Gastrointestin. Liver Dis., 18(3): 337-343. DOI not available.

[6] van de Kant, K.D.G., van der Sande, L.J.T.M., Jöbsis, Q., van Schayk, O.C.P., Dompeling, E. 2012. *Clinical use of exhaled organic compounds in pulmonary diseases: a systematic review*. Respir. Res., 13(1): 117-140. https://doi.org/10.1186/1465-9921-13-117.

[7] Wilson, A.D. 2015. *Advances in electronic-nose technologies for the detection of volatile biomarker metabolites in the human breath*. Metabolites, 5(1): 140-163. https://doi.org/10.3390/metabo5010140.

[8] Horváth, Z., Lázár, N., Gyulai, N., Kollai, M., Losonczy, G. 2009. *Exhaled biomarkers in lung cancer*. Eur. Respir. J., 34(1): 261-275. https://doi.org/10.1183/09031936.00142508/.

[9] Dormont, L., Bessière, J-M., Cohuet, A. 2013. *Human skin volatiles: a review*. J. Chem. Ecol., 39: 569-578. https://doi.org/10.1007/s10886-013-0286-z.

[10] Gallagher, M., Wysocki, C.J., Leyden, J.J., Spielman, A.I., Sun, X., Preti, G. 2008. *Analyses of volatile organic compounds from human skin*. BJD, 159(4): 780-791. https://doi.org/10.1111/j.1365-2133.2008.08748.x.

[11] Taylor, D., Daulby, A., Grimshaw, S., James, G., Mercer, J., Vaziri, S. 2003. *Characterization of the microflora of the human axilla*. Int. J. Cosmet. Sci., 25: 137-145. https://doi.org/10.1046/j.1467-2494.2003.00181.x.

[12] Natsch, A., Derrer, S., Flachsmann, F., Schmid, J. 2006. *A broad diversity of volatile carboxylic acids, released by a bacterial aminoacylase from axilla secretions, as candidate molecules for the determination of human-body odor type*. Chem. Biodivers, 3: 1-20. https://doi.org/10.1002/cbdv.200690015.

[13] Natsch, A., Gfeller, H., Gygax, P., Schmid, J. 2005. *Isolation of a bacterial enzyme releasing axillary malodour and its use as a screening target for novel deodorant formulations*. Int. J. Cosmet. Sci., 25: 115-122. https://doi.org/10.1111/j.1467-2494.2004.00255.x.

[14] James, A.G., Hyliands, D., Johnston, H. 2004. *Generation of volatile fatty acids by axillary bacteria*. Int. J. Cosmet. Sci., 26: 149-156. https://doi.org/10.1111/j.1467-2494.2004.00214.x.

[15] Penn, D.J., Oberzaucher, E., Grammer K., Fischer, G., Soini, H.A., Wiesler, D., Novotny, M.V., Dixon, S.J., Xu, Y., Brereton, R.G. 2007. *Individual and gender fingerprints in human body odour*. J. R. Soc. Interface, 4: 331-340. https://doi.org/10.1098/rsif.2006.0182.

[16] Fierer, N., Hamady, M., Lauber, C.L., Knight, R. 2008. *In the influence of sex, handedness, and washing on the diversity of hand surface bacteria*. PNAS, 105(46): 17994-17999. https://doi.org/10.1073/pnas.0807920105.

[17] Grice, E.A., Kong, H.H., Conlan, S., Deming, C.B., Davis, J., Young, A.C., NISC Comparative Sequencing Program, Bouffard, G.G., Blakesley, R.W., Murray, P.R., Green, E.D., Turner, M.L., Segre, J.A. *Topographical and temporal diversity of the human skin microbiome*. Science, 324(5931): 1190-1192. https:// doi.org/10.1126%2Fscience.1171700.

[18] Fierer, N., Lauber, C.L. Zhou, N., McDonald, D., Costello, E.K., Knight, R. 2010. *Forensic identification using skin bacterial communities*. PNAS, 107(14): 6477-6481. https://doi.org/10.1073/pnas.1000162107.

[19] Penn, D.J., Oberzaucher, E., Grammer K., Fischer, G., Soini, H.A., Wiesler, D., Novotny, M.V., Dixon, S.J., Xu, Y., Brereton, R.G. 2007. *Individual and gender fingerprints in human body odour*. J. R. Soc. Interface, 4: 331-340. https://doi.org/10.1098/rsif.2006.0182.

[20] Kusano, M., Mendez, E., Furton, K.G. 2013. *Comparison of the volatile organic compounds from different biological specimens for profiling potential*. J. Forensic Sci., 58(1): 29-39. https://doi.org/10.1111/j.1556-4029.2012.02215.x.

[21] Curran, A.M., Prada, P.A., Furton, K.G. 2010. *The differentiation of the volatile organic signatures of individuals through SPME-GC/MS of Characteristic Human Scent Compounds*. J. Forensic Sci., 55(1): 50-57. https://doi.org/10.1111/j.1556-4029.2009.01236.x.

[22] Xu, Y., Dixon, S.J., Brereton, R.G., Soini, H.A., Novotny, M.V., Trebesius, K., Bergmaier, I., Oberzaucher, E., Grammer, K., Penn, D.J. 2007. *Comparison of human axillary odor profiles obtained by gas chromatography/mass spectrometry and skin microbial profiles obtained by denaturing gradient gel electrophoresis using multivariate pattern recognition*. Metabolomics, 3(4): 427-437. https://doi.org/10.1007/s11306-007-0054-6.

[23] Cardé, R.T. 2015. *Multi-cue integration: how female mosquitoes locate a human host*. Current Biology, 25: R793-R810. https://doi.org/10.1016/j.cub.2015.07.057.

[24] McMeniman, C.J., Corfas, R.A., Matthews, B.J., Ritchie, S.A., Vosshall, L.B. 2014. *Multimodal integration of carbon dioxide and other sensory cues drives mosquito attraction to human*. Cell, 156(5): 1060-1071. https://doi.org/10.1016/j.cell.2013.12.044.

[25] Schreck, C.E., Gouck, H.K., Smith, N. 1967. *An improved olfactometer for use in studying mosquito attractants and repellents*. J. Econ. Entomol., 60(4): 1188-1190. https://doi.org/10.1093/jee/60.4.1188.

[26] Schreck, C.E., Smith, N., Carlson, D.A., Price, G.D., Haile, D., Godwin, D.R. 1981. *A material isolated from human hands that attracts female mosquitoes*. J. Chem. Ecol., 8(2): 429-438. https://doi.org/10.1007/bf00987791.

[27] Braks, M.A.H., Cork, A., Takken, W. 1997. *Olfactometer studies on the attraction of Anopheles gambiae sensu stricto (Diptera: Culicidae) to human sweat*. Proc. Exper. & Appl. Entomol., N.E.V, Amsterdam, 8: 99-104. DOI not available.

[28] Braks, M.A.H. & Takken, W. 1999. *Incubated human sweat but not fresh sweat attracts the malaria mosquito Anopheles gambiae sensu stricto*. J. Chem. Ecol., 25(3): 663-672. https://doi.org/10.1023/A:1020970307748.

[29] Knols, B.G.J., de Jong, R., Takken, W. 1995. *Differential attractiveness of isolated humans to mosquitoes in Tanzania*. Trans. R. Soc. Trop. Med. Hyg., 89(6): 604-606. https://doi.org/10.1016/0035-9203(95)90406-9.

[30] Lindsay, S.W., Adiamah, J.H., Miller, J.E., Pleass, R.J., Armstrong, J.R.M. 1993. *Variation in attractiveness of human subjects to malaria mosquitoes (Diptera: Culicidae) in the Gambia*. J. Med. Entomol., 30(2): 368-373. https://doi.org/10.1093/jmedent/30.2.368.

[31] Acree, F., Turner, R.B., Gouck, H.K., Beroza, M., Smith, N. 1968. *L-Lactic acid: a mosquito attractant isolated from humans*. Science, 161(3848): 1346-1347. https://doi.org/10.1126/science.161.3848.1346.

[32] Smith, C.N., Smith, N., Gouck, H.K., Weidhaas, D.E., Gilbert, I.J., Mayer, M.S., Smittle, B.J., Hofbauer, A. 1970. *L-Lactic acid as a factor in the attraction of Aedes aegypti (Diptera: Culicidae) to human hosts*. Ann. Entomol. Soc. Am., 63(3): 760-770. https://doi.org/10.1093/aesa/63.3.760.

[33] Braks, M.A.H., Meijerink, J., Takken, W. 2001. *The response of the malaria mosquito, Anopheles gambiae, to two components of human sweat, ammonia and L-lactic acid, in an olfactometer*. Phys. Entomol., 26: 142-148. https://doi.org/10.1046/j.1365-3032.2001.00227.x.

[34] Geier, M. & Boeckh, J. 1999. *A new Y-tube olfactometer for mosquitoes to measure the attractiveness of host odours*. Entomol. Exp. Appl., 92(1): 9-19. https://doi.org/10.1046/j.1570-7458.1999.00519.x.

[35] Smallegange, R.C., Qiu, Y.T., van Loon, J.J.A., Takken, W. 2005. *Synergism between ammonia, lactic acid and carboxylic acids as kairomones in the host-seeking behaviour of the*

malaria mosquito Anopheles gambiae sensu stricto (Diptera: Culicidae). Chem. Senses, 30(2): 145-152. https://doi.org/10.1093/chemse/bji010.

[36] Qiu, Y.T., Smallegange, R.C., van Loon, J.J.A., Takken, W. 2011. *Behavioural responses of Anopheles gambiae sensu stricto to components of human breath, sweat and urine depend on mixture composition and concentration*. Med. Vet. Entomol., 25(3): 247-255. https://doi.org/10.1111/j.1365-2915.2010.00924.x.

[37] Geier, M., Sass, H. & Boeckh, J. 1996. *A search for components in human body odour that attract females of Aedes aegypti.* In: Olfaction in Mosquito-Host Interactions. Brock G.R., Cardew, C. (ed.) Ciba Found. Symp., 200: 132-148. https://doi.org/10.1002/9780470514948.ch11.

[38] Carlson, D.A., Smith, N., Gouck, H.K., Godwin, D.R. 1973. *Yellowfever mosquitoes: compounds related to lactic acid that attract females*. J. Econ. Entomol., 66(2): 329-331. https://doi.org/10.1093/jee/66.2.329.

[39] McGovern, T., Gouck, H.K., Beroza, M., Ingangi, J.C. 1970. *Esters of α-hydroxy-β-phenyl aliphatic acids that attract yellow-fever mosquitoes*. J. Econ. Entomol., 63(6): 2002-2004. https://doi.org/10.1093/jee/66.2.329.

[40] Smallegange, R.C., Qiu, Y.T., Bukovinszkiné-Kiss, G., van Loon, J.J.A., Takken, W. 2009. *The effect of aliphatic carboxylic acids on olfaction-based host-seeking of the malaria mosquito Anopheles gambiae sense stricto*. J. Chem. Ecol., 35: 933-943. https://doi.org/10.1007/s10886-009-9668-7.

[41] Knols B.G.J., van Loon, J.J.A., Cork, A., Robinson, R.D., Adam, W., Meijerink, de Jong, R., Takken, W. 1997. *Behavioural and electrophysiological responses of the female malaria mosquito Anopheles gambiae (Diptera: Culicidae) to Limburger cheese volatiles*. B. Entomol. Res., 87(2): 151-159. https://doi.org/10.1017/S0007485300027292.

[42] Bernier, U.R., Kline, D.L., Schreck, C.E., Yost, R.A., Barnard, D.R. 2002. *Chemical analysis of human skin emanations: comparison of volatiles from humans that differ in attraction of Aedes aegypti (Diptera: Culicidae)*. J. AMCA, 18(3): 186-195.

[43] Bernier, U.R., Booth, M.M., Yost, R.A. 1999. *Analysis of human skin emanations by gas chromatography/mass spectrometry. 1. Thermal desorption of attractants for the yellow fever mosquito (Aedes aegypti) from handled glass beads*. Anal. Chem., 71(1): 1-7. https://doi.org/10.1021/ac980990v.

[44] Bernier, U.R., Kline, D.L., Barnard, D.R., Schreck, C.E., Yost, R.A. 2000. *Analysis of human skin emanations by gas chromatography/mass spectrometry. 2. Identification of*

volatile compounds that are candidate attractants for the yellow fever mosquito (Aedes aegypti). Anal. Chem., 72(4): 747-756. https://doi.org/10.1021/ac990963k.

[45] Cork, A., Park, K.C. 1996. *Identification of electrophysiologically-active compounds for the malaria mosquito, Anopheles gambiae, in human sweat extracts*. Med. Vet. Entomol., 10(3): 269-276. https://doi.org/10.1111/j.1365-2915.1996.tb00742.x.

[46] van den Broek, I., den Otter, C.J. 1999. *Olfactory sensitivities of mosquitoes with different host preferences (Anopheles gambiae s.s., An. arabiensis, An. quadriannulatus, An. m. atroparvus) to synthetic host odours*. J. Insect Physiol., 45(11): 1001-1010. https://doi.org/10.1016/S0022-1910(99)00081-5.

[47] Costantini, C., Birkett, M.A., Gibson, G., Ziesmann, J., Sagnon, N'F., Mohammed, H.A., Coluzzi, M., Pickett, J. A. 2001. *Electroantennogram and behavioural responses of the malaria vector Anopheles gambiae to human-specific sweat components*. Med. Vet. Entomol., 15(3): 259-266. https://doi.org/10.1046/j.0269-283x.2001.00297.x.

[48] Logan, J.G., Birkett, M.A., Clark, S.J., Powers, S., Seal, N.J., Wadhams, L.J., Mordue (Luntz) A.J., Pickett, J. 2008. *Identification of human-derived volatile chemicals that interfere with attraction of Aedes aegypti mosquitoes*. J. Chem Ecol, 34(3): 308-322. https://doi.org/10.1007/s10886-008-9436-0.

[49] Emami, S.N., Lindberg, B.G., Hua, S., Hill, S., Mozuraitis, R., Lehmann, P., Birgersson, G., Borg-Karlson, A-K., Ignell, R., Faye, I. 2017. *A key malaria metabolite modulates vector blood seeking, feeding, and susceptibility to infection*. Science, 355(6329): 1076-1080. https://doi.org/10.1126/science.aah4563.

[50] Cornet, S., Nicot, A., Rivero, A., Gandon, S. 2013. *Malaria infection increases bird attractiveness to uninfected mosquitoes*. Ecol. Lett., 16(3): 323-329. https://doi.org/10.1111/ele.12041.

[51] Cornet, S., Nicot, A., Rivero, A., Gandon, S. 2013. *Both infected and uninfected mosquitoes are attracted toward malaria infected birds*. Malar. J., 12(1): 179. https://doi.org/10.1186/1475-2875-12-179.

[52] Day, J.F., Ebert, K.M., Edman, J.D. 1983. *Feeding patterns of mosquitoes (Diptera: Culicidae) simultaneously exposed to malarious and healthy mice, including a method for separating blood meals from conspecific hosts*. J. Med. Entomol., 20(2): 120-127. https://doi.org/10.1093/jmedent/20.2.120.

[53] De Moraes, C.M., Stanczyk, N.M., Betz, H.S., Pulido, H., Sim, D.G., Read, A.F., Mescher, M.C. 2014. *Malaria-induced changes in host odors enhance mosquito attraction*. PNAS, 111(30): 11079-11084. https://doi.org/10.1073/pnas.1405617111.

[54] Lacroix, R., Mukabana, W.R., Gouagna, L.C., Koella, J.C. 2005. *Malaria infection increases attractiveness of humans to mosquitoes*. PloS Biol, 3(9): e298. https://doi.org/10.1371/journal.pbio.0030298.

[55] Robinson, A. Busula, A.O., Voets, M.A., Beshir, K.B., Caulfield, J.C., Powers, S.J., Verhulst, N.O., Winskill, P., Muwanguzi, J., Birkett, M.A., Smallegange, R.C., Masiga, D.K., Mukabana, W.R., Sauerwein, R.W., Sutherland, C.J., Bousema, T., Pickett, J.A., Takken, W., Logan, J.G., de Boer, J.G. 2017. *Plasmodium-associated changes in human odor attract mosquitoes*. PNAS, 115(18): E4209-E4218. https://doi.org/10.1073/pnas.1721610115.

[56] Lalubin, F., Bize, P., van Rooyen, J., Christe, P., Glaizot, O. 2012. *Potential evidence of parasite avoidance in an avian malaria vector*. Anim. Behav., 84(3): 539-545. https://doi.org/10.1016/j.anbehav.2012.06.004.

[57] de Boer, J.G., Robinson, A., Powers, S.J., Burgers, S.L.G.E., Caulfield, J.C., Birkett, M.A., Smallegange, R.C., van Genderen, P.J.J., Bousema, T., Sauerwein, R.W., Pickett, J.A., Takken, W., Logan, J.G. 2017. *Odours of Plasmodium falciparum-infected participants influence mosquito-host interactions*. Sci. Rep., 7(1): 9283. https://doi.org/10.1038/s41598-017-08978-9.

[58] Busula, A.O., Bousema, T., Mweresa, C.K., Masiga, D., Logan, J.G., Sauerwein, R.W., Verhulst, N.O., Takken, W., de Boer, J.G. 2017. *Gametocytemia and attractiveness of Plasmodium falciparum-infected Kenyan children to Anopheles gambiae mosquitoes*. J. Infec. Dis., 216: 291-295. https://doi.org/10.1093/infdis/jix214.

[59] Batista, E.P.A., Costa, E.F.M., Silva, A.A. 2014. *Anopheles darlingi (Diptera: Culicidae) displays increased attractiveness to infected individuals with Plasmodium vivax gametocytes*. PARASITE VECTOR, 7:251. https://doi.org/10.1186/1756-3305-7-251.

[60] Busula, A.O., Verhulst, N.O., Bousema, T., Takken, W., de Boer, J.G. 2017. *Mechanisms of Plasmodium-enhanced attraction of mosquito vectors*. Trends Parasitol., 33(12): 961-973. https://doi.org/10.1016/j.pt.2017.08.010.

[61] Pulido, H. Stanczyk, N.M., De Moraes, C.M., Mescher, M. 2021. *A unique volatile signature distinguishes malaria infection from other conditions that cause similar symptoms*. Sci. Rep., 11: 139928. https://doi.org/10.1038/s41598-021-92962-x.

[62] Zeng, X-N., Leyden, J.J., Spielman, A.I., Preti, G. 1996. *Analysis of characteristic human female axillary odours: qualitative comparison to males*. J. Chem. Ecol., 22(2): 237-257. https://doi.org/10.1007/bf02055096.

[63] Soini, H.A., Bruce, K.E., Klouckova, I., Brereton, R.G., Penn, D.J., Novotny, M.V. 2006. *In situ surface sampling of biological objects and preconcentration of their volatiles*

for chromatographic analysis. Anal. Chem., 78(20): 7161-7168. https://doi.org/10.1021/ac0606204.

[64] Dormont, L., Bessière, J-M., McKey, D., Cohuet, A. 2013. *New methods for field collection of human skin volatiles and perspectives for their application in the chemical ecology of human-pathogen-vector interactions*. J. Exp. Biol., 216(15): 2783-2788. https://doi.org/10.1242/jeb.085936.

[65] Jiang, R., Cudjoe, E., Bojko, B., Abaffy, T., Pawliszyn, J. 2013. *A non-invasive method for in-vivo skin volatile compounds sampling*. Analytica Chimica Acta, 804(4): 111-119. https://doi.org/10.1016/j.aca.2013.09.056.

[66] Syed, Z., Leal, W. 2009. *Acute olfactory response of Culex mosquitoes to a human- and bird-derived attractant*. PNAS, 106(44): 18803-18808. https://doi.org/10.1073/pnas.0906932106.

[67] Curran, A.M., Rabin, S.I., Prada, P.A., Furton, K.G. 2005. *Comparison of the volatile organic compounds present in human odour using SPME-GC/MS*. J. Chem. Ecol., 31(7): 1607-1619. https://doi.org/10.1007/s10886-005-5801-4.

[68] Caroprese, A., Gabbanini, S., Beltramini, C., Lucchi, E., Valgimigli, L. 2009. *HS-SPME-GC-MS analysis of body odour to test the efficacy of foot deodorant formulations*. Skin Res. Technol., 15(4): 503-510. https://doi.org/10.1111/j.1600-0846.2009.00399.x.

[69] Prada, P.A., Curran, A.M., Furton, K.G., 2011. *The evaluation of human hand odor volatiles on various textiles: a comparison between contact and noncontact sampling methods*. J. Forensic Sci., 56(4): 866-881. https://doi.org/10.1111/j.1556-4029.2011.01762.x.

[70] Mochalski, P., King, J., Unterkofler, K., Hinterhuber, H., Amann, A. 2014. *Emission rates of selected volatile organic compounds from skin of healthy volunteers*. J. Chroma. B, 959(100): 62-70. https://doi.org/10.1016/j.jchromb.2014.04.006.

[71] Mebazaa, R., Mahmoudi, A., Rega, B., Cheikh, R.B., Camel, V. 2010. *Analysis of human male armpit sweat after fenugreek ingestion: instrumental and sensory optimisation of the extraction method.* Food Chem., 120(3): 771-782. https://doi.org/10.1016/j.foodchem.2009.11.009.

[72] Mebazaa, R., Rega, B., Camel, V. 2011. *Analysis of human male armpit sweat after fenugreek ingestion: characterisation of odour active compounds by gas chromatography coupled to mass spectrometry and olfactometry*. Food Chem., 128(1): 227-235. https://doi.org/10.1016/j.foodchem.2009.11.009.

[73] Brooksbank, B.W.L., Brown, R., Gustafsson, J-A. 1974. *The detection of 5 α-androst-16-en-3α-ol in human male axillary sweat*. Experientia, 30(8): 864-865. https://doi.org/10.1007/BF01938327.

[74] Kanda, F., Yagi, E., Fukuda, M., Nakajima, K., Ohta, T. Nakata, O. 1990. *Elucidation of chemical compounds responsible for foot malodour*. Brit J. Dermatol., 122(6): 77-776. https://doi.org/10.1111/j.1365-2133.1990.tb06265.x.

[75] Qiu, Y.T., R.C., Smallegange, R.C., Hoppe, S., Van Loon, J.J.A., Bakker, E.-J., Takken, W. 2004. *Behavioural and electrophysiological responses of the malaria mosquito Anopheles gambiae Giles sensu stricto (Diptera: Culicidae) to human skin emanations*. Med. Vet. Entomol., 18(4): 429-438. https://doi.org/10.1111/j.0269-283x.2004.00534.x.

[76] Roodt, A. P., Naudé, Y., Stoltz, A., Rohwer, E. 2018. *Human skin volatiles: passive sampling and GC X GC-ToFMS analysis as a tool to investigate the skin microbiome and interactions with anthropophilic mosquito disease vectors.* Journal of Chromatography B, 1097-1098: 83-89. https://doi.org/10.1016/j.jchromb.2018.09.002.

[77] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Non-invasive sorptive extraction for the separation of human skin surface chemicals using comprehensive gas chromatography coupled to time-of-flight mass spectrometry: a mosquito-host biting site investigation*. J. Sep. Sci., 43(22): 4202-4215. https://doi.org/10.1002/jssc.202000522.

[78] Wooding, M., Rohwer, E.R., Naudé, Y. 2020. *Chemical profiling of the human skin surface for malaria vector control via a non-invasive sorptive sampler with GC×GC-TOFMS*. Anal. Bioanal. Chem., 412(23): 5759-5777. https://doi.org/10.1007/s00216-020-02799-y.

[79] Turner, C., Parekh, B., Walton, C., Španěl, P., Smith, D., Evans, M. 2008. *An exploratory comparative study of volatile compounds in exhaled breath and emitted by skin using selected ion flow tube mass spectrometry*. Rapid Commun. Mass Spectrom., 22(4): 526-532. https://doi.org/10.1002/rcm.3402.

[80] Martínez-Lozano, P. 2009. *Mass spectrometric study of cutaneous volatiles by secondary electrospray ionization*. Int. J. Mass Spectrom., 282(3): 128-132. https://doi.org/10.1016/j.ijms.2009.02.017.

[81] Martínez-Lozano, P., de la Mora, J.F. 2009. *On-line detection of human skin vapors*. J. Am. Soc. Mass Spectrom., 20(6): 1060-1063. https://doi.org/10.1016/j.jasms.2009.01.012.

[82] Ruzsanyi, V., Mochalski, P., Schmid, A., Wiesenhofer H., Klieber, M., Hinterhuber, H., Amann, A. 2012. *Ion mobility spectrometry for detection of skin volatiles*. J. Chromatgor., B., 911(1): 84-92. https://doi.org/10.1016/j.jchromb.2012.10.028.

[83] Mochalski, P., Unterkofler, K., Hinterhuber, Amann, A. 2014. *Monitoring of skin-borne volatile markers of entrapped humans by selective reagent ionization time of flight mass spectrometry in NO^+ mode*. Anal. Chem., 86(8): 3915-3923. https://doi.org/10.1021%2Fac404242q.

Chapter 7: Methods and materials

7.1) Reagents and chemical standards

Methanol, acetone, acetonitrile and *n*-hexane were purchased from Merck, Pretoria, South Africa. The solution of *n*-alkanes (C_8-C_{28}) used for the calculation of linear retention indices of selected analytes (Chapter 7.9) was purchased from Merck, Pretoria, South Africa. Heptanal (≥95%), octanal (≥95%), nonanal (≥99.5%), trans-2-octenal (94%), trans-2-decenal (≥95%) and 2-octanone (≥99.5%) were purchased from Sigma-Aldrich (Pty) Ltd., Kempton Park, South Africa, and were used as target standards (Chapter 7.7).

7.2) Preparation and conditioning of polydimethylsiloxane (PDMS) sampling loops

The in-house developed sampling loops [1] were prepared by cutting medical-grade silicone elastomer tubing (0.3 mm ID × 0.6 mm OD; SIL-TEK®, Technical Products Inc., Georgia, USA) into lengths of 18 cm which were joined end-to-end with a 1 cm length of uncoated capillary column (0.25 mm ID; SGE Analytical Science, Separation Scientific (Pty), Roodepoort, South Africa). Each loop had an average mass of 0.03-0.035 g.

To ensure analytical cleanliness, the sampling loops were cleaned and conditioned, according to a published method [2]. The loops were sonicated three times in a 1:1 v/v methanol:acetone solution for five minutes, and were then inserted individually into thermal desorption tubes (17.8 cm; 4 mm ID, 6 mm OD), and conditioned under hydrogen gas (100 mL/min) at 280 ˚C, for over twelve hours, using an off-line Gerstel™ thermal desorption (TDS) unit (Chemetrix, Midrand, South Africa). After conditioning, the loops were collectively sonicated three times, for five minutes, in acetonitrile, and were dried, wrapped in heavyweight aluminium foil, and stored in a screw-top Schott glass bottle.

7.3) Ethical considerations

The recruitment, testing and sampling protocol in this study was approved by the Ethics committees of the faculties of the Natural and Agricultural Sciences (reference: NAS036/2019) and Health Sciences (reference 606/2019) of the University of Pretoria, and permission to conduct the study in the government clinics of Masisi and Madimbo was granted by the Limpopo Provincial Department of Health (reference LP-202002-014), as well as by signed consent from the Head Sisters of the respective clinics (Appendix B.1).

7.4) Study location, participant recruitment and sample population

The recruitment, testing and sampling of participants was conducted in two daytime government clinics, Masisi and Madimbo, in the Vhembe district of the Limpopo Province of South Africa, over the period of February to March 2020. Patients visiting the respective clinics for diagnosis or treatment of malaria were enrolled in the study as participants, and provided signed and informed consent to their participation. The inclusion criteria for participation included males and females of the ages 18 to 70. Prior to testing, all participants filled in a questionnaire pertaining to aspects of general lifestyle and health (Appendix B.2). A total of 25 participants were included in the study. Originally, a larger sample size (50-80) was to be obtained, however due to the outbreak of the SARS-CoV2 virus, responsible for the COVID-19 pandemic, and the national lockdown that was implemented on the 27th of March 2020, further recruitment, testing and sampling was discontinued. Although replicates were taken, each was treated as a discrete sample (i.e.: replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning (Chapter 7.10.3), giving a total sample size of 52.

7.5) Determination of participant malaria status

The malaria status of each participant was determined in two ways: 1) by an onsite Rapid Diagnostic Test (RDT; U-Test Malaria, Humor Diagnostica, Hermanstad, South Africa), sensitive to *P. falciparum*, *P. vivax*, *P. malariae* and *P. ovale* plasmodial antigen; 2) offsite

microscope examination assay of thin-film peripheral blood smears. The index or ring finger of each participant was first dabbed with commercial isopropanol swabs prior to administration of the RDT. The finger prick required for the latter was used to collect blood for the thin-smear. Participants were informed of their RDT results onsite, and if found to be positive for infection were immediately referred for treatment to the clinic staff. Microscope analysis was performed at Ampath Laboratories, Pretoria, South Africa.

7.6) Sorptive sampling of cutaneous VOCs

The inner side of the dominant wrist of each participant was daubed with commercial isopropanol swabs to wash the skin of contaminants. Duplicate or triplicate PDMS sampling loops were placed side-by-side on the same area of the inner wrist, and then covered and secured with aluminised Mylar® film (5 cm × 11 cm; Hydroponic, South Africa; previously sonicated in a 1:1 v/v methanol:acetone solution for five minutes, and cleaned with an isopropanol swab prior to use) and Micropore surgical tape (3M; 72 mm; Dis-chem, Pretoria, Gauteng). The purpose of the Mylar® film was to shield the sampling loops from the open-air during the sampling period. Participants were free to move around while the sampling loops were attached to their wrist. After 20-30 minutes, the micropore tape and mylar were removed, the sampling loops collected, and individually wrapped in foil (heavy duty aluminium) and stored in a glass vial (labelled according to participant number and replicate number, e.g.: 1i, 1ii; 2i, 2ii etc).

7.7) Preparation and application of targeted standards

A stock solution (*n*-hexane) was prepared containing 100 µg/mL of the following target standards: heptanal, octanal, nonanal, trans-2-octenal, trans-2-decenal and 2-octanone. These compounds have previously been reported in the literature on malarial volatile emissions (Chapter 6B.4) [3]. From the stock solution, a working solution (*n*-hexane) containing 1 µg/mL of each target standard was prepared. Duplicate sampling loops were spiked, respectively, with 60, 40, 20, 15, 10, 5 and 2.5 ng of the working target standard solution. Each loop was placed on a strip of mylar wiped clean with commercial isopropanol. The requisite volume of working

target standard solution was applied to the surface of the mylar strip, adjacent to each loop, and the mylar strip was folded over and wrapped closed with surgical micropore tape. The mylar/micropore tape package containing the loop with spiked targeted standards was placed in a Schott glass submerged in a water bath at 31 ˚C to (to approximate epidermal surface temperature) for 25 minutes. Each spiked loop was wrapped in foil (heavy duty aluminium) and stored in a glass vial prior to analysis.

7.8) Instrumental and analytical methods: comprehensive GC×GC-TOFMS

Instrumental analysis was performed on a LECO© Pegasus 4D GC×GC-TOFMS fitted with a Gerstel™ TDS unit as an inlet, an Agilent© 7890 chromatograph and a dual quad-jet cryogenic modulator (LECO®, Kempton Park, South Africa), operated by ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®). The hot jets were operated with synthetic air and the cold jets were operated with nitrogen gas cooled with liquid nitrogen (Afrox, South Africa). The primary (1D) column was a capillary Rxi-5MS of length 30 m, 250 µm ID and 0.25 µm film thickness; the secondary (2D) column was a Rxi-17SilMS mid-polar capillary column of length 0.760 m, 250 µm ID and 0.25 µm film thickness.

Thermal desorption of the compounds from the PDMS sampling loops was performed using a Gerstel™ thermal desorption unit (TDS 3, Chemetrix, Midrand, South Africa), with the loops inserted into a glass thermal desorption tube with the open-end directed towards the GC inlet. The heat of desorption was from 30˚C (held for 3 min) to 280˚C (held for 10 min) at 60˚C/min, with a flow rate of 100 mL/min at a vent pressure of 10 psi of helium (ultra-high purity grade; Afrox, Gauteng, South Africa). The TDS transfer line temperature was maintained at 350˚C. Cryogenic focussing of the desorbed compounds occurred at -100˚C, using liquid nitrogen (Afrox, Gauteng, South Africa) in a cooled injection system (Gerstel™ CIS 4) with an empty, baffled and deactivated glass liner. The thermally desorbed compounds were introduced into the inlet *via* a splitless injection (purge flow of 30 mL/min after 90 s, solvent vent mode) by heating the CIS from -100˚C, at 10˚C/s, to 250˚C (and held for the duration of the GC run). The carrier gas (ultra-high purity grade helium [Afrox, Gauteng, South Africa]) flow-rate was constant at 1.4 mL/min.

Note that the 1D column was replaced during the analysis period, resulting in two sets of retention times for some analytes.

The initial temperature for the primary oven was held at 40˚C for 1.5 min, and ramped to 300˚C at a rate of 10˚C/min, with a hold time at this temperature of 8 min. The total run time was 35.5 min. The temperature programme rates for the secondary oven and the modulator were the same as that of the primary oven, but offset by +5˚C and +15˚C respectively. The transfer line to the TOFMS was maintained at a temperature of 300˚C. The modulation period was 3 s, with a hot pulse time of 0.8 s and a cool time between stages of 0.7 s.

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

7.9) Data acquisition and processing

ChromaTOF® software (version 4.51.6.0, optimised for Pegasus®) was used for data acquisition and chromatographic peak alignment. A S/N threshold of 100 was set, and deviations in retention time were bound within the modulation period (3 s), for 1D peaks, and 0.1 s for 2D peaks. Tentative identification of the analyte compounds was achieved by comparison of experimental mass spectra with reference spectra of the National Institute of Standards and Technology (NIST) library (version 2.2), with the minimum similarity threshold for a match set at 75%. Chromatographic peak areas were normalised (in terms of the peak area mean for each variable) during the processing step prior to statistical and machine learning analysis (Chapter 7.10.4).

Retention indices of reported analyte compounds were calculated using the method of the linear temperature programmed retention index, as developed by Van den Dool and Kratz [4], using a solution of *n*-alkanes (C_8-C_{28}; Merck, Pretoria, South Africa) injected into a thermal desorption tube. Due to maintenance of the primary column, two sets of *n*-alkanes were run, with retention indices of selected analytes (Chapter 7.10.6) being calculated accordingly with respect to these.

7.10) Data analysis: statistics and machine learning

7.10.1) Dataset processing prior to statistical analysis

The peak area datasets produced by comprehensive GC×GC-TOFMS analysis were processed in four steps prior to statistical analysis:

1) Contaminant compounds were removed from each individual sample data set, including organosiloxanes, halogens, boronic compounds and metallic complexes.

2) Tentatively identified compounds, with corresponding peak area values, from all samples, were combined into a single data set using the VLOOKUP function of Microsoft® Excel® (version 16.0.14228.20204).

3) Peak area values for compounds which were not detected in a particular sample were assigned a value of zero for that sample.

4) Blank corrections (from duplicate-averaged blank PDMS sampling loops [i.e.: which did not come into contact with skin]) were performed for each replicate. Compounds with resultant negative values (of which there were 676) were removed, as these were contaminants (of no informative import) occurring to a greater extent in the blank than in the samples.

7.10.2) Preliminary statistical analysis: principal component and linear discriminant analysis

The four-step processing resulted in a dataset (N=52) of 3166 compounds. In order to determine the efficacy of the blank correction, and as a preliminary assessment of the variation across the sample data, a Principal Component Analysis (PCA), with a Wide estimation, and a Linear Discriminant Analysis (LDA), with a Wide Linear fit, were performed on the final dataset, using JMP® (Version 15.0.0) statistical software.

7.10.3) Machine learning (regression and classification)

Machine learning was performed using R© computational and statistical software (version 1.3.959) with the Classification and Regression Training (caret) package (version 6.0-86) [5]. The method was based on that of Kuhn, 2008 [6]. The seed function was set at 23 for all computations. Three algorithms were used to construct regression and classification models of the data: an elastic-net regression (using the glmnet algorithm), a random forest (ranger) and a support vector machine (svmPoly). The analysis pipeline outlined in Chapter 2.3 (Figure 53) was followed for each algorithm:

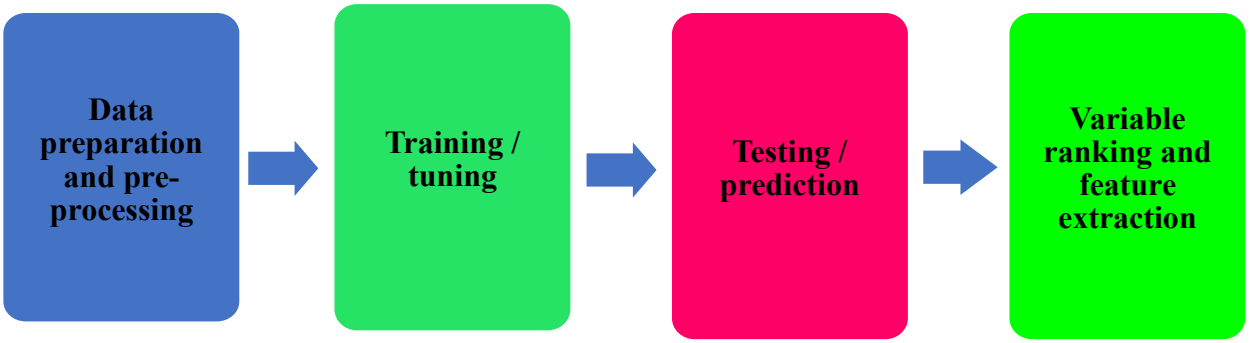


Figure 53: The machine learning pipeline followed for the training and testing of the elastic-net regression (glmnet), random forest (ranger) and support vector machine (svmPoly) algorithms (Chapter 2.3). Adapted from the method of Kuhn, 2008 [6].

7.10.4) Dataset splitting and pre-processing

The dataset (after contaminant removal and blank correction) was shuffled and randomly split with a 0.5 ratio into a training and testing set, such that there was an approximately equal proportion of malaria-positive and -negative samples in each set. For the results of each of the three models to be comparable, the same train/test split was used for the training and testing of each model. After the preliminary statistical analysis, and prior to machine learning, the training dataset was pre-processed, using pre-processing functions of the caret package [7], in two steps:

1) The data was normalised, i.e.: centred (by subtracting the peak area mean of each variable from the sample peak area values) and scaled (by dividing the result of centring by the standard deviation of the peak area mean).

2) Variables that had zero or near-zero variance were removed from the dataset. Pre-processing resulted in the removal of 2038 compound variables, and 1128 variables were retained, centred and scaled. The same pre-processed dataset was used for the training of each of the three models. Although replicates were taken, each was treated as a discrete sample (i.e.: replicates were not averaged), in order to obtain a sufficiently large sample population for train/test splitting during machine learning.

7.10.5) Model tuning and training

The parameters and coefficients of each model were computed and optimised using the resampling trainControl function of caret to perform a five-fold, five-times-repeated cross-validation. The function selects those model parameters with the highest AUC/ROC values as determined by cross-validation.

For the elastic-net regression, using the glmnet algorithm of the caret package, the tuning parameters are alpha (α) and lambda (λ); alpha is equal to the fractional contribution of the ridge-regression penalty to the total ridge-lasso penalty on the regression coefficients of the model, and lambda is the magnitude of the penalty (Chapter 2.4.1). Five values of alpha, between zero and one, and ten values of lambda, between 0.0001 and one, were used to tune the model.

For the random forest, using the ranger algorithm, the tuning parameters are mtry (the number of randomly selected variables used at each nodal split), the split rule for the splitting of each node, and the minimal node size [6]. The values of mtry used to tune the model were 2, 3 and 5; the split rules used were gini and extratrees, and the minimal node sizes used were 1, 3 and 5.

For the support vector machine, using the svmPoly algorithm, the tuning parameters are the degree and scale of the kernel function, and C (Chapter 2.4.3). The values of C used to tune the model were 0.01, 0.1, 1 and 10, and values of 1-3 were set for the degree and scale.

7.10.6) Variable importance and feature selection

Predictors included in model training were ranked in terms of their relative importance (Chapter 2.3.4). For the glmnet algorithm, variables are ranked according to the values of the regression coefficients of the final tuned model [7]. For the ranger algorithm, the predictors are ranked according to predictive accuracy from out-of-bag resampling [7]. For the svmPoly algorithm, the AUC/ROC from in-training resampling is used to rank the variables [7]. For each model, the twenty highest ranking compounds are listed, and retention indices (Chapter 7.9) for these compounds, where applicable, were calculated. Peak area values of the top variables were normalised (i.e.: centred and scaled) and visualised as a heatmap, using the heatmap.2 function of the gplots package (version 3.1.3).

References

[1] Roodt, A. P., Naudé, Y., Stoltz, A., Rohwer, E. 2018. *Human skin volatiles: passive sampling and GC X GC-ToFMS analysis as a tool to investigate the skin microbiome and interactions with anthropophilic mosquito disease vectors.* Journal of Chromatography B, 1097-1098: 83-89. https://doi.org/10.1016/j.jchromb.2018.09.002.

[2] Triñanes, S., Pena, M.T., Casais, M.C., Mejuto, M.C. 2015. *Development of a new sorptive extraction method based on simultaneous direct and headspace sampling modes for the screening of polycyclic aromatic hydrocarbons in water samples.* Talanta, 132: 433-442. https://doi.org/10.1016/j.talanta.2014.09.044.

[3] Busula, A.O., Bousema, T., Mweresa, C.K., Masiga, D., Logan, J.G., Sauerwein, R.W., Verhulst, N.O., Takken, W., de Boer, J.G. 2017. *Gametocytemia and attractiveness of Plasmodium falciparum-infected Kenyan children to Anopheles gambiae mosquitoes*. J. Infec. Dis., 216: 291-295. https://doi.org/10.1093/infdis/jix214.

[4] Van den Dool, H., Kratz, P.D. *A generalization of the retention index system including linear temperature programmed gas-liquid partition chromatography*. J. Chromatogr., 11: 463-471. https://doi.org/10.1016/S0021-9673(01)80947-X.

[5] Kuhn, M. 2019. *The caret Package*. Available from: http://topepo.github.io/caret/index.html [Accessed: 23/11/2021].

[6] Kuhn, M. 2008. *Building predictive models in R using the caret package*. J. Stat. Softw., 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

[7] Kuhn, M. 2018. *Caret: classification and regression training*. Available from: https://cran.r-project.org/web/packages/caret/index.html [Accessed: 23/11/2021].

Chapter 8: Results and discussion

8.1) Determination of malaria-infection status

The malaria-status was determined for each participant using the rapid diagnostic test (RDT) and microscopy. Of the twenty-five participants included, three showed at least one positive assay. The first (#4) tested positive by RDT for either *Plasmodium vivax (P. vivax)*, *P. malariae* or *P. ovale*, but tested negative by microscopy. The second (#17) tested positive for *P. falciparum* by RDT and microscopic analysis, and had been infected with malaria three times previously (Appendix B.2). The third participant (#18) tested positive for *P. falciparum* by the RDT administered by a healthcare worker at the Madimbo clinic, but showed a negative result on a second RDT administered by the researcher, as well as a negative result by microscopy. This case may be a false positive, since the test line was observed to show faintly, and is possibly due to the fact that the patient may have been on antimalarial medication for the past 4-7 days, as they had stated in the general health and lifestyle questionnaire [Appendix B.2])[7]. However, due to the paucity of the sample size for malaria-positive samples, and the discontinuation of participant recruitment and sampling due to the national lockdown that was implemented on the 27th of March 2020, this sample is classified as positive for the purpose of machine learning analysis. The outcome of participant #4 is statistically less likely, since *P. falciparum* accounts for over 90% of cases in southern Africa [1]. The participant's status was negative by microscopy, however, the prolonged storage of the blood smears during the national lockdown period may have compromised the quality of the smear.

8.2) Chromatographic data from comprehensive GC×GC-TOFMS

Profiles of the cutaneous VOCs of each participant were obtained using comprehensive GC×GC-TOFMS. Two-dimensional total ion chromatograms (TICs) for malaria-positive (POS) participants, and two malaria-negative (NEG) participants, are presented in Figures

[7] The participant may have misunderstood the question, or answered it incorrectly, since the health worker at the clinic provided the patient with medication on the day the RDT was administered, and the participant presumably visited the clinic on the day for the purpose of receiving diagnosis and treatment. The participant had been infected with malaria on three previous occasions, and may have been referring to medication for a previous infection.

54-63, with selected peaks indicated. A pattern of dense peak clusters (±300-1200s) is common across samples of both categories (POS/NEG). There is a large number of significantly broadened unknown peaks (MS similarity match of <750), and peaks due to siloxanes (originating from the column and PDMS loops) and semi- to non-volatile fatty acids (particularly at 1D retention times >1000s). The use of a membrane between the epidermis and the PDMS sampling loop could prevent contamination of the latter from semi- to non-volatile components.

There is no observable difference in gross peak structure between POS and NEG cases. Compound selected peaks common to the chromatograms presented include toluene, 1,5-heptdien-3-yne, furfural, 2(5H)-furanone and 1,3,5,7-cyclooctatetraene. Toluene has been previously reported as a key compound associated with malaria-infection [2]. Notable compound peaks in chromatograms for POS patients include isothiazole, isoamyl cyanide, 2-methylbutanoic anhydride, *n*-decanoic acid (#17i; Figure 57), (E)-2-octen-1-ol (#17i and #18i; Figure 57 and Figure 59) and (E,E)-2,4-decadienal (#18i; Figure 59). As discussed in Chapter 8.7, (E)-2-octen-1-ol; 2-methylbutanoic anhydride; (E,E)-2,4-decadienal; isothiazole and isoamyl cyanide are ranked as top compounds by machine learning. The compound diethyl phthalate (Figure 51, Figure 61 and Figure 63) is likely a contaminant. The compound 2-ethyl-1-hexanol, present in POS cases #17i and #18i (Figure 57 and Figure 59) has been reported as a putative marker of malaria-status, specifically for febrile children with diarrhea [2]. However, it was found also to occur with comparable relative abundance in malaria-negative cases.

Note that due to the long modulation period (3 s; Chapter 7.8) the chromatograms have unused separation space in the second dimension which could be used in future analyses by reducing the modulation period to 2 s.

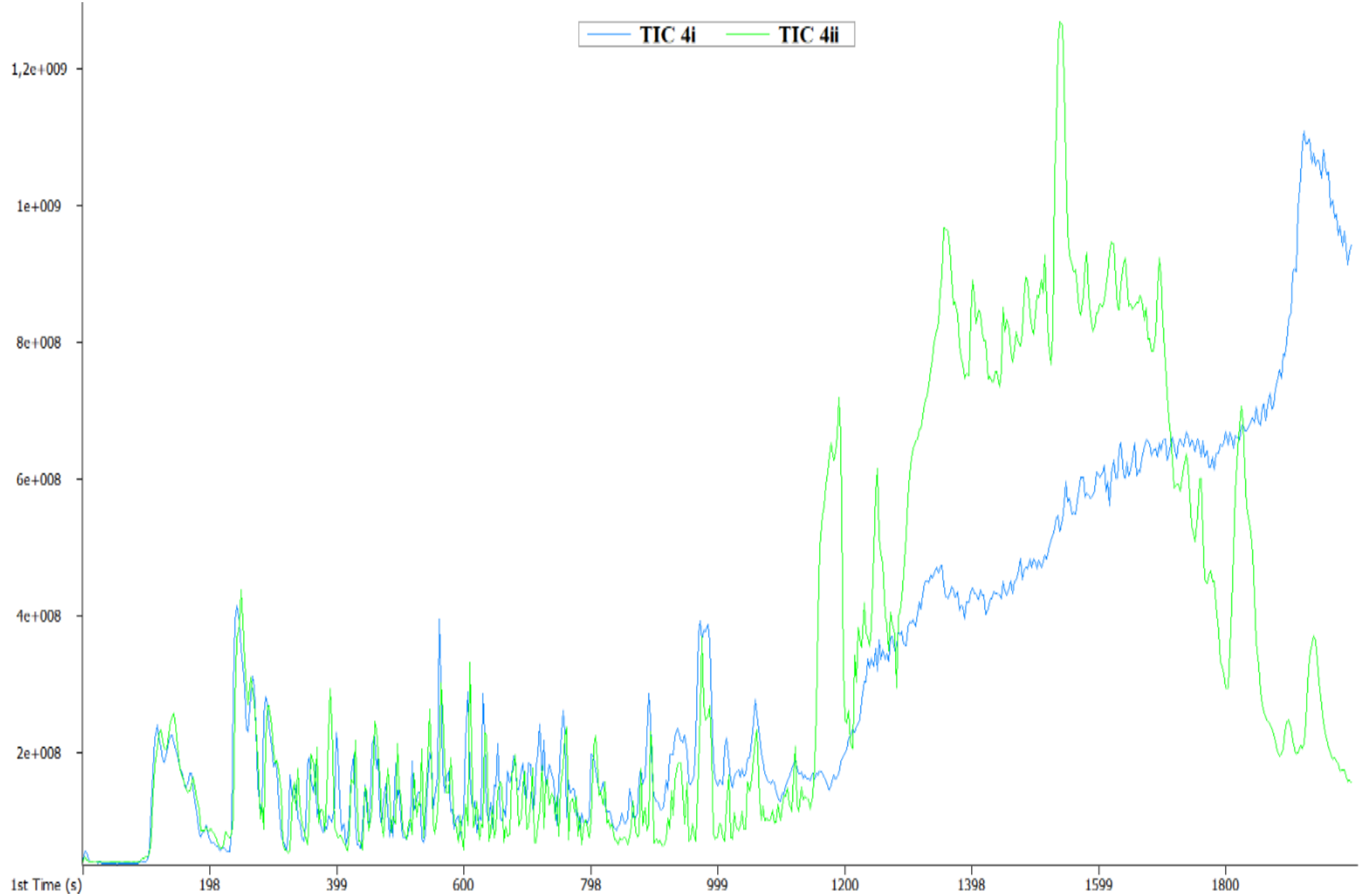


Figure 54: 1D TIC overlay of replicate cutaneous extractions from a malaria-positive participant (#4i and 4ii). Peaks of interest are in the 1D retention time region of 300-1000 s.

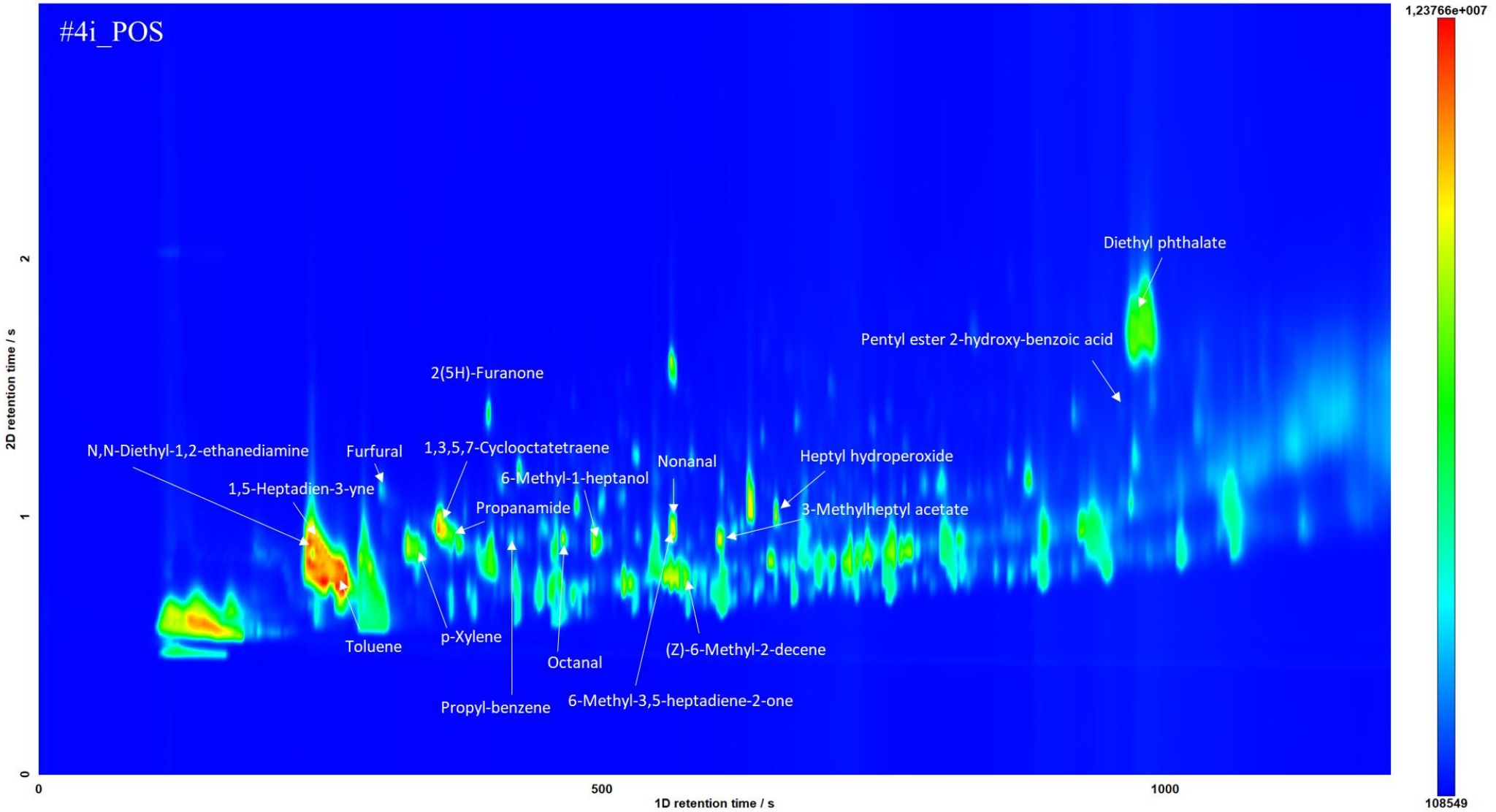


Figure 55: 2D TIC contour plot of a replicate cutaneous extraction from a malaria-positive participant (#4i). Peaks of interest are in the 1D retention time region of 300-1000 s. Selected peaks are labelled either because they are top-ranking variables (Chapter 8.7), are reported in literature [2] or because they are prominent selectable peaks on the 2D chromatogram. Many visible peaks are below the MS similarity threshold (<75%), or correspond to siloxanes, which were removed from the dataset.

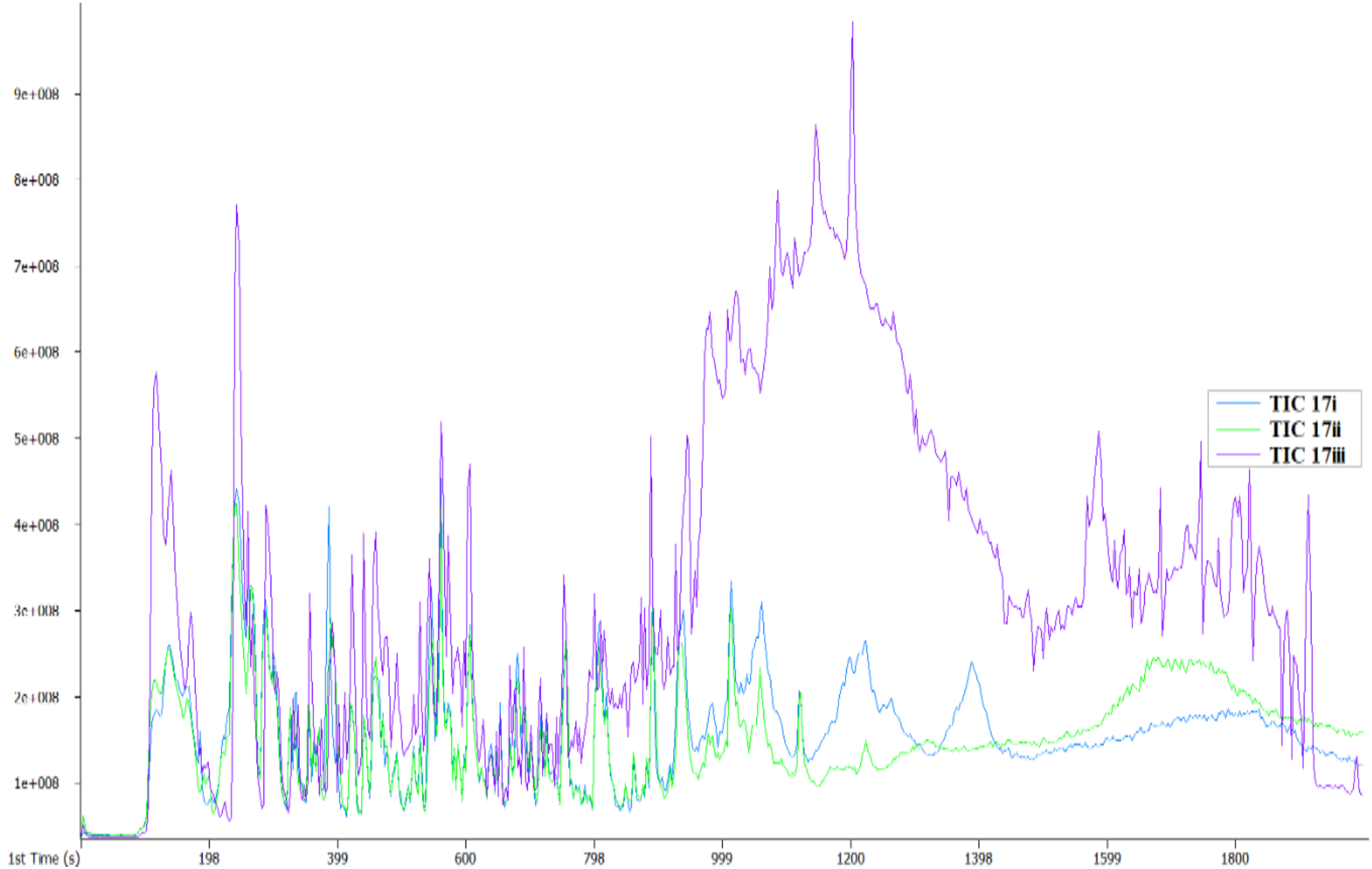


Figure 56: 1D TIC overlay of replicate cutaneous extractions from a malaria-positive participant (#17i, 17ii and 17iii). Peaks of interest are in the 1D retention time region of 300-1000 s.

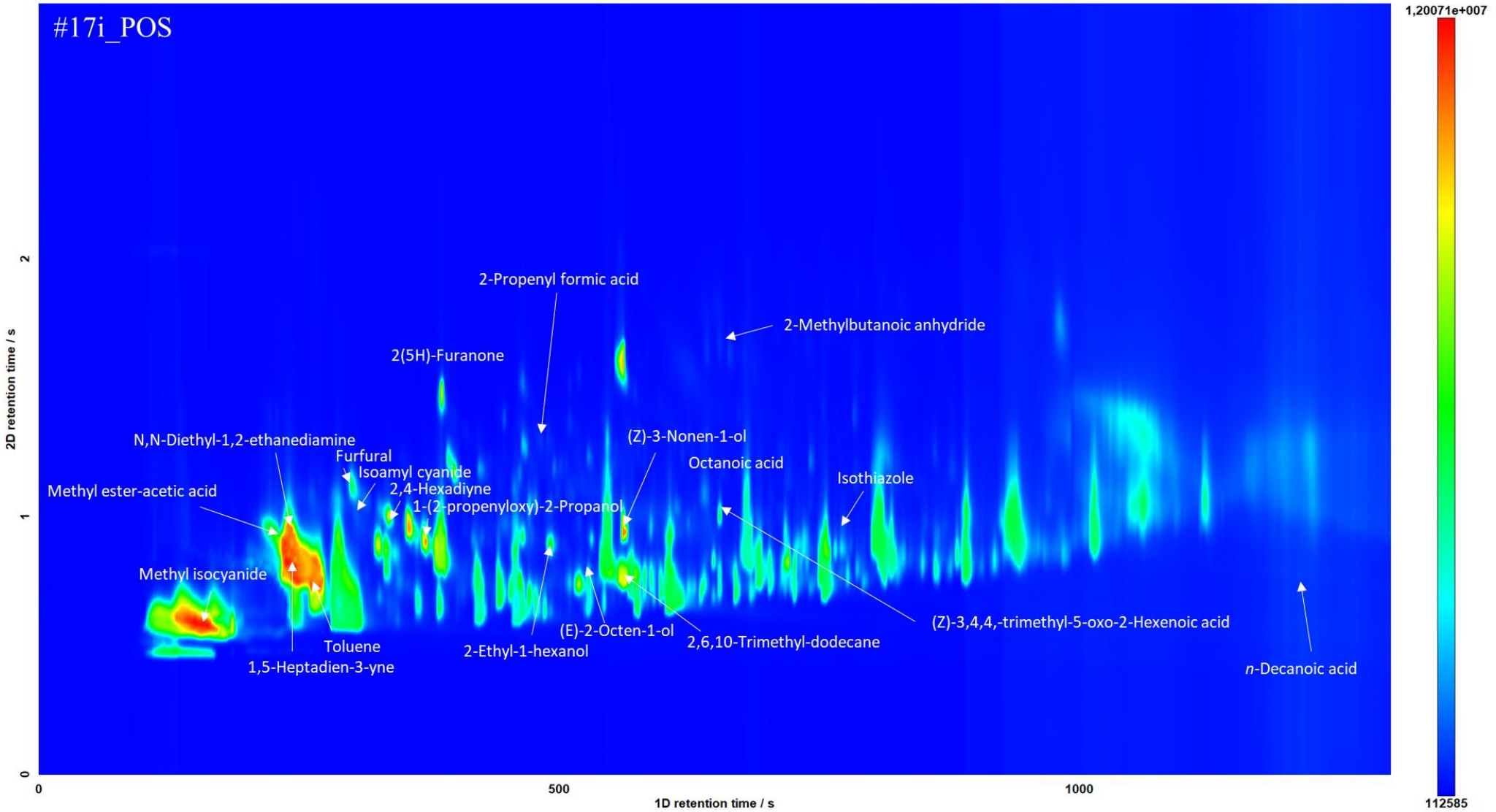


Figure 57: 2D TIC contour plot of a replicate cutaneous extraction from a malaria-positive participant (#17i). Peaks of interest are in the 1D retention time region of 300-1000 s. Selected peaks are labelled either because they are top-ranking variables (Chapter 8.7), are reported in literature [2] or because they are prominent selectable peaks on the 2D chromatogram. Many visible peaks are below the MS similarity threshold (<75%), or correspond to siloxanes, which were removed from the dataset.

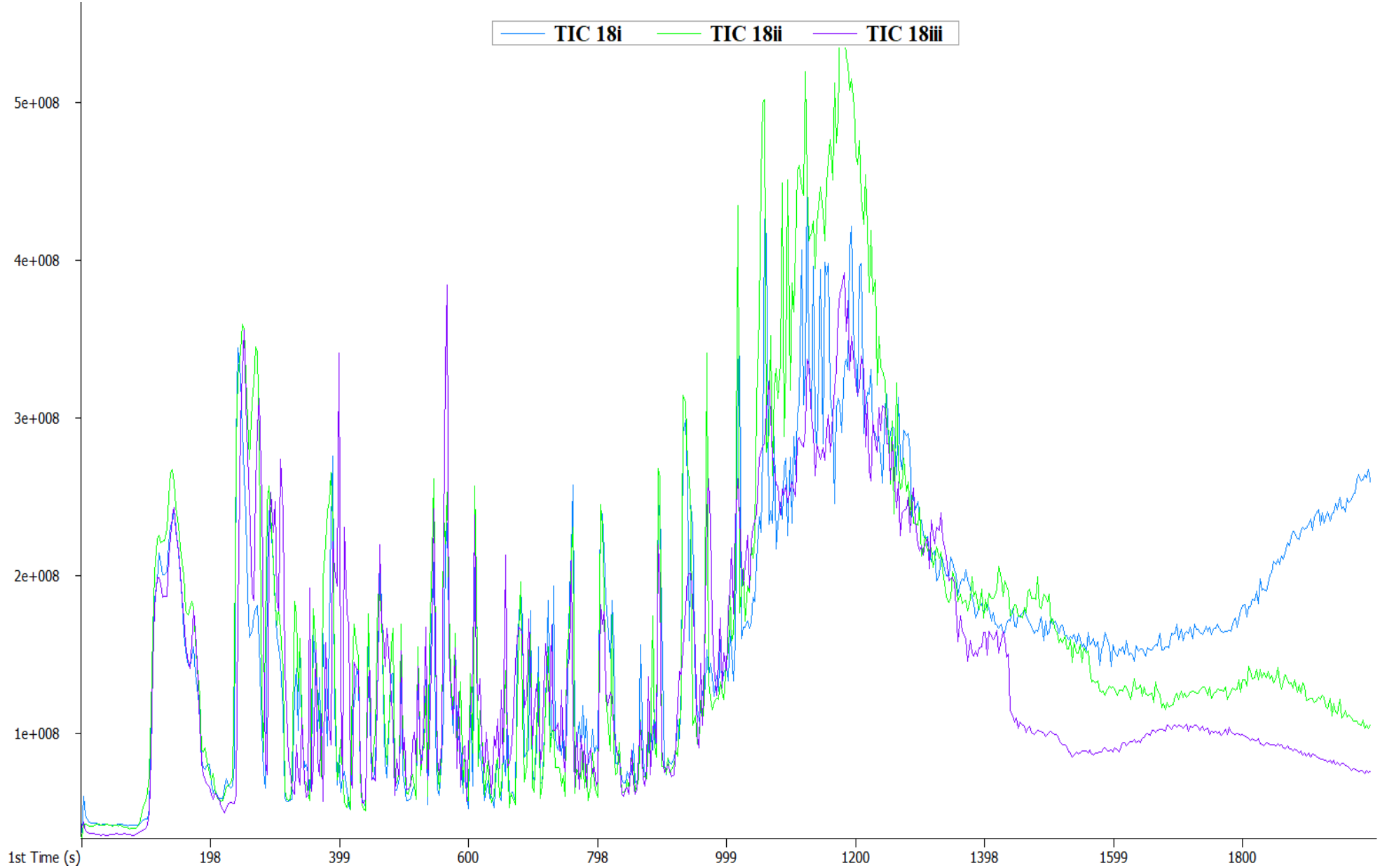


Figure 58: 1D TIC overlay of replicate cutaneous extractions from a malaria-positive participant (#18i, 18ii and 18iii). Peaks of interest are in the 1D retention time region of 300-1000 s.

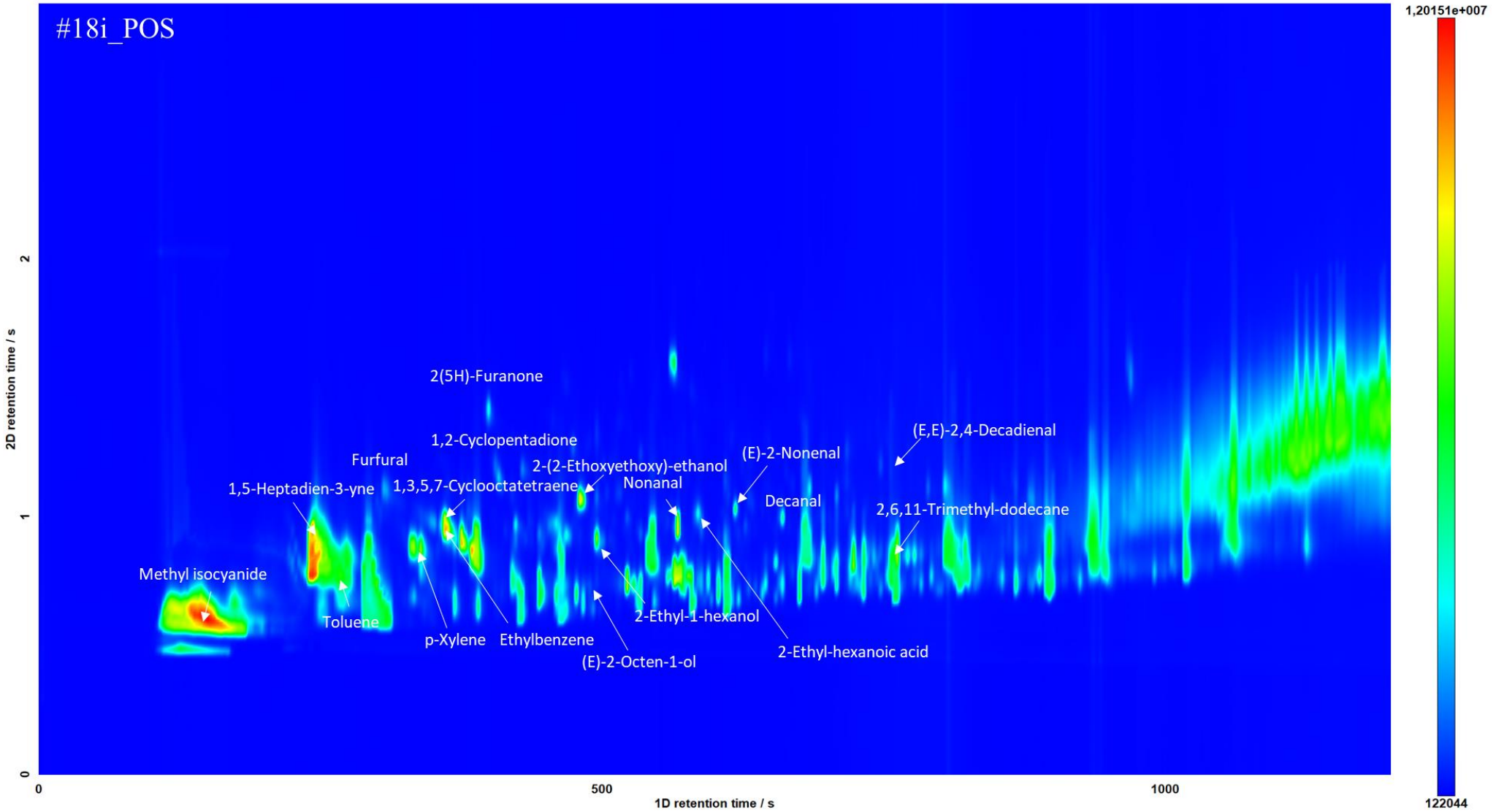


Figure 59: 2D TIC contour plot of a replicate cutaneous extraction from a malaria-positive participant (#18i). Peaks of interest are in the 1D retention time region of 300-1000 s. Selected peaks are labelled either because they are top-ranking variables (Chapter 8.7), are reported in literature [2] or because they are prominent selectable peaks on the 2D chromatogram. Many visible peaks are below the MS similarity threshold (<75%), or correspond to siloxanes, which were removed from the dataset.

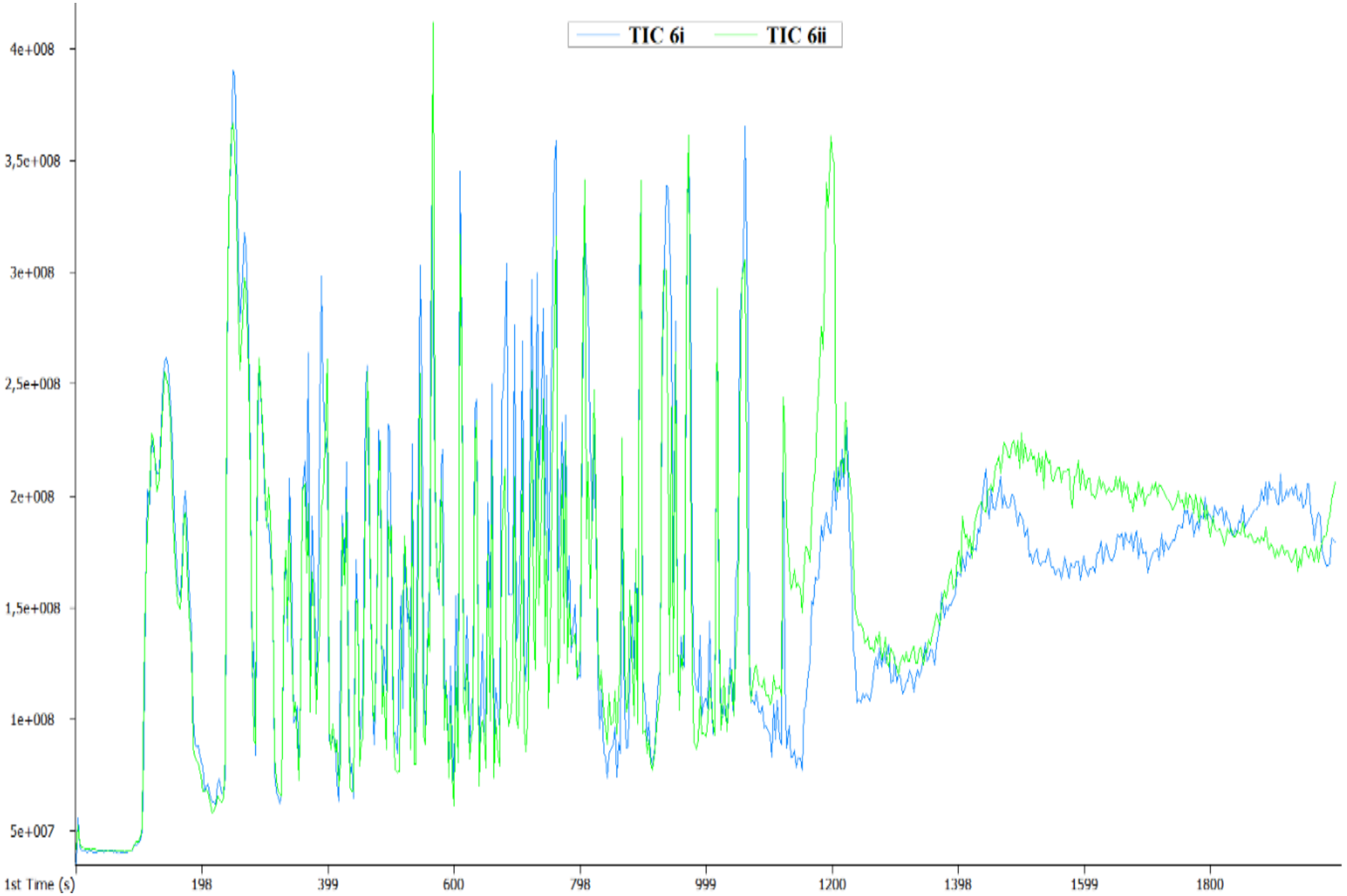


Figure 60: 1D TIC overlay of replicate cutaneous extractions from a malaria-negative participant (#6i and 6ii). Peaks of interest are in the 1D retention time region of 300-1000 s.

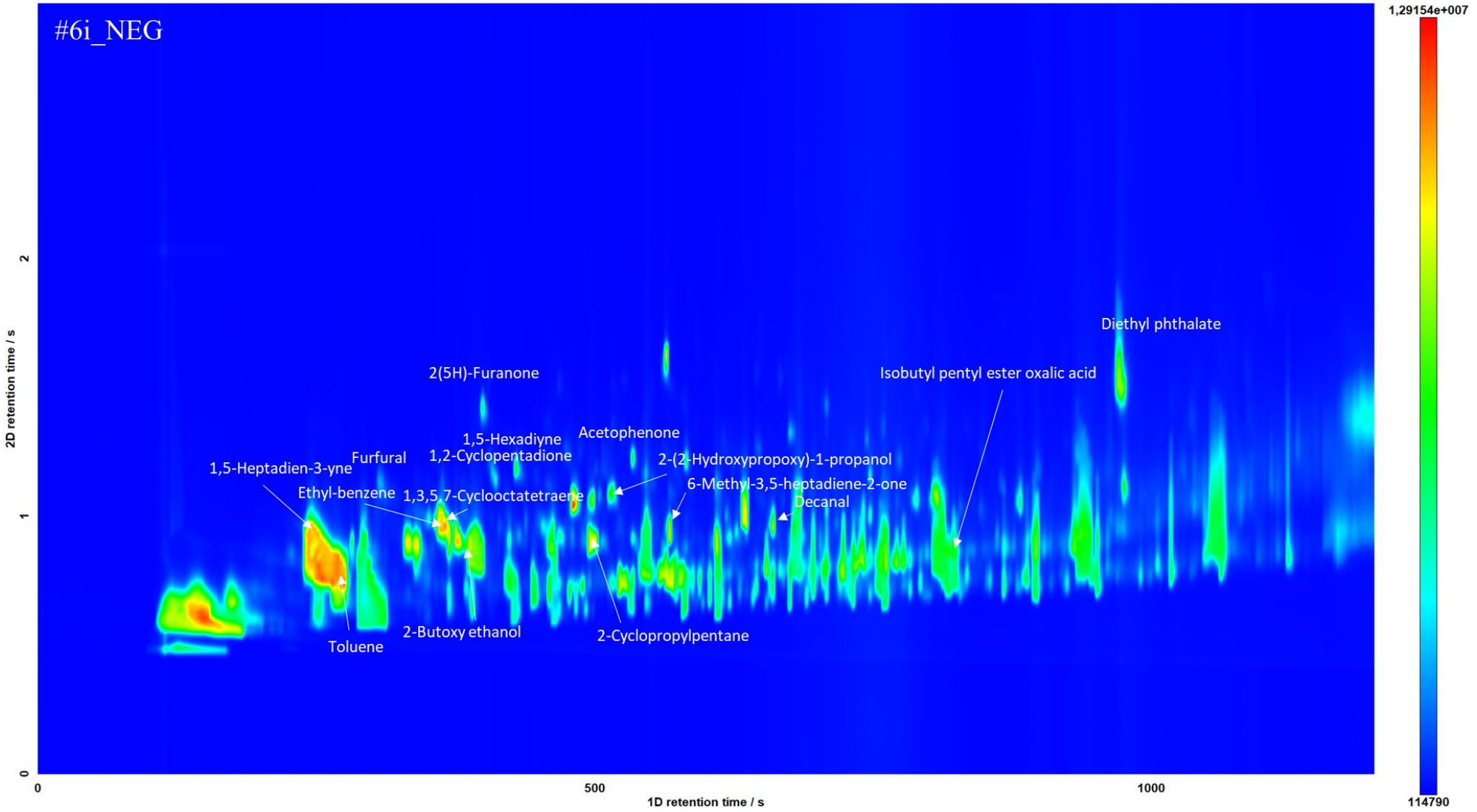


Figure 61: 2D TIC contour plot of a replicate cutaneous extraction from a malaria-negative participant (#6i). Peaks of interest are in the 1D retention time region of 300-1000 s. Selected peaks are labelled either because they are top-ranking variables (Chapter 8.7), are reported in literature [2] or because they are prominent selectable peaks on the 2D chromatogram. Many visible peaks are below the MS similarity threshold (<75%), or correspond to siloxanes, which were removed from the dataset.

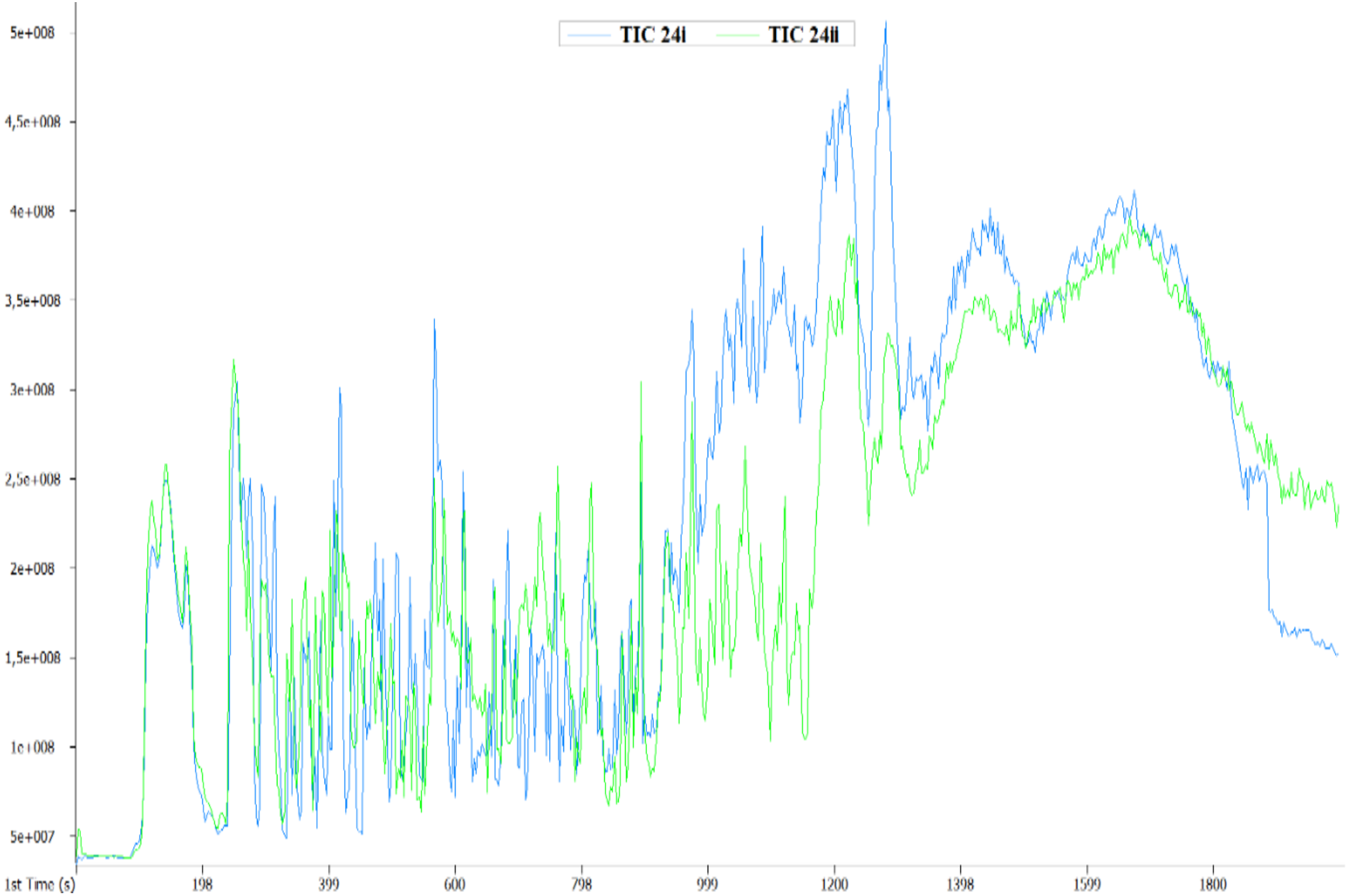


Figure 62: 1D TIC overlay of replicate cutaneous extractions from a malaria-negative participant (#24i and 24ii). Peaks of interest are in the 1D retention time region of 300-1000 s.

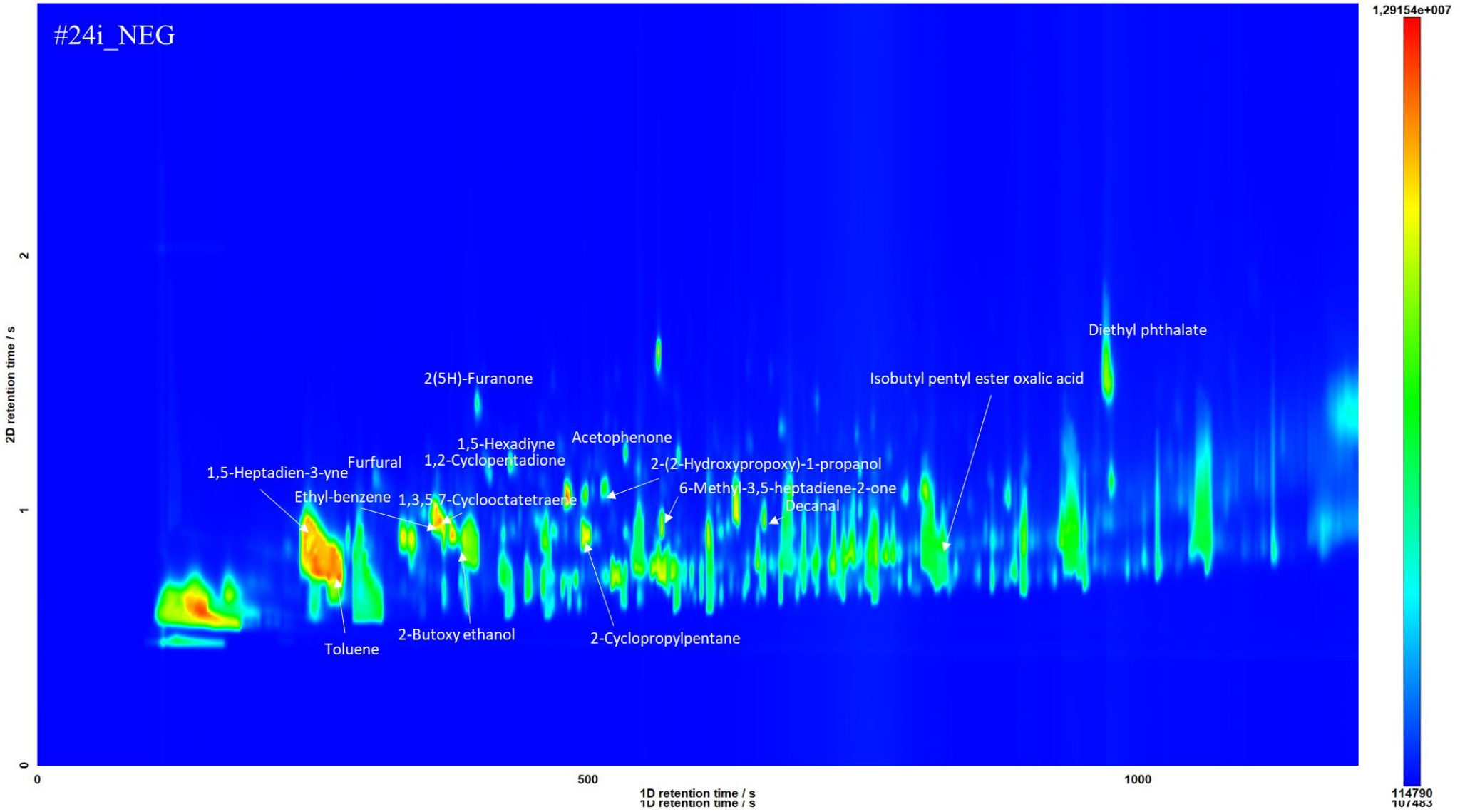


Figure 63: 2D TIC contour plot of a replicate cutaneous extraction from a malaria-negative participant (#24i). Peaks of interest are in the 1D retention time region of 300-1000 s. Selected peaks are labelled either because they are top-ranking variables (Chapter 8.7), are reported in literature [2] or because they are prominent selectable peaks on the 2D chromatogram. Many visible peaks are below the MS similarity threshold (<75%), or correspond to siloxanes, which were removed from the dataset.

Figure 64 is an overlay of the TICs of the blank PDMS loops and two replicate cutaneous samples from different healthy participants. Peaks from each TIC, including the blanks, are densely distributed across the full retention time range, which suggests contamination of the sample loops (although this is not easily ascertained by simple visual inspection). This may be due to the fact that the samples were stored for a prolonged period during the national lockdown instituted in February 2020. Note that blank corrections were performed (Chapter 7.10.1).

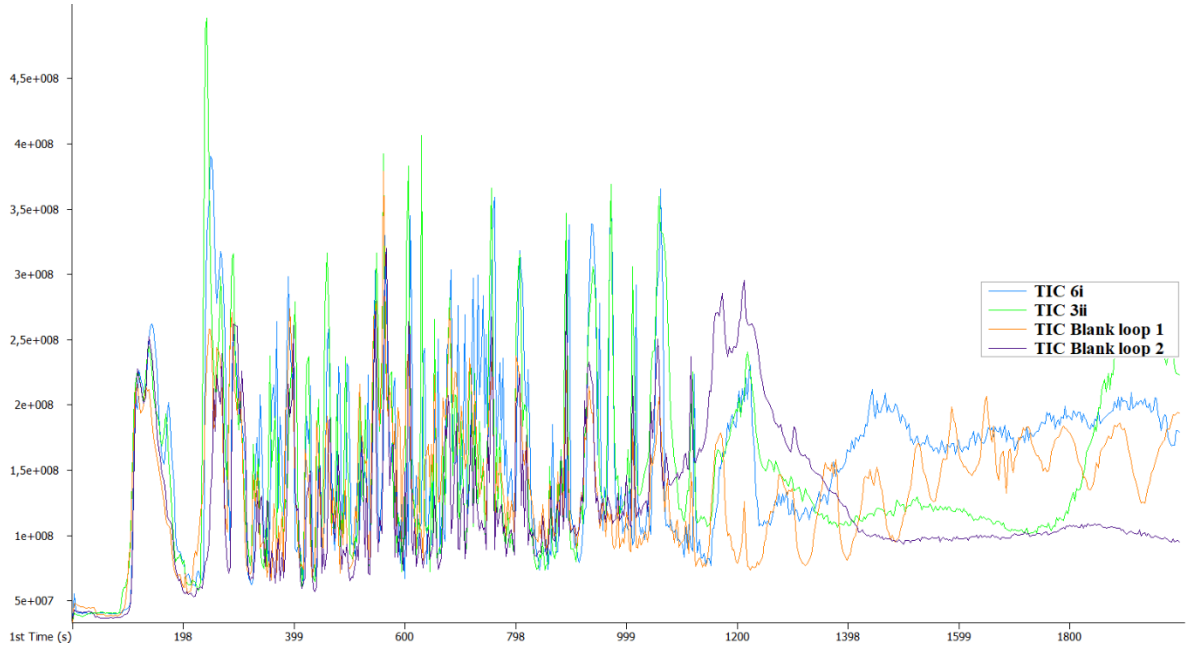


Figure 64: Overlay of the 1D TICs of two blank PDMS loops and cutaneous VOC samples from two healthy participants (6i and 3ii).

8.3) Preliminary statistical analysis: PCA and LDA

As a preliminary assessment of the variation between samples in the cutaneous VOC profiles of the participants, principal component analysis (PCA) and linear discriminant analysis (LDA) were applied to the full dataset, described by 3166 variables (compounds), prior to processing for supervised learning.

Figure 65 is a PCA score plot for samples corresponding to the blank PDMS loop and the cutaneous samples, and Figure 66 is an amplified view of the same plot. PCA decomposes the full dataset into 51 components that capture 99.92% of the variance. Tabulated eigenvalues are provided in Appendix B.3. The first component, which accounts for the greatest variation, represents only 4.88% of the cumulative total, suggesting that the variation is not strongly influenced by a small number of compounds.

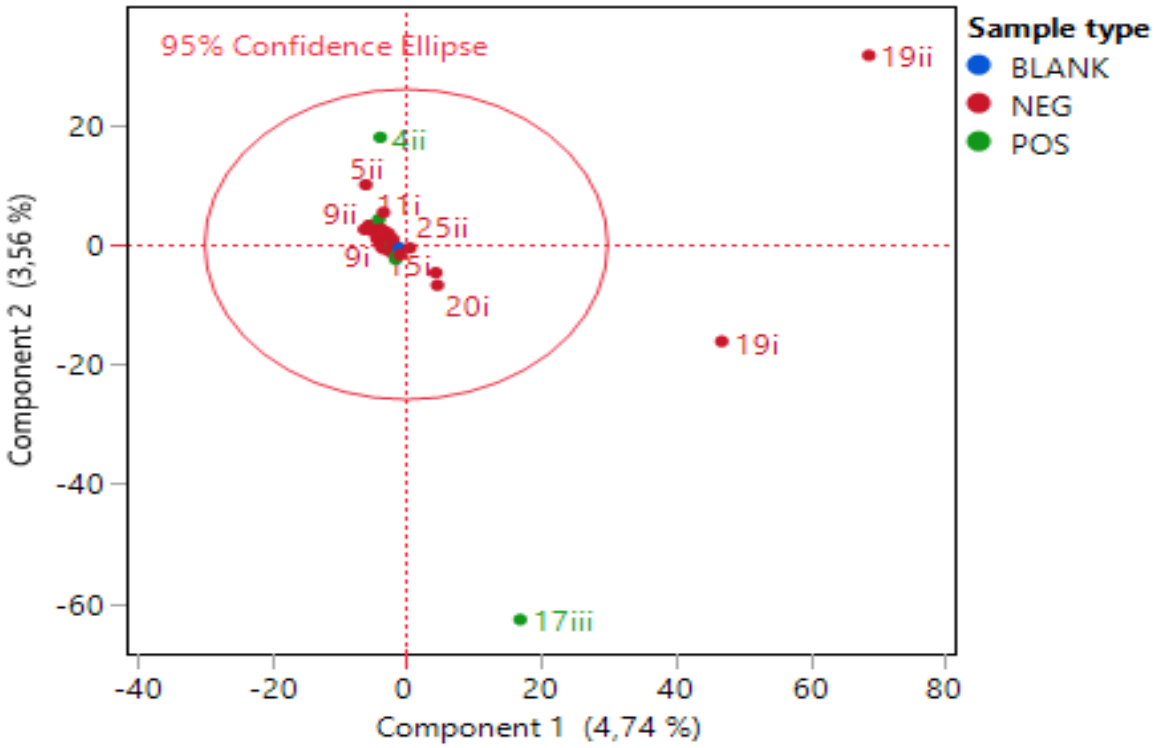


Figure 65: Score plot for the first two principal components of the blank loop and the cutaneous samples (for malaria-negative and -positive individuals).

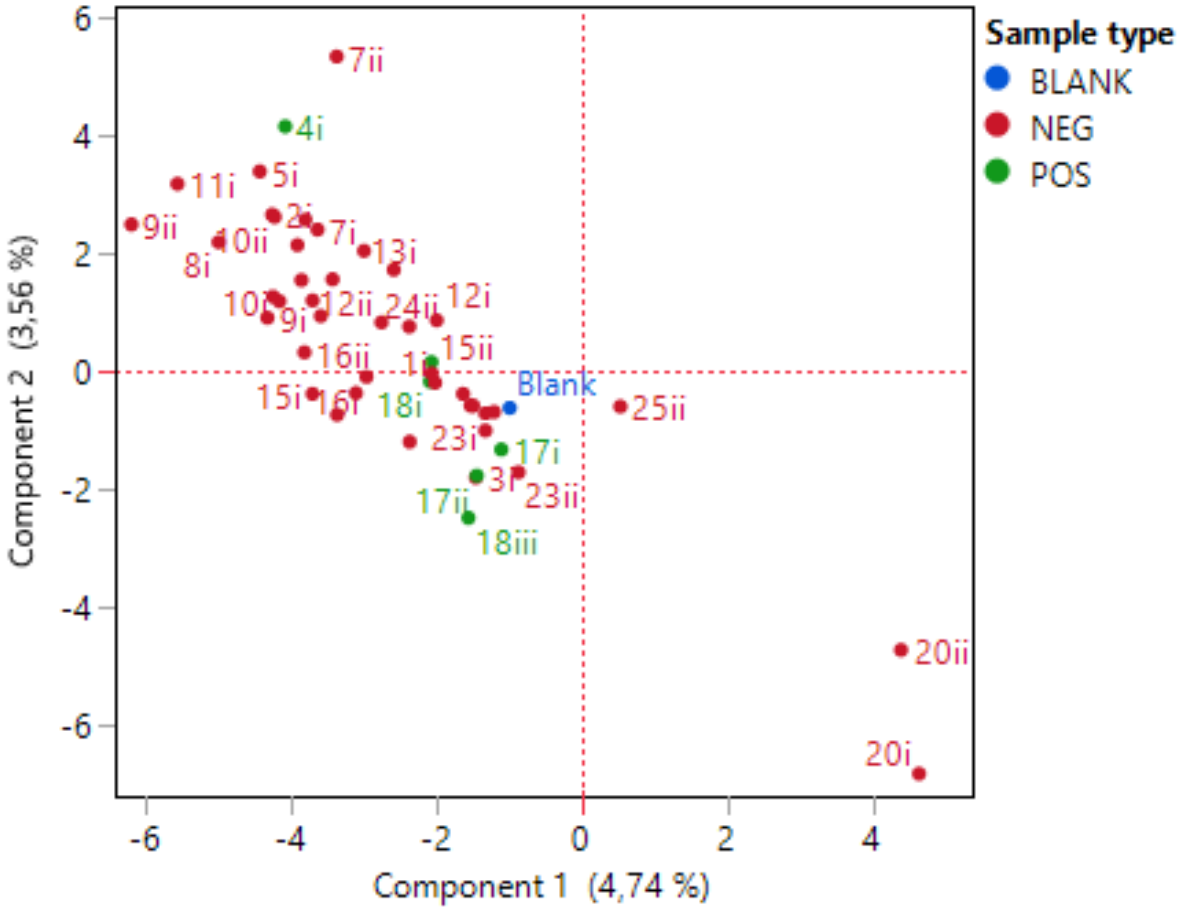


Figure 66: Amplified view of the region around the origin in the score plot in Figure 65.

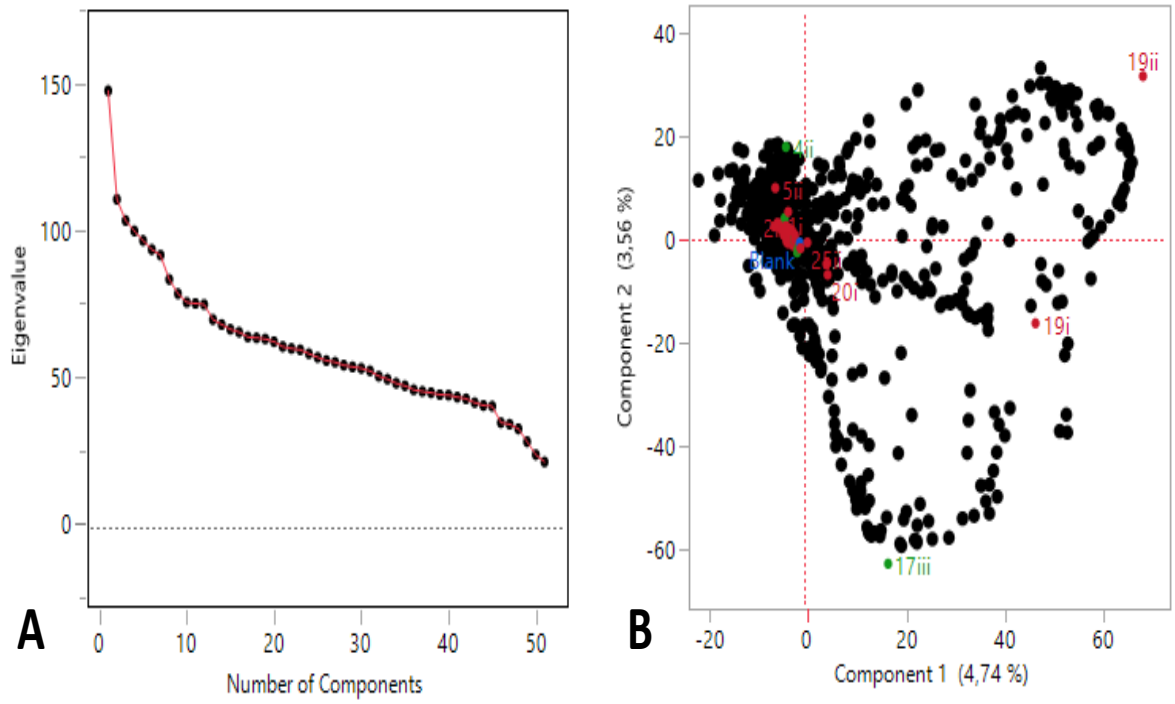


Figure 67: A) Scree plot of the eigenvalues for the 50 principal components; B) PCA biplot, overlaying the score plot in Figure 65 and the loading plot for the variables (point labels for the latter not provided for the sake of clarity).

The point for the blank is observed to be located in close proximity to the majority of samples, which suggests contamination of the sample loops, and possible loss and/or degradation of VOCs during the prolonged storage time. Overall, most of the samples cluster close together at low values of the first principal component, with little differentiation between samples corresponding to positive and negative cases for *Plasmodium* infection. However, one replicate for a positive sample (#17iii) is observed to fall distally from the main cluster, outside of the 95% confidence ellipse (95% CI) (Figure 65), and another replicate from another positive case (#4ii) lies removed from the main cluster, but still within the 95% confidence ellipse. Two replicate samples from a malaria-negative participant (#19i and 19ii) fall outside of the 95% CI, which is potentially noteworthy considering that the participant reported having had malaria more than three times in the past, and also reported suffering from high blood-pressure (Appendix B.2).

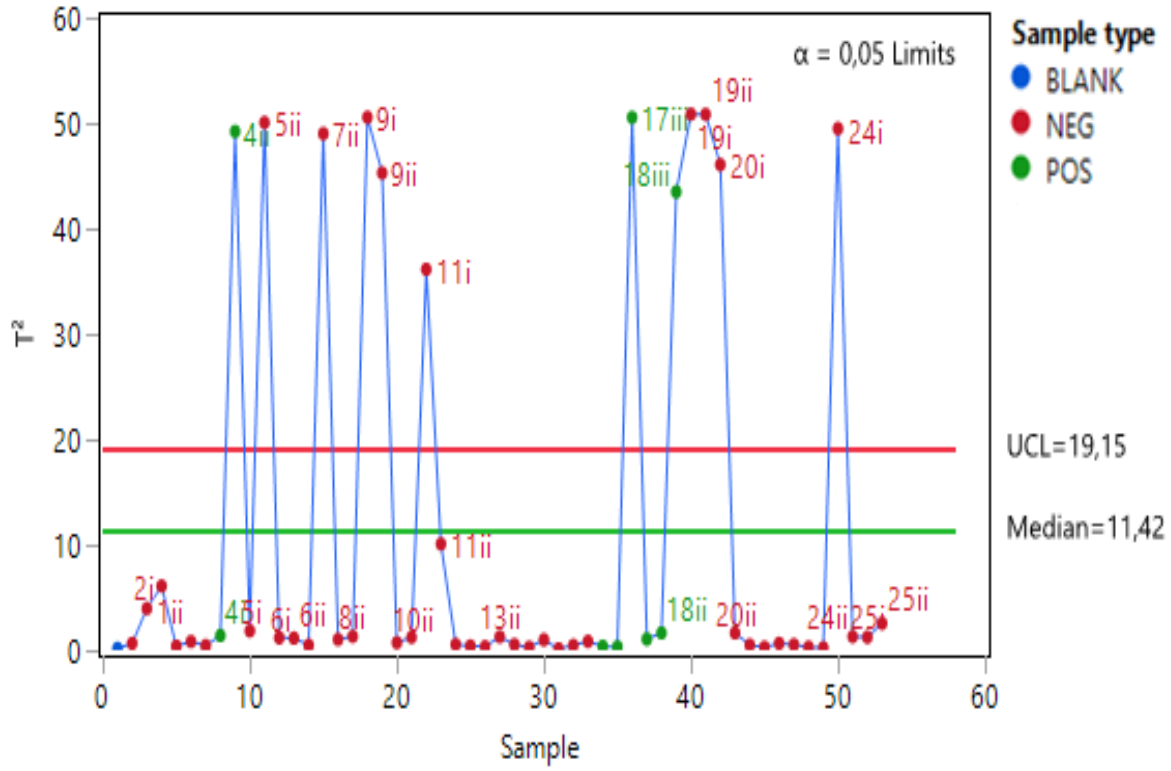


Figure 68: Outlier analysis using Hotelling's T^2 statistic (α=0.05); UCL = upper confidence limit.

A multivariate outlier test using Hotelling's T^2 statistic reveals 12 outliers at a significance threshold of α = 0.05, including notably three replicates of the positive cases: #4ii, 17iii and 18iii (Figure 68), and two points for replicates of a negative case (#19). Since an explicit test for a hypothesis of difference is not in this case the purpose of the PCA, and since biological variation is expected, the points identified as outliers were retained.

LDA plots the data according to the maximal separability between samples of different classification. Figure 69 is a canonical plot for the supervised clustering of the cutaneous samples (positive and negative) and the blank loop. Separable grouping of the negative samples from the positive samples (distributed along a diagonal of the canonical axes) is apparent, however, there is substantial overlap of the 50% confidence ellipses of the two groups, as well as with that of the blank. The 95% confidence ellipse for the negative group encompasses 50% of the positive cases.

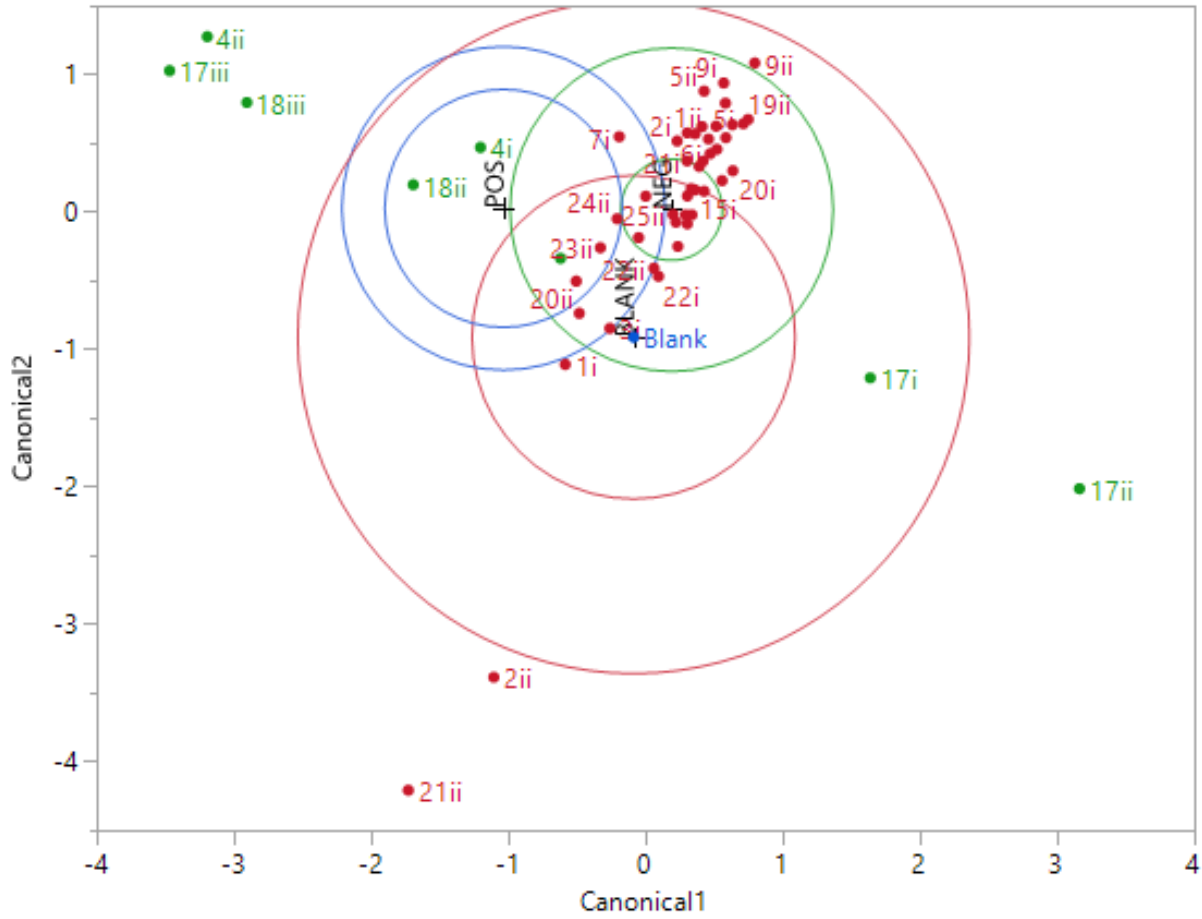


Figure 69: Canonical plot for the blank and cutaneous samples (malaria-positive and -negative) with 95% (outer) and 50% (inner) confidence ellipses shown.

8.4) Data pre-processing for machine learning

The full dataset of cutaneous VOC profiles (after the removal of contaminant peaks, as described in Chapter 7.10.1), consists of 3166 compounds identified by comparison of experimental mass-spectra with reference spectra from the NIST database. The near-zero variance removal function of the pre-processing step removes 2038 variables, and centres and scales 1128 variables. The removal of uninformative variables, in this case 64% of the variables, simplifies the dataset, and reserves only informative predictors for modelling. A PCA can be included in the pre-processing step, which results in a model consisting of 23 principal components accounting for 95% of the variance, compared to the 51 components for 99.92% of the variance for the PCA performed in the preliminary statistical analysis (Chapter 8.3) on the full dataset. Again, this demonstrates the utility of the pre-processing step in eliminating null, or uninformative features from the data.

8.5) Training, tuning and model selection

As discussed in Chapter 2.3.2, optimal predictive performance requires prior tuning of the model parameters. Essentially, tuning involves the assessment of the AUC/ROC for a number of models, constrained by different values of their coefficients/parameters, via in-training resampling, and the selection of the model with the highest AUC/ROC. Each of the three algorithms used in this study were trained and tuned over a range of parameters. Figure 70, Figure 72 and Figure 73 are plots of the AUC/ROC for different parameters of the elastic-net regression, random forest and support-vector machine. The tabulated data is included in Appendix B.4.

For the glmnet (Figure 70), the tuning parameters are alpha (α), or the mixing percentage, which corresponds to the ratio of the ridge-to-lasso penalty type, and lambda (λ), which corresponds to the magnitude of the penalty. A ridge regression penalty is taken as the square of the coefficients, whereas a lasso penalty is taken as the absolute value of the coefficients (Chapter 2.4.1). Consequently, coefficients in a lasso regression can reach zero and become excluded, producing a sparser model, whereas those of a ridge regression can only tend towards zero, and are thus retained. In this case, the optimal model, with an AUC/ROC of 0.66, is a predominantly ridge type ($\alpha=0.05$) with a large penalty ($\lambda=1$). This may imply a large number of outliers or correlated predictors, which can result in coefficients of large magnitude (Chapter 2.2). Notably, the accuracy of models with a high lasso-regression composition (values of alpha greater than 0.5) decreases at higher values of lambda due to models becoming overly sparse from variable reduction. In contrast, the accuracy of models with a greater ridge-type penalty, which retain most or all of the variables, are not as much affected. A pure ridge-regression ($\alpha=0$) demonstrates constant performance across all values of lambda. This may imply that model accuracy is more sensitive to variable reduction than to the magnitude of the coefficients.

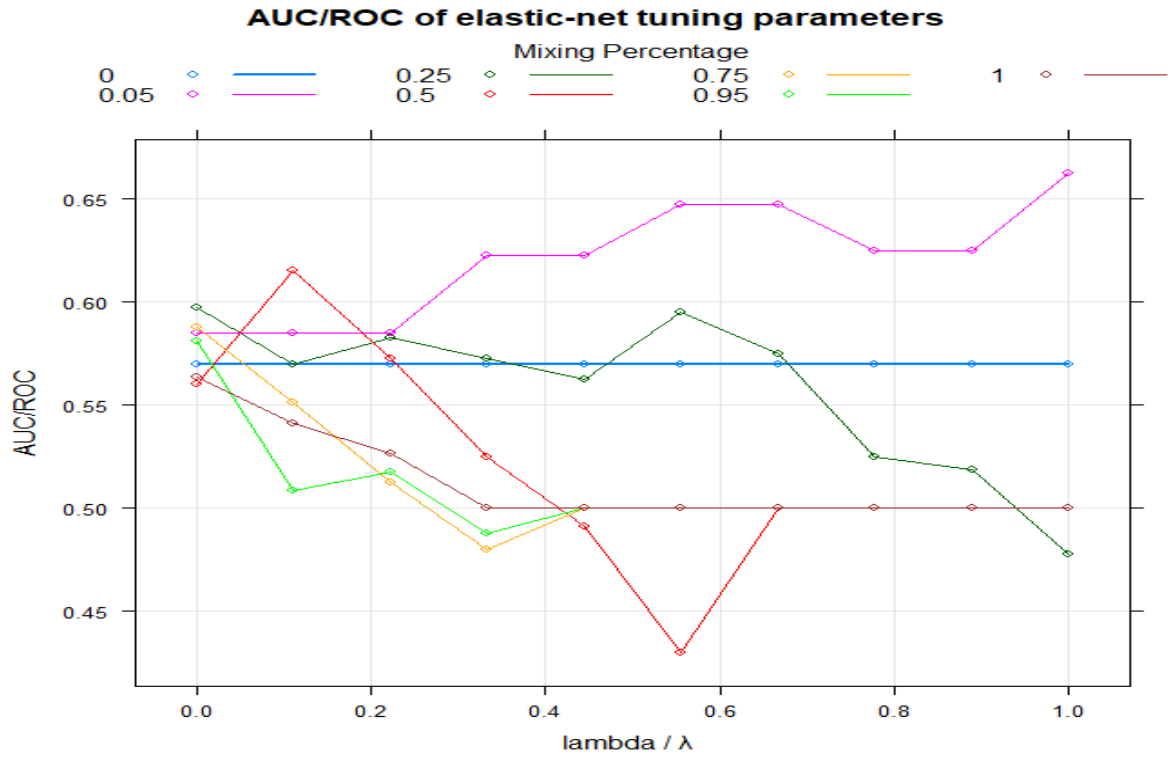


Figure 70: AUC/ROC of the elastic-net tuning parameters; mixing percentage = alpha (α).

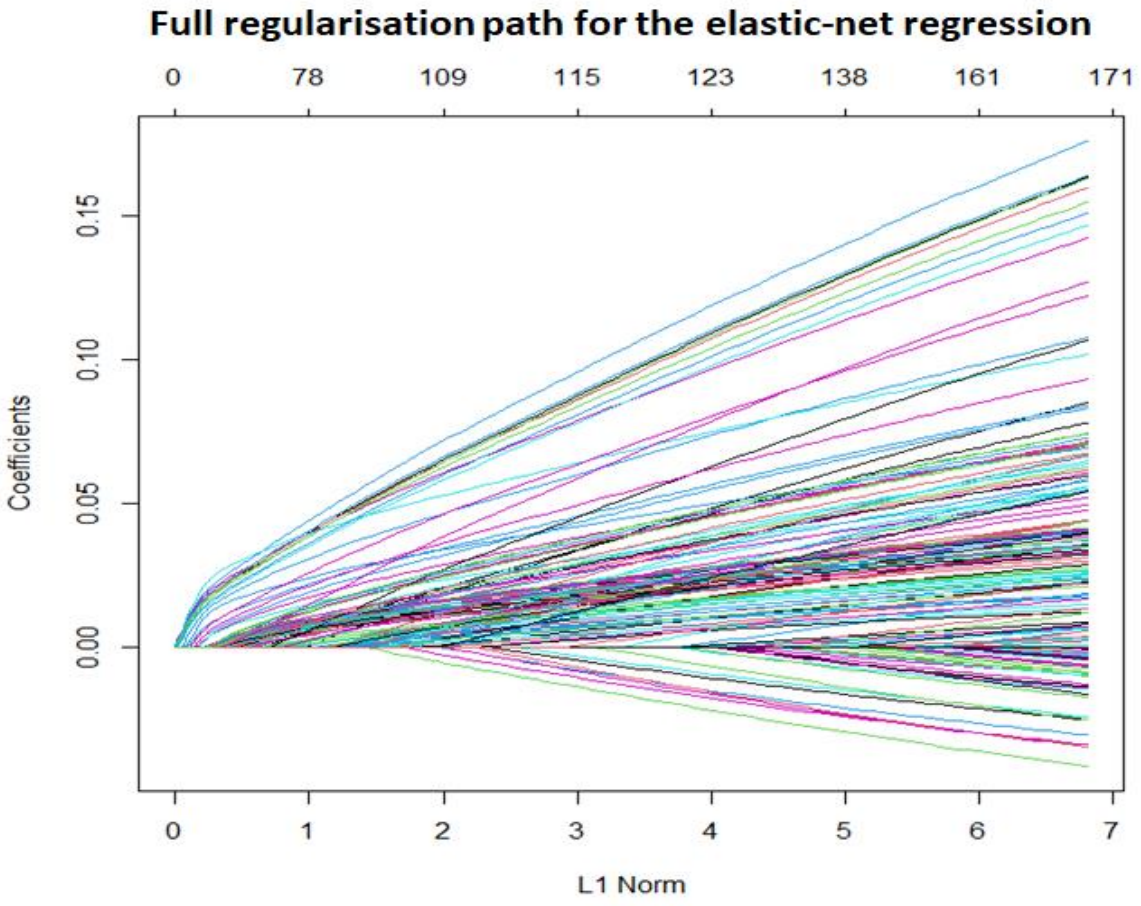


Figure 71: Full regularisation (tuning) path for the elastic-net regression. Model complexity decreases from right to left as the value of lambda (λ) is increased.

Figure 71 plots the full regularisation path of the elastic-net along the L_1 norm. High values of the L_1 norm equate to low values of lambda. An increase in lambda decreases the coefficients, until they reach zero and are excluded from the model. Model complexity thus decreases with an increase in lambda, until, at sufficiently high penalties, the intercept-only model (a regression equation consisting of zero predictors) is reached. Model performance can deteriorate with oversimplification, as reflected by a general drop in the AUC/ROC (Figure 71) as lambda increases. Given the large magnitude of the optimum value of lambda in this case (albeit at a low value of alpha), this could mean that a sparser model yields more accurate predictions, however, the complexity of the sample TICs (Chapter 8.2) and the large number of variables retained (1128) after near-zero variance pre-processing could suggest a large number of outliers and/or correlated predictors necessitating a high regularisation penalty.

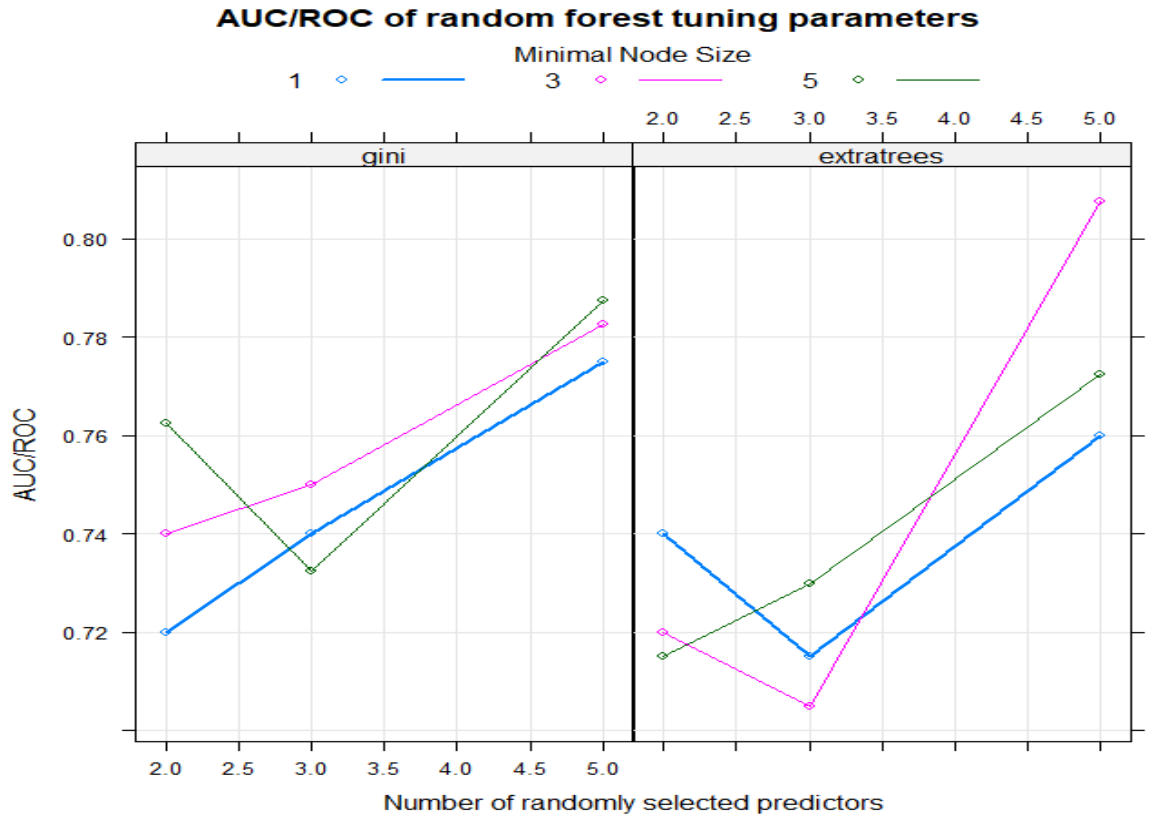


Figure 72: AUC/ROC of the random forest tuning parameters: the splitting rule (gini, left; extratrees, right), the number of randomly selected predictors at each node split, and the minimal node size.

The ranger algorithm for the random forest is defined by three parameters: 1) the splitting rule (gini or extratrees), 2) the minimal number of nodes at each split (minimal node size), and 3) the number of randomly selected predictors for each node (mtry). The results of tuning for different random forests are summarised in Figure 72. The model of optimal AUC/ROC (0.81), by cross-validation, in terms of these specifications, is one split according to the extratrees rule

(although the gini rule appears to produce more stable AUC/ROC values), with a minimal node size of three, and five randomly selected predictors for each node. The large values for the latter two parameters imply that forests of more complex trees produce improved predictions, and that modelling in this case requires multiple features.

The svmPoly algorithm for the support-vector machine has three parameters: the cost (C), the degree of the polynomial function, and the scale of the function (Chapter 2.4.3). Overall, a second-degree polynomial, gives the highest AUC/ROC values, and the optimal model, with an AUC/ROC of 0.78, is a second-degree polynomial with a scale of two, and C=1 (Figure 73). In terms of regularisation, the high value of C is a result analogous to that seen for the elastic-net (Figure 70), where a high value of lambda is selected as optimal. However, a comparable AUC/ROC can be obtained at a low value of C for a polynomial of the same degree but at a higher scale (Figure 73).

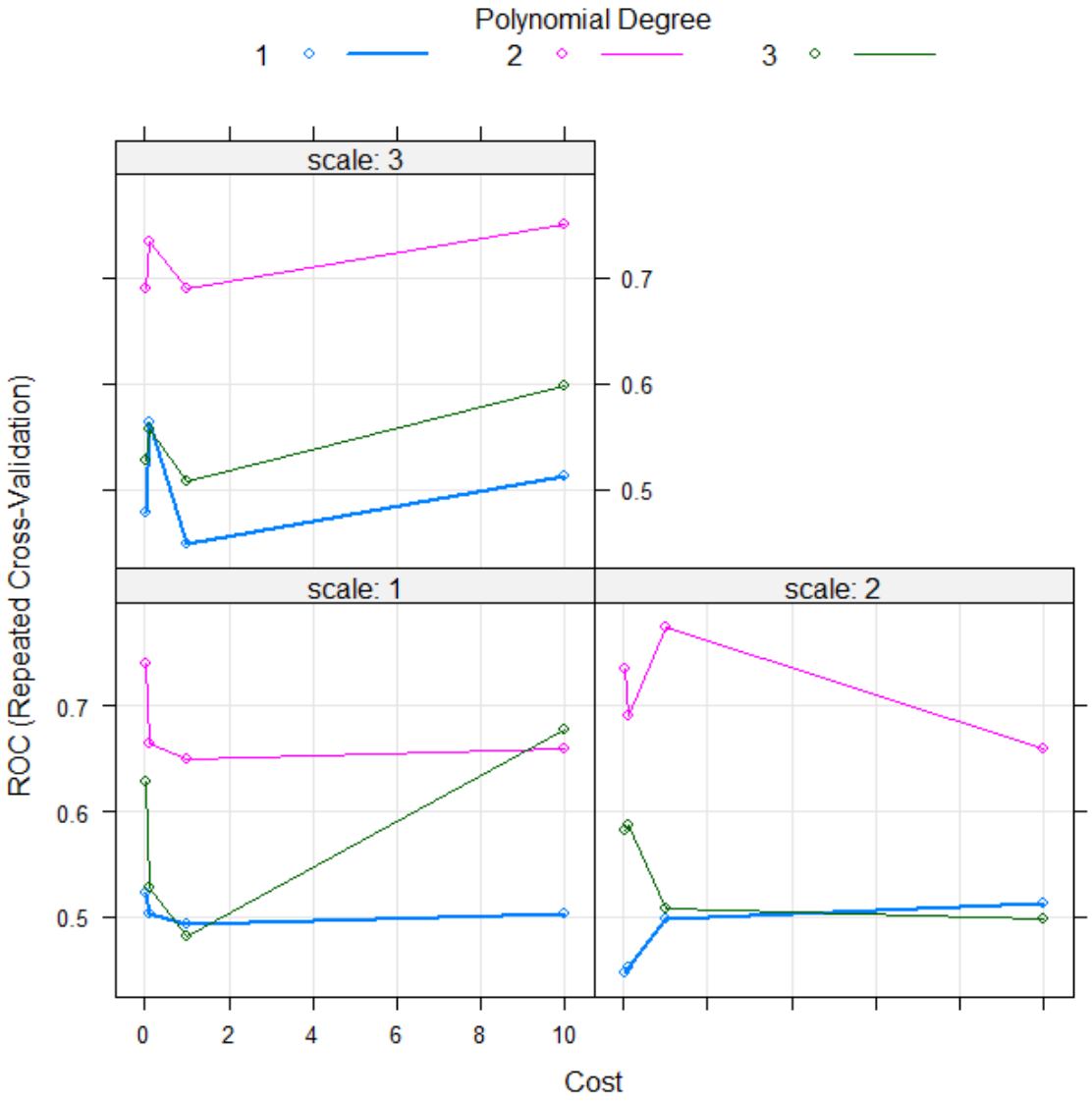


Figure 73: AUC/ROC of the support-vector machine (polynomial kernel) tuning parameters: the polynomial degree, the scale and cost (C).

8.6) Testing and prediction

In order to assess the predictive accuracy of a model, it must be applied to unknown observations in a testing set. In other words, it must be able to predict the category of samples that it was not trained on, and that were not used as input in its construction. Prediction involves inputting sample data, and calculating the probability that the given sample falls under a given category, in this case negative or positive for malaria-status. Table 5 lists the probabilities and corresponding predicted categories for each of the 26 samples in the testing set. The predictions of each model are summarised in a confusion matrix (Figure 74) with corresponding performance statistics.

Table 5: computed probabilities of patient samples in the testing set being either malaria-positive or -negative, for the elastic-net, random forest and support-vector machine.

	Elastic-net				Random forest			Support-vector machine		
	Status probability		Predicted status	Actual status						
Sample	NEG	POS			NEG	POS		NEG	POS	
10i	0,9027	0,0973	NEG	NEG	0,8766	0,1234	NEG	0,8574	0,1426	NEG
2ii	0,1077	0,8923	POS	NEG	0,8107	0,1893	NEG	0,3032	0,6968	POS
20i	0,9364	0,0636	NEG	NEG	0,8704	0,1296	NEG	1,0000	0,0000	NEG
4i	0,6775	0,3225	NEG	POS	0,8502	0,1498	NEG	0,3362	0,6638	POS
7ii	0,7762	0,2238	NEG	NEG	0,8692	0,1308	NEG	0,7881	0,2119	NEG
8ii	0,8927	0,1073	NEG	NEG	0,8767	0,1233	NEG	0,8463	0,1537	NEG
24i	0,7774	0,2226	NEG	NEG	0,8487	0,1513	NEG	0,8296	0,1704	NEG
9ii	0,9276	0,0724	NEG	NEG	0,8678	0,1322	NEG	0,8702	0,1298	NEG
15ii	0,9224	0,0776	NEG	NEG	0,8552	0,1448	NEG	0,8513	0,1487	NEG
12ii	0,8941	0,1059	NEG	NEG	0,8613	0,1387	NEG	0,8828	0,1172	NEG
8i	0,6215	0,3785	NEG	NEG	0,8764	0,1236	NEG	0,7792	0,2208	NEG
18iii	0,4847	0,5153	POS	POS	0,8032	0,1968	NEG	1,0000	0,0000	NEG
2i	0,8900	0,1100	NEG	NEG	0,8111	0,1889	NEG	0,8555	0,1445	NEG
18ii	0,8883	0,1117	NEG	POS	0,7810	0,2190	NEG	0,9269	0,0731	NEG
1i	0,8836	0,1164	NEG	NEG	0,8190	0,1810	NEG	0,8282	0,1718	NEG
24ii	0,8222	0,1778	NEG	NEG	0,8447	0,1553	NEG	0,8155	0,1845	NEG
16ii	0,9318	0,0682	NEG	NEG	0,8573	0,1427	NEG	0,8533	0,1467	NEG
21ii	0,9368	0,0632	NEG	NEG	0,8466	0,1534	NEG	0,8326	0,1674	NEG
25ii	0,9273	0,0727	NEG	NEG	0,8331	0,1669	NEG	0,9877	0,0123	NEG
17iii	0,8470	0,1530	NEG	POS	0,8083	0,1917	NEG	0,9694	0,0306	NEG
6ii	0,8951	0,1049	NEG	NEG	0,8527	0,1473	NEG	0,8905	0,1095	NEG
13ii	0,9205	0,0795	NEG	NEG	0,8746	0,1254	NEG	0,9044	0,0956	NEG
22ii	0,9165	0,0835	NEG	NEG	0,8600	0,1400	NEG	0,8805	0,1195	NEG
12i	0,9244	0,0756	NEG	NEG	0,8877	0,1123	NEG	0,8564	0,1436	NEG
23ii	0,9243	0,0757	NEG	NEG	0,8636	0,1364	NEG	0,8633	0,1367	NEG
6i	0,8207	0,1793	NEG	NEG	0,8372	0,1628	NEG	0,8231	0,1769	NEG

		ACTUAL					
		POS	NEG	POS	NEG	POS	NEG
PREDICTED	POS	1	1	0	0	1	1
	NEG	3	21	4	22	3	21

Accuracy	0.8462	0.8462	0.8462
95% CI	(0.6513, 0.9564)	(0.6513, 0.9564)	(0.6513, 0.9564)
No information rate	0.8462	0.8462	0.8462
P-value (Acc > NIR)	0.6293	0.6293	0.6293
McNemar's Test P-Value	0.6171	0.1336	0.6171
Sensitivity	0.25000	0.0000	0.25000
Specificity	0.95455	1.0000	0.95455
Positive predictive value (PPV)	0.5000	N/A	0.5000
Negative predictive value (NPV)	0.87500	0.8462	0.87500
Prevalence	0.15385	0.15385	0.15385
Detection rate	0.03846	0.0000	0.03846
Detection prevalence	0.07692	0.0000	0.07692
Balanced accuracy	0.60227	0.5000	0.60227
	A) EN	B) RF	C) SVM

Figure 74: Confusion matrix and associated statistics for the predictions of the tuned models: A) elastic-net (EN), B) random forest (RF), and C) support-vector machine (SVM). POS = malaria-positive; NEG = malaria-negative.

The no-information rate (NIR) is equivalent to the percentage of samples in the majority category (in this case, malaria-negative), or in other words, the percentage of true negatives and false positives in the testing set. The NIR can be interpreted as the predictive accuracy obtained by a null model that classifies every sample as negative. For a model to be considered predictive, it must have a higher predictive success rate than is obtained from null prediction, which is to say, the accuracy of the model must be greater than the NIR [3]. Since there are only a few positive cases, the NIR in this case is high (84.6%). The accuracy of each model is equal to the NIR, and thus these models are non-predictive. The p-values (for the accuracy being greater than the NIR) are greater than 0.05 for all cases, indicating a good probability of model predictions being due to chance, especially given the large imbalance in the proportion of negative to positive samples.

The key statistics are the sensitivity and the specificity (Figure 74) [4]. The sensitivity is the proportion of true positive cases of those predicted to be positive, and is thus a measure of the ability of the model to correctly classify malaria-positive cases. The sensitivity is 0% for the random forest, and 25% for the elastic-net and support-vector machine, both of which make

one correct true-positive classification. Related to the sensitivity is the detection rate (the proportion of true positives of all samples in the testing set) and the detection prevalence (the percentage of true and false positives), which in this case is 0% for the random forest, and very low for the elastic-net and support-vector machine. Since the sensitivity is zero for the random forest, its positive predictive value (PPV) is undefined.

The specificity is the proportion of the true negative cases of all the cases predicted to be negative, and in this case measures the ability of the model to correctly classify malaria-negative cases. The specificity, and consequently the negative predictive value (NPV) for each model is high, but since the sample population is predominantly malaria-negative (giving a high NIR) this statistic is not useful.

8.7) Variable importance ranking and feature selection

The final step of the machine learning pipeline is to identify the top-ranking model variables, although due to the poor predictive power of the models in this case, these cannot be considered to be key predictors, or features, characteristic of the categorical types in question. The algorithms of the caret package have inbuilt functions available for the ranking of variable importance (Chapter 2.3.4). The ranking functions of the elastic-net and the random forest are based on the parameters of the models themselves. For the elastic-net regression, the absolute values of the regression coefficients are used to rank predictors. For the random forest, normalised mean differences in the accuracy of predictors, determined by bootstrapping, are computed. The variable importance function of the support-vector machine is not based on the model itself, but determines the AUC/ROC for each predictor over a range of thresholds [4].

Table 6 lists the twenty top-ranking compounds for the 3 models, as well as their scaled scores[8]. A wide variety of tentatively identified compounds are selected as top features, including mid- to long-chain (C_{10}-C_{18}) saturated and unsaturated fatty acids; unsaturated alcohols, esters, aldehydes and ketones and alkylbenzenes. In addition, there are a number of nitrogenous compounds, including a pyrazole amine (1H-pyrazol-3-amine); a nitrile (isoamyl cyanide); an azole (isothiazole); amides and substituted pyridines.

[8] The full scaled-score outputs for the pre-processed set of 1131 variables are available upon request.

Table 6: List of the top predictor compounds for the elastic-net, random forest and support vector machine models.

Elastic-net		Random forest		Support-vector machine	
Predictor compound	**Scaled score**				
2-Octen-1-ol, (E)-	100	Isoamyl cyanide	100	Dicyclohexyl phthalate	100
2-Hexenoic acid, 3,4,4-trimethyl-5-oxo-, (Z)-	92,06	Dodecanoic acid	99,96	Benzoic acid, 2-hydroxy-, pentyl ester	84,21
1H-Pyrazol-3-amine	90,8	2-Hexenoic acid, 3,4,4-trimethyl-5-oxo-, (Z)-	95,64	2-Octen-1-ol, (E)-	82,89
n-Decanoic acid	89,69	n-Decanoic acid	76,54	Tetracosane	78,95
2-Methylbutanoic anhydride	89,58	Formic acid, 2-propenyl ester	63,88	2-Piperidinone	78,95
Isothiazole	87,14	Oxalic acid, isobutyl pentyl ester	52,14	Undecane, 2,6-dimethyl-	77,63
2-Piperidinone	83,93	2-Octen-1-ol, (E)-	51,45	2,4-Nonadienal	77,63
Isoamyl cyanide	83,67	9-Hexadecenoic acid	50,77	Pentadecanoic acid	76,32
Dodecanoic acid	81,11	2-Methylbutanoic anhydride	49,8	Propanamide	73,68
Oxalic acid, isobutyl pentyl ester	71,95	Phthalic acid, 6-ethyl-3-octyl butyl ester	47,21	Benzene, (2-methylpropyl)-	73,68
1-Nonen-3-ol	64,84	Ethanone, 1-(2-methyl-1-cyclopenten-1-yl)-	45,82	Pyridine, 2,4-dimethyl-	71,05
Formic acid, 2-propenyl ester	64,21	2,4-Decadienal, (E,E)-	43,81	Butanamide, 3-methyl-	68,42
Propanoic acid, ethenyl ester	55,42	Pyridine, 3-ethyl-4-methyl-	43,32	Z-10-Pentadecen-1-ol	68,42
Ethanone, 1-(2-methyl-1-cyclopenten-1-yl)-	50,88	Isothiazole	43,13	6-Methyl-3,5-heptadiene-2-one	67,11
1,2-Ethanediamine, N,N-diethyl-	47,72	Indene	43	Cyclodecane	65,79
Naphthalene, 1,2,3,4-tetrahydro-1,4-dimethyl-	46,44	Butane, 2,2-dimethyl	42,02	Dimethyl ether	65,79
9-Hexadecenoic acid	44,74	2-Piperidinone	40,83	Isopropyl acetate	65,79
Cyclodecane	39,94	Pentadecane, 2,6,10,14-tetramethyl-	40,42	Naphthalene, 2,3,6-trimethyl-	63,16
3-Pentanol, 2,4-dimethyl,	39,75	Pyridine, 2,4-dimethyl-	39,39	Cyclopentaneacetic acid, 3-oxo-2-pentyl-, methyl ester	61,84
Propanal	37,83	Pentadecanoic acid	38,68	Benzene, propyl-	60,53

The compounds with the highest scores are (E)-2-octen-1-ol, isoamyl cyanide and dicyclohexyl phthalate. The latter is an industrial product, and thus likely to be a contaminant. The C_8 unsaturated alcohol (E)-2-octen-1-ol is the only top predictor in Table 6 listed for all three algorithms, with scaled scores of 50 to 100. Another unsaturated alcohol of similar chain length, 1-nonen-3-ol, is listed for the elastic-net. These compounds have not been previously reported in the context of malaria-marker investigations, however, another alcohol, 2-ethyl-1-hexanol, shown in Figure 55 and Figure 57, for positive cases #4i and #17i, has been reported as a putative marker of malaria-status, specifically for febrile children with diarrhea [2]. However, this compound was found also to occur with comparable relative abundance in malaria-negative cases.

An unsaturated C_8 ketone, 6-methyl-3,5-heptadiene-2-one, listed for the support-vector machine, is similar to the C_8 ketone, 2-octanone, which has previously been reported to be associated with the presence of microscopic gametocytes [5].

The unsaturated C_9 aldehyde, 2,4-nonadienal is ranked highly by the support-vector machine, and is an unsaturated counterpart to nonanal, which has been found in elevated amounts in individuals infected with *Plasmodium* [5], but has also been found to be associated with underlying conditions and other diseases in malaria-free cases [2].

Two alkylbenzenes — (2-methylpropyl)-benzene and propyl-benzene — are also listed by the support-vector machine, and are similar to other alkylbenzenes that have been reported as compounds associated with malaria-infection, such as toluene, 1-ethyl-3-methyl- ethylbenzene and 1,2,4-trimethyl-benzene [2]. Alkylbenzenes, however, are likely industrial contaminants.

The compounds (E)-2-octen-1-ol; 2-methylbutanoic anhydride; (E,E)-2,4-decadienal; isothiazole and isoamyl cyanide are indicated on the chromatograms in Figure 57 and Figure 59.

8.8) Retention indices of top-ranking compounds

Median retention indices (RIs) for selected top-ranking compounds were calculated from a homologous series of n-alkanes (C_6-C_{28}), as described in Chapter 7.9, and are reported in Table 7, along with the unique CAS number[9] for the tentatively identified compound, the mass spectral similarity to the NIST reference spectrum, and the scaled variable importance score for each of the three machine learning models. RIs were calculated from the equation of Kovats retention index, as well as from a least-squares linear regression equation of the linear retention indices of reference n-alkanes.

Experimental values indicated with a cross symbol (†) are within ±50 units of the literature values. These include: (E)-2-octen-1-ol; pentadecanoic acid; 2-methylbutanoic anhydride; 2-hydroxy-benzoic acid, pentyl ester; formic acid, 2-propenyl ester; propanoic acid, ethenyl ester; 1,2,3,4-tetrahydro-1,4-dimethyl-naphthalene; 2,4-nonadienal; 6-methyl-3,5-heptadiene-2-one; 3-oxo-2-pentyl-cyclopentaneacetic acid, methyl ester; 2,5-dimethyl undecane; (2-methylpropyl)-benzene; propyl-benzene; isoamyl cyanide; 3-methyl butanamide; 2-piperidinone; 3-ethyl-4-methyl-pyridine and 2,4-dimethyl-pyridine. For some compounds (indicated with an asterisk[*]), such as Z-10-Pentadecen-1-ol, the experimental RI is not available, and estimated values are cited.

The retention times for some of the compounds with high variable scores, including N,N-diethyl-1,2-ethanediamine; dicyclohexyl phthalate; dimethyl ether and isopropyl acetate fall outside of the retention time range of the reference n-alkane series, and thus are reported as <800. Note that due to the necessity of replacing the 1D column during the analysis (Chapter 7.8), two values are applicable in most instances, as indicated in Table 7.

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

Table 7: Median retention indices (RI) and variable importance scores of selected top predictor compounds of malaria-status. 1D RI values in red are for samples run on the first 1D column, prior to replacement (c.f.: Chapter 7.8). EN = elastic-net; RF = random forest; SVM = support vector-machine. R^2 (first GC column) = 0.9934; R^2 (second replacement GC column) = 0.9938.

Tentative identification	CAS number	Molecular formula	Chemical class	MW (g/mol)	1D time (m)	2D time (s)	1D RI Kovats (nonpolar)	1D RI Regression (nonpolar)	RI (lit.) nonpolar (NIST)	MS similarity match
2-Octen-1-ol, (E)-	18409-17-1	C8H16O	Unsaturated alcohol	128	528	0,94	1039†	1073†	1055; 1052	750-825
					532,5	0,94	1040†	1072†		
1-Nonen-3-ol	21964-44-3	C9H18O	Unsaturated alcohol	142	441	0,89	891	958	1058	757-851
					444	0,89	986	953		
Z-10-Pentadecen-1-ol	245485 *NIST	C15H30O	Unsaturated alcohol	226	795	1,09	1289	1424	1763 *estimated	758-827
					1191	1,18	1870	1957		
					1260	1,48	2021	2037		
n-Decanoic acid	334-48-5	C10H20O2	Saturated fatty acid	172	795	1,11	1425	1425	1344; 1350	752-934
					813	1,14				
					1194	0,6	1871	1961		
					1203	0,78	1882	1962		
Dodecanoic acid	143-07-7	C12H24O2	Saturated fatty acid	200	1140	1,07	1829	1879		
					1251	0,86	2018	2038	1556; 1554	752-943
Pentadecanoic acid	1002-84-2	C15H30O2	Saturated fatty acid	242	1137	0,98	1827†	1885†	1848	762-900
					1461	1,09	2277	2301		
9-Hexadecenoic acid	2091-29-4	C16H30O2	Unsaturated fatty acid	254	1248	1,72	2012	2021	1916,29; 1924	751-774
					1830	1,31	2532	2787		
2-Hexenoic acid, 3,4,4-trimethyl-5-oxo-, (Z)-	14919-56-3	C9H14O3	Keto acid	170	660	1,14	1208	1247	1309	752-754
2-Methylbutanoic anhydride	1468-39-9	C10H18O3	Acid anhydride	186	633	1,57	1090	1207†	1190 *estimated	755-856
					613,5	1,53	1084	1185†		
Oxalic acid, isobutyl pentyl ester	309370 *NIST	C11H20O4	Oxalic acid ester	216	816	1,01	1298	1453	1385*estimated	782-848
Benzoic acid, 2-hydroxy-, pentyl ester	2050-08-0	C12H16O3	Benzoic acid ester	208	939	1,31	1498	1618†	1574; 1579	750-955
					958,5	1,4	1495	1640		
Formic acid, 2-propenyl ester	1838-59-1	C4H6O2	Unsaturated ester/aldehyde	86	486	1,395	1015†	1017†	1014	1014
					487,5	1,4	1014†	1011†		
Propanoic acid, ethenyl ester	105-38-4	C5H8O2	Unsaturated ester/aldehdyde	100	435	1,24	1057†	940	1092	775-777
					561	1,53	1218	1116†		
Ethanone, 1-(2-methyl-1-cyclopenten-1-yl)-	3168-90-9	C8H12O	Cyclic unsaturated ketone	124	570	1,17	1062	1128	1504	769
					570	1,11	1060	1122		
Naphthalene, 1,2,3,4-tetrahydro-1,4-dimethyl-	4175-54-6	C12H16	Naphthalene derivative	160	774	1,185	1275†	1396	1283,1; 1258	759-779
2,4-Nonadienal	6750 03 04	C9H14O	Unsaturated aldehyde	138	663	1,11	1210†	1251	1200; 1169	829-901
					664,5	1,06	1207	1249		
2,4-Decadienal, (E,E)-	25152-84-5	C10H16O	Unsaturated aldehyde	152	750	1,17	1263	1365		
					751,5	1,13	1262	1366	1288; 1291	763-929

Tentative identification	CAS number	Molecular formula	Chemical class	MW (g/mol)	1D time (m)	2D time (s)	1D RI Kovats (nonpolar)	1D RI Regression (nonpolar)	RI (lit.) nonpolar (NIST)	MS similarity match
6-Methyl-3,5-heptadiene-2-one	1604-28-0	C8H12O	Unsaturated ketone	124	562,5	1,105	1058†	1118†	1074,9	772-932
					567	1,095	1058†	1118†		
Cyclopentaneacetic acid, 3-oxo-2-pentyl-, methyl ester	24851-98-7	C13H22O3	Cyclic ketone ester	226	1018,5	1,52	1649†	1725	1648; 1649	757-861
					1020	1,595	1641†	1721		
Undecane, 2,5-dimethyl-	17301-22-3	C13H28	Saturated hydrocarbon	184	705	0,805	1233†	1304	1210; 1218	752-904
Cyclodecane	293-96-9	C10H20	Saturated macrocycle	140	468	0,76	1004	994	1127; 1156	760-915
					504	0,8	1024	1033		
Benzene, (2-methylpropyl)-	538-93-2	C10H14	Alkylbenzene	134	480	0,91	1009†	1001†	1001; 997	752-913
					495	0,92	1020†	1029†		
Benzene, propyl-	103-65-1	C9H12	Alkylbenzene	120	418,5	0,91	882	929†	944,4; 933,5	753-948
					426	0,9	882	928†		
1H-Pyrazol-3-amine	1225387-53-0	C3H5N3	Pyrazole derivative	83	472,5	1,41	1005	991	1064* estimated	763-783
					477	1,35	1009	1006		
Isoamyl cyanide	542-54-1	C6H11N	Nitrile	97	306	0,975	823†	<800	814; 848	765-806
					310,5	0,96	818†	<800		
Propanal	123-38-6	C3H6O	Aldehyde	58	135	1,31	<800	<800	472; 468	751-828
					132	0,55	<800	<800		
Propanamide	79-05-0	C3H7NO	Amide	73	361,5	1,105	854	853		
					364,5	1,11	850	846	984	768-904
Butanamide, 3-methyl-	541-46-8	C5H11NO	Amide	101	475,5	1,185	1009†	1004†		
					477	1,165	1007†	997†	1018,5	757-850
Isothiazole	288-16-4	C3H3NS	Azole	85	774	1,035	1277	1397	743* estimated	768-775
2-Piperidinone	675-20-7	C5H9NO	Piperidine ketone	99	649,5	1,635	1147†	1229	1173,5; 2060	758-922
					651	1,63	1202†	1235		
Pyridine, 3-ethyl-4-methyl-	529-21-5	C8H11N	Alkylpyridine	121	489	1,05	1017†	1021†	1011	759-889
					493,5	1,06	1017†	1019†		
Pyridine, 2,4-dimethyl-	108-47-4	C7H9N	Alkylpyridine	107	396	1,02	871†	899 †	903,9; 916	753-875
					403,5	1,025	871†	898†		

†RI value falls within ±50 units of one and/or two of the literature values.

8.9) Relative abundance of top-ranking compounds

The normalised peak area values of the top-ranking compounds are plotted as a heatmap, in Figure 75, representing their relative abundance across all replicate samples. Since the final models have no predictive value, the suite of top compounds is not reliable in distinguishing malaria-positive from malaria-negative cases, and in many cases the compounds are present in samples from both. However, the following compounds are present at overall greater relative abundance in samples of positive cases: (E)-2-octen-1-ol; 9-hexadecenoic acid; *n*-decanoic acid; 9-hexadecenoic acid; (Z)-3,4,4-trimethyl-5-oxo-2-hexenoic acid; 2-methylbutanoic anhydride; (E,E)-2-2,4-decadienal; propenyl ester formic acid; 1-(2-methyl-1-cyclopenten-1-yl)-ethanone; 1,2,3,4-tetrahydro-1,4-dimethyl-naphthalene; N,N-diethyl-1,2-ethanediamine; isothiazole and isoamyl cyanide. The collective presence of these compounds could potentially indicate *Plasmodium* infection. Compounds of a higher carbon number than C_{10}-C_{12} are less volatile, however, they may still potentially act as kairomones to Anopheline vectors that determine, at a proximal distance to the host, whether or not the mosquito will initiate landing and blood-feeding (c.f.: Chapter 6B.3) [6]. Notably, the compound *n*-decanoic and 9-hexadecenoic acid (present at a comparatively higher relative abundance in malaria-positive patient #17 than in the negative participants) is a saturated carboxylic acids with rancid odours reminiscent of foot malodour, a scent profile which has been found to enhance host attraction to female Anopheline vectors [7, 8]. In addition, 2-methylbutanoic anhydride (also present at greater relative abundance in #17) is the anhydride of 2-methylbutanoic acid, which is an isomer of 3-methylbutanoic acid (isovaleric acid)— a known component of foot malodour [9]. The presence of these compounds may thus be associated with enhanced vector attraction in relation to malaria-infection status, and are thus also potential markers of malaria-infection.

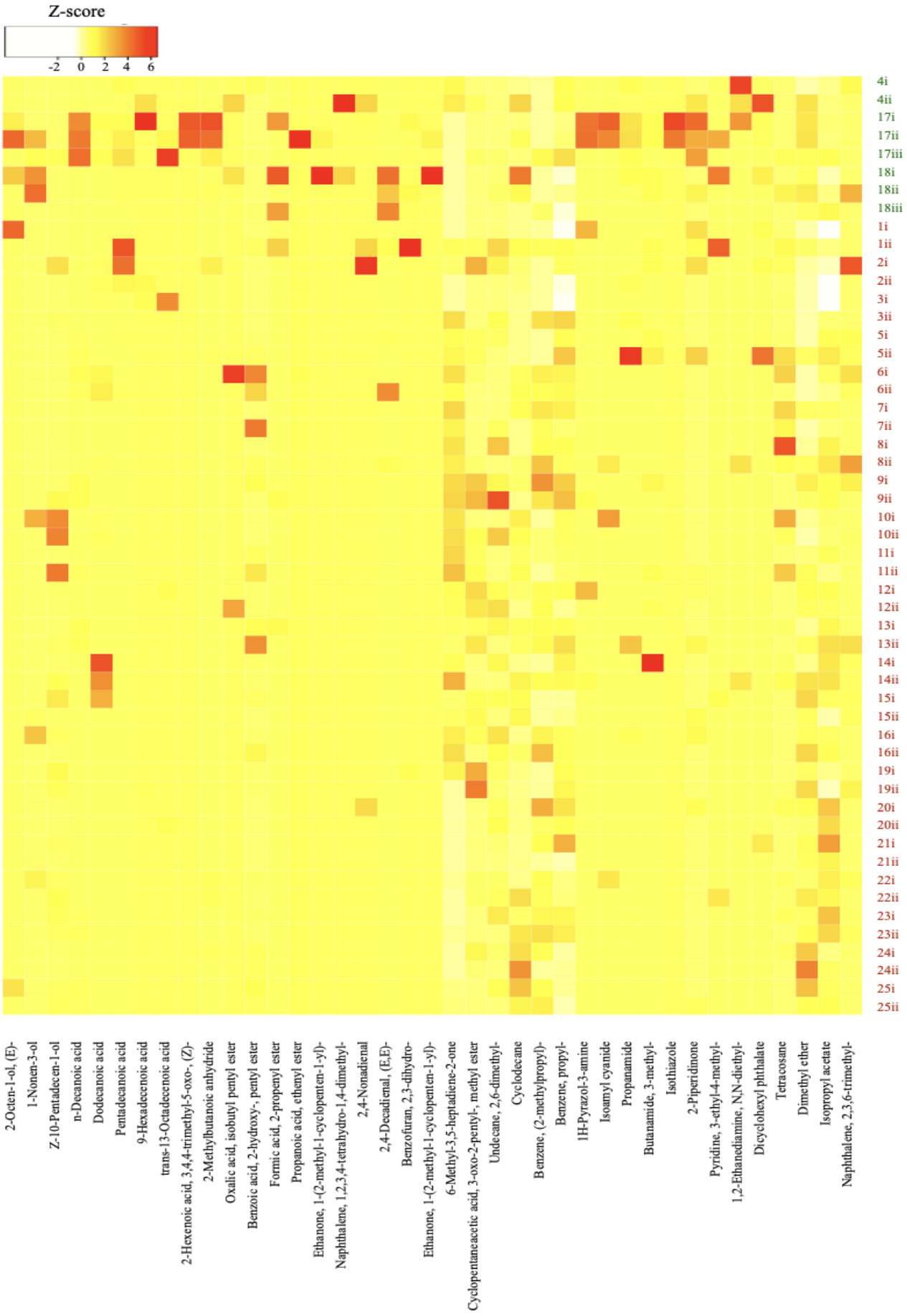


Figure 75: Heatmap of sample-wise relative abundance of the compounds ranked as top variables by machine learning (Table 7). Green = malaria-POS; red = malaria-NEG. The colour intensity scale depicts the relative abundance Z-score.

8.10) Targeted standard analysis

A targeted analysis for compounds previously reported to be associated with *Plasmodium* infection was performed using external standards (Chapter 7.7). These compounds (Chapter 6B.4) include heptanal, (E)-2-octenal, 2-octanone, octanal, nonanal and (E)-2-decenal [5]. Masses (ng) of these were estimated for the malaria-positive cases using least-squares linear regression, the results of which are summarised in Table 8. Five malaria-negative cases (#3, #6, #13, #24 and #25) were chosen randomly as controls. After the removal of outliers (by inspection), higher R^2 values were obtained.

The limit of detection (LOD) and quantification (LOQ) for each compound was calculated using the signal-to-noise of the lowest-concentration standard (S/N of 3 for the LOD, and S/N of 10 for the LOQ)[10].

Overall, there is little difference between the malaria-positive and -negative cases in terms of quantified masses, although octanal is found at a higher mass (7.3 ng) in #17 than the malaria-negative cases. This is not true for the other two malaria-positive cases (#4 and #18). However, these two patients were found to be positive only by RDT, and not by RDT and microscopy (Chapter 8.1).

The retention times for the samples are shifted (albeit within overall agreement) compared to those of the standards, which may be due to aging of the column in the time period between the analysis of the samples and the standards (instrumental maintenance after failure of the turbomolecular pump of the TOF mass spectrometer occurred prior to the standard analysis).

[10] Values for the LOD and LOQ calculated using the standard error and the intercept from least-squares linear regression were found to be unreasonably high.

Table 8: Limit of detection (LOD), limit of quantification (LOQ), retention times (RT), R^2 values, and interpolated masses (from least-squares linear regression) of targeted standards (simulated matrix matched calibration) for malaria-positive and -negative patients. Masses are in units of nanograms (ng). Retention times are in units of seconds (s). Grey blocks indicate that the particular compound is not present (<LOD) in the sample. Pink blocks indicate samples in which blank quantities were higher than sample quantities.

Target compound	LOD (ng)	LOQ (ng)	Malaria-positive			Malaria-negative					1D Retention time (s)		Mass range (ng)	Linear regression equation	R^2
			#4	#17	#18	#3	#5	#6	#13	#24	#25	Standards	Samples		$y = mx + c$	
Heptanal	0,0041	0,014	14,9	11,2	9,3					15,4	17,3	360; 363; 366	363; 369; 372	2,5-60	y = 3E+06x + 7E+07	0,9601
(E)-2-Octenal	0,0063	0,021	4,0		4,7	2,2	3,5	3,8				513; 516	510; 519; 522	2,5-60	y = 5E+06x + 2E+07	0,9871
2-Octanone	0,0004	0,0014		25,3	23,8	23,4						447; 450	453; 456	5-60	y = 8E+06x + 2E+08	0,9275
Octanal	0,0025	0,0085	0,9	7,3	1,8	0,9	0,9	1,5	0,4	2,6	1,1	459; 462	459; 465; 468; 471	2,5-60	y = 1E+07x + 5E+07	0,9476
Nonanal	0,0014	0,0045	2,9	15,3	10,0	15,1	12,8	11,0	17,4	18,1	18,8	555; 558	561; 564; 567	2,5-60	y = 5E+06x + 1E+08	0,9699
(E)-2-Decenal	0,0025	0,0083	1,8	0,8	1,0	1,7	0,4	5,0	0,7	1,7		696; 699	702; 705; 708	2,5-60	y = 4E+06x + 7E+06	0,9835

8.11) Limitations of the study and future considerations

There are two major limitations to the study which would have to be addressed in future investigations. The first is the small number of malaria-positive participants, which significantly limits the reliability of the results. In practice, a balanced positive/negative sample size may not always be possible, however, there should be a sufficient positive representation for the statistical models and predictions to be meaningful. This is particularly important considering the splitting of the dataset into training and testing sets, which should each contain enough positive samples for predictions to be statistically meaningful, and for the NIR (Chapter 8.6) not to be much greater than 50-60%. Fluctuations in predictive accuracy are exaggerated with smaller sample sizes, since changes in the random training-test splits can alter the composition of the splits, which may in turn affect the predictive accuracy (by affecting which variables are included in the final model).

The second limitation is the lack of PCR testing for the confirmation of malaria-status. Though microscopic assay on collected peripheral blood smears was performed (Chapter 7.5), PCR is the most sensitive and specific diagnostic assay (Chapter 6A.2.2), and is useful in detecting potential false negative or positive results by RDT, or by clarifying conflicting results, as in the case of participant #18, who tested negative by the RDT administered by the on-site investigator, and by microscopic assay, but positive by the RDT administered by the clinic staff (Chapter 8.1).

A future study could investigate potential underlying differences between the VOC profiles of malaria-positive patients with asexual gametocytes and positive patients without gametocytes, using quantitative nucleic acid sequence-based amplification (QT-NASBA) [5, 10]. In terms of machine learning, this could be achieved with multiclass modelling and prediction, or by parallel dual-class modelling, whereby the case of malaria-positive/negative is modelled and tested on the one hand, and that of malaria-positive with/without gametocytes is tested on the other hand.

8.12) Conclusion

The cutaneous VOCs of malaria-negative and -positive individuals visiting two local clinics in Limpopo province, South Africa, were extracted using non-invasive PDMS sampling loops

adhered to the surface of the epidermis. These samplers are small, light-weight and portable, and thus easy to apply by trained non-medical personnel. The samples were analysed by GC×GC-TOFMS, which was used to construct individual cutaneous VOC profiles for machine learning, in an attempt to construct predictive models of malaria-status, and to identify potential markers of infection by the *Plasmodium* parasite.

The full set of variables, after blank correction, consists of a high dimensional set of 3166 compounds tentatively identified by mass-spectral comparison of experimental and reference spectra. The removal of near-zero variance predictors prior to machine learning reduces the dimensionality of the data set by 64.37%, and this pared dataset of 1128 variables can be used as input to construct machine learning models.

The cutaneous VOC profiles, obtained by comprehensive GC×GC-TOFMS, are complex, showing dense peak distribution along the 2D separation space. A wide variety of organic species ($C_{\pm 3}$-$_{30}$), of all major functional groups, are present on the surface of the epidermis. Though there is inter-individual variation in the VOC profiles, the chromatograms across all samples have similar peak patterns.

PCA and LDA were performed on the full dataset as a preliminary assessment of the variation in the data. For PCA, one point for a malaria-positive case falls outside of the cluster of points for negative cases, but there is otherwise little distinction between points of the two categories.

Three machine learning algorithms (an elastic-net regression, a random forest and a support-vector machine) were used to construct models of the training data and to make predictions on the testing data. Optimised models were selected by cross-validation. The models have no predictive value, showing poor accuracy and sensitivity.

The top-ranking variables listed by each of the models include C_{3}-$_{21}$ compounds of a wide variety of chemical classes. Among these notably, are unsaturated and oxygenated C_{6-8} species, including alcohols, ketones and aldehydes, similar to compounds previously reported [2, 5]. Notably, the unsaturated alcohol, 2-octen-1-ol, (E)-, is ranked highly by all three algorithms, and is present at significantly greater relative abundance for the malaria-positive cases. A group of C_{10-18} saturated and unsaturated fatty acids are also among the top variables. Compounds such as *n*-decanoic acid, 9-hexadecenoic acid and 2-methylbutanoic anhydride (associated with rancid malodour) may have potential use as markers of infection, and though such carboxylic acids are semi-volatile, they may nevertheless function as proximal-distance kairomones to Anopheline vectors, and further investigation of their potential role in this regard is thus justified.

RI values for the top predictors, where applicable, were calculated from a homologous series of n-alkanes (C_6-C_{28}). For a number of the compounds, experimental and literature values are in agreement within ±50 RI units.

The targeted analysis (for compounds previously found to be associated with malaria-infection) shows no overall significant differences, in terms of quantified masses, between the malaria-positive and -negative cases. However, octanal, which was detected at a comparably higher mass for one positive case, may be a potential indicator of infection-status.

References

[1] Snow, R.W. 2015. *Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa*. BMC Med, 13:23. https://doi.org/10.1186%2Fs12916-014-0254-7.

[2] Pulido, H. Stanczyk, N.M., De Moraes, C.M., Mescher, M. 2021. *A unique volatile signature distinguishes malaria infection from other conditions that cause similar symptoms*. Sci. Rep., 11: 139928. https://doi.org/10.1038/s41598-021-92962-x.

[3] Kuhn, M. 2008. *Building predictive models in R using the caret package*. J. Stat. Softw., 28(5): 1-26. https://doi.org/10.18637/jss.v028.i05.

[4] Kuhn, M. 2019. *The caret Package*. Available from: http://topepo.github.io/caret/index.html [Accessed: 23/11/2021].

[5] Robinson, A. Busula, A.O., Voets, M.A., Beshir, K.B., Caulfield, J.C., Powers, S.J., Verhulst, N.O., Winskill, P., Muwanguzi, J., Birkett, M.A., Smallegange, R.C., Masiga, D.K., Mukabana, W.R., Sauerwein, R.W., Sutherland, C.J., Bousema, T., Pickett, J.A., Takken, W., Logan, J.G., de Boer, J.G. 2017. *Plasmodium-associated changes in human odor attract mosquitoes*. PNAS, 115(18): E4209-E4218. https://doi.org/10.1073/pnas.1721610115.

[6] Wooding, M., Dodgen, T., Rohwer, E.R., Naudé, Y. 2020. *Mass spectral studies on the human skin surface for mosquito vector control applications*. J. Mass Spectrom., 56(2): e4686. https://doi.org/10.1002/jms.4686.

[7] Knols B.G.J., van Loon, J.J.A., Cork, A., Robinson, R.D., Adam, W., Meijerink, de Jong, R., Takken, W. 1997. *Behavioural and electrophysiological responses of the female malaria mosquito Anopheles gambiae (Diptera: Culicidae) to Limburger cheese volatiles*. B. Entomol. Res., 87(2): 151-159. https://doi.org/10.1017/S0007485300027292.

[8] Qiu, Y.T., R.C., Smallegange, R.C., Hoppe, S., Van Loon, J.J.A., Bakker, E.-J., Takken, W. 2004. *Behavioural and electrophysiological responses of the malaria mosquito Anopheles gambiae Giles sensu stricto (Diptera: Culicidae) to human skin emanations*. Med. Vet. Entomol., 18(4): 429-438. https://doi.org/10.1111/j.0269-283x.2004.00534.x.

[9] Kanda, F., Yagi, E., Fukuda, M., Nakajima, K., Ohta, T. Nakata, O. 1990. *Elucidation of chemical compounds responsible for foot malodour*. Brit J. Dermatol., 122(6): 77-776. https://doi.org/10.1111/j.1365-2133.1990.tb06265.x.

[10] Schneider, P., Schoone, G., Schallig, H., Verhage, D., Telgt, D., Eling, W., Sauerwein, R. 2004. *Quantification of Plasmodium falciparum gametocytes in differential stages of development by quantitative nucleic acid sequence-based amplification*. Mol Biochem Parasitol., 137(1): 35-41. https://doi.org/10.1016/j.molbiopara.2004.03.018.

Chapter 9: Conclusion - the applicability of using GC×GC-TOFMS and machine learning for identifying predictive markers in complex samples of biogenic VOCs

The findings of this dual study illustrate the applicability of comprehensive GC×GC-TOFMS, in combination with machine learning, in identifying putative chemical markers — in the form of biogenic volatile organic compounds (VOCs) — as a tool of classification and prediction of discrete biological states.

In the case of the first study (*Identifying predictive volatile markers of genus for southern African Plectranthus and Coleus using GC×GC-TOFMS and machine learning* [Chapters 3-5]), a suite of sesquiterpenes, as well as other C_6-C_8 compounds, were tentatively identified as candidate markers of genus for a group of southern African *Plectranthus* and *Coleus* species.

In the case of the second study (*Identifying predictive volatile markers of malaria infection from human skin using GC×GC-TOFMS and machine learning* [Chapters 6-8]), VOCs originating from the human epidermis were used in an attempt to predict the malaria-infection status of participants.

Comprehensive GC×GC-TOFMS is well suited to the exploratory analysis of complex samples due its two-dimensional chromatographic separation capability, its increased peak capacity (compared to single column GC-MS) and its high resolution. Using comprehensive GC×GC-TOFMS, complex analyte profiles, consisting of over a thousand compounds, were constructed from the foliar and epidermal samples, of the two respective studies, which could be subsequently analysed for potential compounds of interest.

The complex chemical profiles garnered by comprehensive GC×GC-TOFMS represent substantially high-dimensional datasets which require multivariate statistics and machine learning for analysis. Such large datasets can be made more amenable to statistical analysis by processing, including the removal of impertinent compounds and predictors of near-zero variance. Unsupervised statistical methods (PCA and LDA) were employed as preliminary tools of investigation in order to assess the extent of variation and clustering of discrete groups. Three machine learning algorithms (an elastic-net regression, a random forest and a support-vector machine) were used to model training data (using cross-validation), and optimised models were selected to predict the category of samples (genus, in the case of the first study; and malaria-status, in the case of the second). The ranking of top variables by each model was used to select putative markers, and their calculated retention indices were compared with literature values.

The findings of these two preliminary investigations show promise as methods of biomarker discovery, particularly in the case of the *Plectranthus/Coleus* study where an accuracy of up to 90% was obtained (with a sensitivity of up to 100%) and a suite of sesquiterpenes (including α- and β-cubebene, β-ylangene, β-copaene, γ-cadinene and isogermacrene D) was identified as being potentially characteristic of genus *Coleus*. Though predictive models were not obtained in the second study, the list of top variables nevertheless suggested certain compounds, including alcohols (such as (E)-2-octen-1-ol), two nitrogen species (N,N-diethyl-1,2-ethanediamine and isoamyl cyanide), a sulphur species (isothiazole), two short-to long-chain carboxylic acids (*n*-decanoic acid and 9-hexadecenoic acid), as being potentially characteristic of a malaria-positive state. The failure to obtain predictive models from the cutaneous VOC dataset highlights the importance of obtaining a sufficiently large and representative sample population, characterised by good quality data, in order to obtain meaningful statistical results.

Regardless of the particular content, data and findings of the two studies, the common method of combining a sensitive analytical technique (GC×GC-TOFMS) with advanced methods of multivariate statistics (machine learning) demonstrate a promising means of complex sample analysis for the discrimination of discrete biological states.

For corroboration of the results garnered from such non-targeted approaches, targeted and quantitative analyses using certified reference standards are required. Future investigations would also be well served by larger sample populations, and where appropriate and practicable, independently-sampled populations for the training and testing sets. In the future, such a combination of approaches could lead to the identification of biogenic markers for which more refined and targeted techniques could be developed, such as on-site diagnostic tests for particular diseases, or volatilome-based recognition tools of genera or species for ecological applications.

Appendix A.1: Eigenvalues of the principal components for the full foliar VOC dataset

Eigenvalues

Number	Eigenvalue	Percentage	Cumulative percentage	20 40 60 80	Singular Value
1	150,24	8,3979	8,3979		83,132
2	137,68	7,6960	16,094		79,582
3	88,24	4,9324	21,026		63,711
4	75,46	4,2180	25,244		58,916
5	74,42	4,1599	29,404		58,509
6	69,85	3,9044	33,309		56,684
7	64,71	3,6171	36,926		54,559
8	60,71	3,3936	40,319		52,847
9	57,59	3,2194	43,539		51,472
10	56,24	3,1436	46,682		50,862
11	53,38	2,9838	49,666		49,553
12	50,78	2,8384	52,504		48,331
13	50,32	2,8129	55,317		48,113
14	48,45	2,7081	58,026		47,208
15	43,58	2,4359	60,461		44,773
16	42,79	2,3921	62,854		44,368
17	40,23	2,2489	65,102		43,020
18	39,98	2,2345	67,337		42,882
19	38,91	2,1749	69,512		42,307
20	38,13	2,1314	71,643		41,881
21	37,37	2,0889	73,732		41,461
22	35,84	2,0031	75,735		40,601
23	32,99	1,8441	77,579		38,956
24	32,86	1,8369	79,416		38,881
25	29,71	1,6607	81,077		36,968
26	29,01	1,6215	82,698		36,529
27	28,75	1,6071	84,306		36,367
28	28,54	1,5954	85,901		36,234
29	27,20	1,5206	87,422		35,375
30	24,35	1,3609	88,783		33,466
31	22,45	1,2549	90,037		32,136
32	20,34	1,1367	91,174		30,584
33	18,20	1,0172	92,191		28,932
34	16,48	0,9210	93,112		27,531
35	15,61	0,8725	93,985		26,795
36	15,19	0,8491	94,834		26,435
37	14,15	0,7912	95,625		25,516
38	13,61	0,7608	96,386		25,022
39	12,95	0,7238	97,110		24,406
40	11,38	0,6361	97,746		22,879
41	10,84	0,6059	98,352		22,330
42	9,83	0,5493	98,901		21,261
43	8,40	0,4694	99,370		19,654
44	5,51	0,3079	99,678		15,919
45	4,86	0,2717	99,950		14,954

Appendix A.2: AUC/ROC (by cross-validation) and related statistics for model tuning parameters, for the elastic-net, random forest and support-vector machine

Elastic-net (glmnet)

α	λ	AUC/ROC	Sensitivity	Specificity	AUC/ROC SD	Sens SD	Spec SD
0.00	0.0001	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.1112	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.2223	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.3334	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.4445	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.5556	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.6667	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.7778	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	0.8889	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.00	1.0000	0.6627778	0.4600000	0.9400000	0.2849626	0.3544688	0.1433721
0.05	0.0001	0.7466667	0.4866667	0.9000000	0.2478781	0.3433225	0.1863390
0.05	0.1112	0.7366667	0.4866667	0.8800000	0.2416121	0.3433225	0.2013841
0.05	0.2223	0.7500000	0.5000000	0.8800000	0.2555644	0.3298428	0.2013841
0.05	0.3334	0.7450000	0.5000000	0.8400000	0.2739575	0.3600411	0.2229848
0.05	0.4445	0.7183333	0.4666667	0.8000000	0.2763133	0.3435921	0.2357023
0.05	0.5556	0.7133333	0.5266667	0.7600000	0.2615294	0.3287180	0.2409472
0.05	0.6667	0.7111111	0.5600000	0.7400000	0.2637670	0.2921821	0.2409472
0.05	0.7778	0.6988889	0.6000000	0.7400000	0.2817631	0.3298428	0.2409472
0.05	0.8889	0.6988889	0.6000000	0.7200000	0.2626286	0.3298428	0.2392117
0.05	1.0000	0.6722222	0.6400000	0.6866667	0.2936441	0.3143188	0.2859358
0.25	0.0001	0.7000000	0.5933333	0.6266667	0.2956083	0.3472111	0.3343041
0.25	0.1112	0.7400000	0.6933333	0.6666667	0.3026886	0.3143188	0.3154949
0.25	0.2223	0.7522222	0.6733333	0.6866667	0.2905534	0.3418360	0.3203008
0.25	0.3334	0.7722222	0.6733333	0.6866667	0.2896757	0.3418360	0.3203008
0.25	0.4445	0.7788889	0.6133333	0.7066667	0.2903940	0.3810317	0.3237512
0.25	0.5556	0.7855556	0.6333333	0.6733333	0.2404985	0.3600411	0.3135815
0.25	0.6667	0.7455556	0.5866667	0.6400000	0.2536615	0.3473444	0.2621139
0.25	0.7778	0.7155556	0.5533333	0.6400000	0.2460177	0.3392530	0.2621139
0.25	0.8889	0.7222222	0.4866667	0.6400000	0.2357023	0.3295620	0.2792185
0.25	1.0000	0.7777778	0.4400000	0.6600000	0.2213525	0.3691833	0.3172509
0.50	0.0001	0.7205556	0.7133333	0.6466667	0.2768596	0.2826790	0.3092703
0.50	0.1112	0.7855556	0.6200000	0.7133333	0.2416986	0.3519785	0.2867442
0.50	0.2223	0.7855556	0.5866667	0.7066667	0.2404985	0.3473444	0.2733537
0.50	0.3334	0.7466667	0.5333333	0.6400000	0.2184754	0.3263150	0.2621139
0.50	0.4445	0.7733333	0.4466667	0.7133333	0.2177915	0.3559026	0.2657972
0.50	0.5556	0.6566667	0.4466667	0.6933333	0.1981103	0.3810317	0.3143188
0.50	0.6667	0.5150000	0.8000000	0.1800000	0.1040833	0.4082483	0.3785939
0.50	0.7778	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.50	0.8889	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.50	1.0000	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.75	0.0001	0.7605556	0.6000000	0.6733333	0.2346524	0.3118048	0.2783882
0.75	0.1112	0.7711111	0.5533333	0.7066667	0.2168447	0.3392530	0.2512930
0.75	0.2223	0.7522222	0.4800000	0.7533333	0.2081666	0.3446684	0.2458922
0.75	0.3334	0.7472222	0.4600000	0.7333333	0.2304258	0.3704352	0.2846375
0.75	0.4445	0.5083333	0.8000000	0.1800000	0.1102396	0.4082483	0.3785939
0.75	0.5556	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.75	0.6667	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.75	0.7778	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.75	0.8889	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.75	1.0000	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	0.0001	0.7944444	0.6000000	0.7733333	0.2243819	0.3298428	0.2447599
0.95	0.1112	0.8233333	0.4800000	0.7933333	0.2169336	0.3736705	0.2419060
0.95	0.2223	0.7833333	0.4333333	0.8000000	0.2381448	0.3535534	0.2357023
0.95	0.3334	0.5166667	0.6733333	0.2800000	0.1284253	0.4342938	0.3840573
0.95	0.4445	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	0.5556	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	0.6667	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	0.7778	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	0.8889	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
0.95	1.0000	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	0.0001	0.7733333	0.5066667	0.8000000	0.2338783	0.3452053	0.2545875
1.00	0.1112	0.8200000	0.4266667	0.7933333	0.2120229	0.3666667	0.2419060
1.00	0.2223	0.7700000	0.4066667	0.7800000	0.2298349	0.3157883	0.2392117
1.00	0.3334	0.5083333	0.8000000	0.1800000	0.1102396	0.4082483	0.3785939
1.00	0.4445	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	0.5556	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	0.6667	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	0.7778	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	0.8889	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625
1.00	1.0000	0.5000000	0.8800000	0.1200000	0.0000000	0.3316625	0.3316625

Random forest (ranger)

mtry	minimal node size	splitting rule	AUC/ROC	Sensitivity	Specificity	AUC/ROC SD	Sens SD	Spec SD
2	1	gini	0.9166667	0.6466667	0.8200000	0.1463285	0.3736705	0.2542236
2	3	gini	0.9133333	0.6933333	0.8266667	0.1511530	0.3654019	0.2229848
2	5	gini	0.9200000	0.7066667	0.8066667	0.1586255	0.3704352	0.2531286
2	1	extratrees	0.9200000	0.7066667	0.8400000	0.1432913	0.3704352	0.2281163
2	3	extratrees	0.9066667	0.5933333	0.7933333	0.1654819	0.3911048	0.2690862
2	5	extratrees	0.8733333	0.6733333	0.7733333	0.1926304	0.3741657	0.2716547
3	1	gini	0.9266667	0.6066667	0.8666667	0.1508464	0.3875373	0.2256677
3	3	gini	0.9133333	0.6866667	0.8400000	0.1586255	0.3828499	0.2281163
3	5	gini	0.9233333	0.6533333	0.8666667	0.1538849	0.3846595	0.2041241
3	1	extratrees	0.9233333	0.6866667	0.8200000	0.1538849	0.3674235	0.2542236
3	3	extratrees	0.9133333	0.6533333	0.8200000	0.1567907	0.3565524	0.2542236
3	5	extratrees	0.9000000	0.6800000	0.8400000	0.1784060	0.3994209	0.2281163
5	1	gini	0.9133333	0.6733333	0.8400000	0.1657614	0.3615809	0.2229848
5	3	gini	0.9233333	0.7000000	0.8533333	0.1699673	0.3600411	0.2273030
5	5	gini	0.9200000	0.6733333	0.8333333	0.1586255	0.3893252	0.2357023
5	1	extratrees	0.8833333	0.6000000	0.8133333	0.1717961	0.4025382	0.2221528
5	3	extratrees	0.9000000	0.6933333	0.8200000	0.1631575	0.3356861	0.2353091
5	5	extratrees	0.9000000	0.7066667	0.7800000	0.1559024	0.3704352	0.2440021

Support-vector machine (svmPoly)

C	Degree	Scale	AUC/ROC	Sensitivity	Specificity	AUC/ROC SD	Sens SD	Spec SD
0.01	1	1	0.6094444	0.4666667	0.8600000	0.2961331	0.3263150	0.2134375
0.10	1	1	0.6016667	0.4666667	0.8600000	0.2948384	0.3263150	0.2134375
1.00	1	1	0.6094444	0.4666667	0.8800000	0.2983235	0.3263150	0.2013841
10.00	1	1	0.6094444	0.4666667	0.8600000	0.2961331	0.3263150	0.2134375
0.01	1	2	0.6116667	0.4666667	0.8600000	0.2954854	0.3263150	0.2134375
0.10	1	2	0.6094444	0.4666667	0.8600000	0.2961331	0.3263150	0.2134375
1.00	1	2	0.6144444	0.4666667	0.8600000	0.2952622	0.3263150	0.2134375
10.00	1	2	0.6166667	0.4666667	0.8600000	0.2945733	0.3263150	0.2134375
0.01	1	3	0.6188889	0.4800000	0.8466667	0.2956377	0.3274480	0.2353091
0.10	1	3	0.6216667	0.4666667	0.8600000	0.2979720	0.3263150	0.2134375
1.00	1	3	0.6044444	0.4666667	0.8600000	0.2969138	0.3263150	0.2134375
10.00	1	3	0.6166667	0.4666667	0.8600000	0.2945733	0.3263150	0.2134375
0.01	2	1	0.4911111	0.4600000	0.6933333	0.2893115	0.3128424	0.3622205
0.10	2	1	0.4738889	0.4000000	0.7333333	0.2937897	0.3154949	0.3367877
1.00	2	1	0.4594444	0.4800000	0.6400000	0.2940795	0.2978317	0.3869396
10.00	2	1	0.4327778	0.4600000	0.6933333	0.2866825	0.3128424	0.3622205
0.01	2	2	0.5650000	0.5533333	0.6533333	0.3026673	0.3107637	0.3816388
0.10	2	2	0.4633333	0.4066667	0.7466667	0.2959311	0.3007706	0.3267404
1.00	2	2	0.4855556	0.4733333	0.6800000	0.3103444	0.2792185	0.3565524
10.00	2	2	0.5627778	0.5066667	0.6400000	0.2922206	0.3315229	0.3869396
0.01	2	3	0.4638889	0.4333333	0.6933333	0.2926845	0.3080705	0.3622205
0.10	2	3	0.4650000	0.4066667	0.7733333	0.2926054	0.3007706	0.3038335
1.00	2	3	0.5200000	0.4933333	0.6800000	0.2954201	0.2984528	0.3565524
10.00	2	3	0.4094444	0.4200000	0.5333333	0.2801439	0.3047464	0.4110736
0.01	3	1	0.4555556	0.3933333	0.6733333	0.2881456	0.3294215	0.3583656
0.10	3	1	0.3733333	0.3933333	0.5600000	0.2624633	0.3294215	0.3815174
1.00	3	1	0.4505556	0.4266667	0.7133333	0.2838763	0.3230354	0.3351340
10.00	3	1	0.4272222	0.3600000	0.6333333	0.2900440	0.2992274	0.3757708
0.01	3	2	0.4888889	0.5066667	0.5600000	0.2908111	0.3315229	0.4079079
0.10	3	2	0.4400000	0.3933333	0.6733333	0.2890569	0.3294215	0.3280131
1.00	3	2	0.4700000	0.4400000	0.6400000	0.2907470	0.3258891	0.3747839
10.00	3	2	0.5044444	0.4266667	0.6666667	0.2969571	0.3230354	0.3632416
0.01	3	3	0.4972222	0.4800000	0.6200000	0.2943549	0.3309638	0.3680529
0.10	3	3	0.4588889	0.4066667	0.6666667	0.2924779	0.3007706	0.3333333
1.00	3	3	0.4750000	0.4533333	0.6133333	0.2936715	0.2569407	0.3686813
10.00	3	3	0.4444444	0.3800000	0.6933333	0.2849762	0.2907525	0.3179797

Appendix B.1: Official permissions for conducting research in the government clinics of Masisi and Madimbo, Limpopo province

Permission from the Limpopo Provincial Government DOH

LIMPOPO
PROVINCIAL GOVERNMENT
REPUBLIC OF SOUTH AFRICA

Department of Health

Ref : LP- 202002 - 014
Enquires : Ms PF Mahlokwane
Tel : 015-293 6028
Email : Kurhula.Hlomane@dhsd.limpopo.gov.za

Elsabe de Kock

PERMISSION TO CONDUCT RESEARCH IN DEPARTMENTAL FACILITIES

Your Study Topic as indicated below;

The detection of skin volatile organic compounds associated with malaria infection: an analysis using novel non-invasive sportive sampling with comprehensive GCxGC-ToFMS.

1. Permission to conduct research study as per your research proposal is hereby Granted.
2. Kindly note the following:
   a. Present this letter of permission to the institution supervisor/s a week before the study is conducted.
   b. In the course of your study, there should be no action that disrupts the routine services, or incur any cost on the Department.
   c. After completion of study, it is mandatory that the findings should be submitted to the Department to serve as a resource.
   d. The researcher should be prepared to assist in the interpretation and implementation of the study recommendation where possible.
   e. The approval is only valid for a 1-year period.
   f. If the proposal has been amended, a new approval should be sought from the Department of Health
   g. Kindly note that, the Department can withdraw the approval at any time.

Your cooperation will be highly appreciated

Head of Department **Date** 19/03/2020

Private Bag X9302 Polokwane
Fidel Castro Ruz House, 18 College Street. Polokwane 0700. Tel: 015 293 6000/12. Fax: 015 293 6211.
Website: http/www.limpopo.gov.za

The heartland of Southern Africa – Development is about people!

Permission from the Masisi clinic

I, Matshate FS, Matron / Head Sister of the Masisi clinic, hereby grant approval for the researchers of the University of Pretoria Institute for Sustainable Malaria Control (UP ISMC), Mr Daniel T. Pretorius and Professor Anton Stoltz, to conduct malaria research at my clinic. I understand that participation is voluntary and that participation may be stopped at any time.

Signature	Name & Surname	Date
	Matshate FS	29/01/2019.

Permission from the Madimbo clinic

I, MAFUMWA FORTUNATE, Matron / (Head Sister) of the MADIMBO clinic, hereby grant approval for the researchers of the University of Pretoria Institute for Sustainable Malaria Control (UP ISMC), Mr Daniel T. Pretorius and Professor Anton Stoltz, to conduct malaria research at my clinic. I understand that participation is voluntary and that participation may be stopped at any time.

Signature	Name & Surname	Date
CHP	Fortunate Mafumwa	29.01.2020

Appendix B.2: Participant responses to the general health and lifestyle questionnaire

Question	Participant number (#1-5)				
	#1	#2	#3	#4	#5
Gender	Female	Female	Female	Female	Female
Age	38	23	29	35	31
Race	African	African	African	African	African
Do you smoke or use any kind of tobacco product?	No	No	No	No	No
How regularly do you drink alcohol?	Never	Never	Never	Rarely	Occasionally
Do you use any recreational drugs?	No	No	No	No	No
Do you smoke a vaporiser?	No	No	No	No	No
Have you showered or bathed in the past 24 hours?	Yes	Yes	Yes	Yes	Yes
Do you use body lotions or creams?	Yes	Yes	Yes	Yes	Yes
Do you use deodorant, perfume, cologne or any other fragrances?	Yes	Yes	Yes	Yes	Yes
Are you vegan?	Yes	Yes	No	No	No
Are you vegetarian?	Yes	Yes	No	No	No
Do you eat garlic regularly, or have you eaten any garlic in the past 24 hours?	No	No	No	No	No
Do you eat exotic or spicy foods regularly, or	No	No	No	No	Yes

have you eaten exotic or spicy food in the past 24 hours?					
Do you take vitamin supplements?	No	No	Yes	No	No
Are you currently on any medication? If you ticked yes, please specify the medication in the space provided.	Yes; Glycomin	No	No	No	No
Do you have any of the following diseases (HIV/TB/Cholera)? If you choose “Yes”, Please mark which one/s you have. If you have any disease, illness or condition that is not listed above, please write it down in the space below	Diabetes	N/A	N/A	N/A	N/A
Do you have any form of cancer? If you choose “Yes”, please write down what type of cancer you have.	No	No	No	No	No
How many times have you had malaria in the past?	1 time	1 time	0 times	0 times	1 time
Are you currently on any anti-malarial treatment?	No	No	No	No	No

If you chose “Yes”, for how long have you been on the treatment?	No	No	No	No	No

Question	**Participant number (#6-10)**				
	#6	**#7**	**#8**	**#9**	**#10**
Gender	Female	Female	Female	Female	Female
Age	18	38	36	26	47
Race	African	African	African	African	African
Do you smoke or use any kind of tobacco product?	No	No	No	No	No
How regularly do you drink alcohol?	Never	Never	Never	Never	Never
Do you use any recreational drugs?	No	No	No	No	No
Do you smoke a vaporiser?	No	No	No	No	No
Have you showered or bathed in the past 24 hours?	Yes	Yes	Yes	Yes	Yes
Do you use body lotions or creams?	Yes	Yes	Yes	Yes	Yes
Do you use deodorant, perfume, cologne or any other fragrances?	Yes	Yes	Yes	Yes	Yes
Are you vegan?	Yes	No	No	No	No
Are you vegetarian?	Yes	No	No	No	N/A
Do you eat garlic regularly, or have you	No	Yes	No	No	N/A

eaten any garlic in the past 24 hours?					
Do you eat exotic or spicy foods regularly, or have you eaten exotic or spicy food in the past 24 hours?	Yes	No	No	Yes	N/A
Do you take vitamin supplements?	Yes	Yes	No	Yes	N/A
Are you currently on any medication? If you ticked yes, please specify the medication in the space provided.	No	Yes; Antiretroviral	No	Yes; Antiretroviral	No
Do you have any of the following diseases (HIV/TB/Cholera)? If you choose “Yes”, Please mark which one/s you have. If you have any disease, illness or condition that is not listed above, please write it down in the space below.	N/A	Yes; HIV	N/A	Yes; HIV; depression	No
Do you have any form of cancer? If you choose “Yes”, please write down what type of cancer you have.	No	No	No	No	No
How many times have you had malaria in the past?	1 time	1 time	0 times	0 times	3 times

Are you currently on any anti-malarial treatment?	No	No	No	No	No
If you chose "Yes", for how long have you been on the treatment?	N/A	N/A	N/A	N/A	N/A

Question	**Participant number (#11-15)**				
	#11	**#12**	**#13**	**#14**	**#15**
Gender	Female	Female	Female	Female	Female
Age	24	31	27	80	46
Race	African	African	African	African	African
Do you smoke or use any kind of tobacco product?	No	No	No	No	No
How regularly do you drink alcohol?	Never	Never	Never	Never	Never
Do you use any recreational drugs?	No	No	No	No	No
Do you smoke a vaporiser?	No	No	No	No	No
Have you showered or bathed in the past 24 hours?	Yes	Yes	Yes	Yes	Yes
Do you use body lotions or creams?	Yes	Yes	Yes	Yes	Yes
Do you use deodorant, perfume, cologne or any other fragrances?	Yes	Yes	Yes	No	Yes
Are you vegan?	No	No	No	No	No
Are you vegetarian?	No	No	No	No	No
Do you eat garlic regularly, or have you	No	No	No	No	Yes

eaten any garlic in the past 24 hours?					
Do you eat exotic or spicy foods regularly, or have you eaten exotic or spicy food in the past 24 hours?	No	Yes	Yes	No	Yes
Do you take vitamin supplements?	No	No	Yes	N/A	No
Are you currently on any medication? If you ticked yes, please specify the medication in the space provided.	No	No	Yes; paracetamol; Allergex	No	Yes; high blood-pressure medication
Do you have any of the following diseases (HIV/TB/Cholera)? If you choose “Yes”, Please mark which one/s you have. If you have any disease, illness or condition that is not listed above, please write it down in the space below	No	No	No	No	Yes; HIV
Do you have any form of cancer? If you choose “Yes”, please write down what type of cancer you have.	No	No	No	No	No
How many times have you had malaria in the past?	0 times	0 times	0 times	2 times	1 time

Are you currently on any anti-malarial treatment?	No	No	No	No	No
If you chose "Yes", for how long have you been on the treatment?	N/A	N/A	N/A	N/A	N/A

Question	**Participant number (#16-20)**				
	#16	**#17**	**#18**	**#19**	**#20**
Gender	Male	Male	Female	Female	Female
Age	18	32	25	69	47
Race	African	African	African	African	African
Do you smoke or use any kind of tobacco product?	No	Yes	No	No	No
How regularly do you drink alcohol?	Never	Occasionally	Never	Never	Never
Do you use any recreational drugs?	No	No	No	No	No
Do you smoke a vaporiser?	No	Yes	No	No	No
Have you showered or bathed in the past 24 hours?	Yes	Yes	N/A	Yes	Yes
Do you use body lotions or creams?	Yes	Yes	Yes	Yes	Yes
Do you use deodorant, perfume, cologne or any other fragrances?	Yes	Yes	Yes	Yes	Yes
Are you vegan?	No	No	No	No	No
Are you vegetarian?	No	No	No	No	No
Do you eat garlic regularly, or have you	Yes	No	No	No	No

eaten any garlic in the past 24 hours?					
Do you eat exotic or spicy foods regularly, or have you eaten exotic or spicy food in the past 24 hours?	No	No	Yes	Yes	Yes
Do you take vitamin supplements?	Yes	N/A	No	N/A	No
Are you currently on any medication? If you ticked yes, please specify the medication in the space provided.	No	No	No	No	No
Do you have any of the following diseases(HIV/TB/Cholera)? If you choose “Yes”, Please mark which one/s you have. If you have any disease, illness or condition that is not listed above, please write it down in the space below	No	No	No	No; high blood-pressure	No; patient reported having contracted cholera 9 years ago
Do you have any form of cancer? If you choose “Yes”, please write down what type of cancer you have.	No	No	No	No	No
How many times have you had malaria in the past?	1 time	3 times	3 times	More than 3 times	0 times
Are you currently on any anti-malarial treatment?	N/A	No	Yes†	No	No

If you chose "Yes", for how long have you been on the treatment?	N/A	N/A	4-7 days†	N/A	N/A

† The researcher is of the opinion that the participant may have misunderstood the question, or answered it incorrectly, since the nurse at the clinic provided the patient with medication on the day that the RDT was administered, and the participant presumably visited the clinic on the day for the purpose of receiving diagnosis and treatment.

Question	**Participant number (#21-25)**				
	#21	**#22**	**#23**	**#24**	**#25**
Gender	Female	Female	Male	Female	Male
Age	21	27	49	21	25
Race	African	African	African	African	African
Do you smoke or use any kind of tobacco product?	No	No	No	No	No
How regularly do you drink alcohol?	Never	Never	Never	Never	Never
Do you use any recreational drugs?	No	No	No	No	No
Do you smoke a vaporiser?	No	No	No	No	No
Have you showered or bathed in the past 24 hours?	Yes	Yes	Yes	Yes	N/A
Do you use body lotions or creams?	Yes	Yes	Yes	Yes	Yes
Do you use deodorant, perfume, cologne or any other fragrances?	Yes	Yes	No	Yes	Yes
Are you vegan?	No	No	No	No	No
Are you vegetarian?	No	No	No	No	No

Do you eat garlic regularly, or have you eaten any garlic in the past 24 hours?	Yes	No	No	No	No
Do you eat exotic or spicy foods regularly, or have you eaten exotic or spicy food in the past 24 hours?	Yes	Yes	No	Yes	No
Do you take vitamin supplements?	Yes	No	No	Yes	Yes
Are you currently on any medication? If you ticked yes, please specify the medication in the space provided.	No	Yes; ulcer medication	Yes; antiretroviral	Yes; pregnancy medication	No
Do you have any of the following diseases (HIV/TB/Cholera)? If you choose “Yes”, Please mark which one/s you have. If you have any disease, illness or condition that is not listed above, please write it down in the space below	No	No	Yes; HIV	No	No
Do you have any form of cancer? If you choose “Yes”, please write down what type of cancer you have.	No	No	No	No	No

How many times have you had malaria in the past?	0 times	0 times	Never	0 times	0 times
Are you currently on any anti-malarial treatment?	No	No	No	No	No
If you chose “Yes”, for how long have you been on the treatment?	N/A	N/A	N/A	N/A	N/A

Appendix B.3: Eigenvalues of the principal components for the full cutaneous VOC dataset

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent	20 40 60 80	Singular Value
1	148,96	4,7363	4,7363		88,010
2	111,99	3,5607	8,2970		76,310
3	104,77	3,3315	11,628		73,812
4	101,21	3,2181	14,847		72,545
5	97,96	3,1148	17,961		71,372
6	94,90	3,0175	20,979		70,248
7	92,98	2,9564	23,935		69,533
8	84,73	2,6941	26,629		66,377
9	79,88	2,5399	29,169		64,449
10	76,93	2,4460	31,615		63,248
11	76,56	2,4344	34,050		63,097
12	76,18	2,4223	36,472		62,940
13	71,07	2,2599	38,732		60,794
14	69,36	2,2056	40,937		60,058
15	67,76	2,1544	43,092		59,357
16	66,76	2,1226	45,214		58,918
17	65,16	2,0719	47,286		58,209
18	64,84	2,0617	49,348		58,067
19	64,47	2,0500	51,398		57,901
20	63,54	2,0204	53,418		57,482
21	61,87	1,9671	55,385		56,719
22	61,23	1,9470	57,332		56,428
23	60,77	1,9323	59,265		56,215
24	59,45	1,8903	61,155		55,600
25	58,18	1,8498	63,005		55,002
26	57,12	1,8163	64,821		54,501
27	56,55	1,7982	66,619		54,228
28	55,54	1,7660	68,385		53,741
29	54,97	1,7479	70,133		53,465
30	54,47	1,7319	71,865		53,220
31	53,60	1,7043	73,569		52,793
32	51,79	1,6467	75,216		51,894
33	50,77	1,6143	76,830		51,381
34	49,41	1,5711	78,401		50,689
35	48,54	1,5434	79,945		50,240
36	47,16	1,4995	81,444		49,521
37	46,61	1,4822	82,926		49,233
38	46,16	1,4676	84,394		48,991
39	45,51	1,4471	85,841		48,647
40	45,35	1,4419	87,283		48,560
41	44,67	1,4204	88,703		48,196
42	44,12	1,4027	90,106		47,896
43	42,79	1,3606	91,467		47,172
44	41,84	1,3302	92,797		46,642
45	41,58	1,3221	94,119		46,499
46	35,98	1,1442	95,263		43,257
47	35,46	1,1274	96,391		42,939
48	33,89	1,0775	97,468		41,978
49	29,52	0,9386	98,407		39,179
50	25,09	0,7976	99,204		36,117
51	22,62	0,7192	99,924		34,297

Appendix B.4: AUC/ROC (by cross-validation) and related statistics for model tuning parameters, for the elastic-net, random forest and support-vector machine

Elastic-net (glmnet)

α	λ	AUC/ROC	Sensitivity	Specificity	AUC/ROC SD	Sens SD	Spec SD
0.00	0.0001	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.1112	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.2223	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.3334	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.4445	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.5556	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.6667	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.7778	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	0.8889	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.00	1.0000	0.57000	0.886	0.25	0.38023538	0.18457158	0.4442617
0.05	0.0001	0.58500	0.662	0.30	0.35359060	0.27167689	0.4701623
0.05	0.1112	0.58500	0.680	0.30	0.35359060	0.27424138	0.4701623
0.05	0.2223	0.58500	0.688	0.30	0.35359060	0.26231343	0.4701623
0.05	0.3334	0.62250	0.688	0.30	0.30628289	0.26231343	0.4701623
0.05	0.4445	0.62250	0.688	0.30	0.30628289	0.26231343	0.4701623
0.05	0.5556	0.64750	0.696	0.30	0.29445534	0.26256745	0.4701623
0.05	0.6667	0.64750	0.696	0.30	0.28721668	0.26256745	0.4701623
0.05	0.7778	0.62500	0.704	0.30	0.29132185	0.24954959	0.4701623
0.05	0.8889	0.62500	0.704	0.30	0.29132185	0.24954959	0.4701623
0.05	1.0000	0.66250	0.704	0.30	0.27714475	0.24954959	0.4701623
0.25	0.0001	0.59750	0.672	0.35	0.31054917	0.27953831	0.4893605
0.25	0.1112	0.57000	0.706	0.20	0.32134585	0.25303162	0.4103913
0.25	0.2223	0.58250	0.738	0.05	0.31675784	0.25011664	0.2236068
0.25	0.3334	0.57250	0.766	0.05	0.32584909	0.26485845	0.2236068
0.25	0.4445	0.56250	0.820	0.05	0.31492480	0.22407216	0.2236068
0.25	0.5556	0.59500	0.864	0.05	0.35314378	0.20285463	0.2236068
0.25	0.6667	0.57500	0.872	0.05	0.35743973	0.18712741	0.2236068
0.25	0.7778	0.52500	0.906	0.05	0.33658502	0.17872232	0.2236068
0.25	0.8889	0.51875	0.956	0.00	0.33667052	0.10735455	0.0000000
0.25	1.0000	0.47750	0.992	0.00	0.33862143	0.04000000	0.0000000
0.50	0.0001	0.56000	0.706	0.15	0.31313357	0.27207536	0.3663475
0.50	0.1112	0.61500	0.818	0.15	0.32811263	0.21354157	0.3663475
0.50	0.2223	0.57250	0.852	0.05	0.36902254	0.18567445	0.2236068
0.50	0.3334	0.52500	0.880	0.05	0.33658502	0.17911821	0.2236068
0.50	0.4445	0.49125	0.930	0.00	0.33188804	0.14505746	0.0000000
0.50	0.5556	0.43000	1.000	0.00	0.11109503	0.00000000	0.0000000
0.50	0.6667	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.50	0.7778	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.50	0.8889	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.50	1.0000	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	0.0001	0.58750	0.802	0.20	0.36377950	0.23650229	0.4103913
0.75	0.1112	0.55125	0.862	0.15	0.32801988	0.18668155	0.3663475
0.75	0.2223	0.51250	0.906	0.00	0.28487070	0.17872232	0.0000000
0.75	0.3334	0.48000	0.972	0.00	0.25152168	0.07783530	0.0000000
0.75	0.4445	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	0.5556	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	0.6667	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	0.7778	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	0.8889	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.75	1.0000	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	0.0001	0.58125	0.866	0.10	0.29579009	0.19349849	0.3077935
0.95	0.1112	0.50875	0.882	0.10	0.28897220	0.17788105	0.3077935
0.95	0.2223	0.51750	0.948	0.00	0.25727571	0.09517528	0.0000000
0.95	0.3334	0.48750	1.000	0.00	0.03847419	0.00000000	0.0000000
0.95	0.4445	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	0.5556	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	0.6667	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	0.7778	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	0.8889	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
0.95	1.0000	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.0001	0.56375	0.898	0.10	0.29351264	0.14823967	0.3077935
1.00	0.1112	0.54125	0.898	0.10	0.28942718	0.14823967	0.3077935
1.00	0.2223	0.52625	0.972	0.00	0.25974114	0.07783530	0.0000000
1.00	0.3334	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.4445	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.5556	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.6667	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.7778	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	0.8889	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000
1.00	1.0000	0.50000	1.000	0.00	0.00000000	0.00000000	0.0000000

Random forest (ranger)

mtry	minimal node size	splitting rule	AUC/ROC	Sensitivity	Specificty	AUC/ROC SD	Sens SD	Spec SD
2	1	gini	0.7200	1	0	0.4372161	0	0
2	3	gini	0.7400	1	0	0.4175744	0	0
2	5	gini	0.7625	1	0	0.4251548	0	0
2	1	extratrees	0.7400	1	0	0.4405738	0	0
2	3	extratrees	0.7200	1	0	0.4274773	0	0
2	5	extratrees	0.7150	1	0	0.4316736	0	0
3	1	gini	0.7400	1	0	0.4175744	0	0
3	3	gini	0.7500	1	0	0.3973597	0	0
3	5	gini	0.7325	1	0	0.4085517	0	0
3	1	extratrees	0.7150	1	0	0.4316736	0	0
3	3	extratrees	0.7050	1	0	0.4113201	0	0
3	5	extratrees	0.7300	1	0	0.4133942	0	0
5	1	gini	0.7750	1	0	0.3795773	0	0
5	3	gini	0.7825	1	0	0.3584598	0	0
5	5	gini	0.7875	1	0	0.3467424	0	0
5	1	extratrees	0.7600	1	0	0.3796120	0	0
5	3	extratrees	0.8075	1	0	0.3023047	0	0
5	5	extratrees	0.7725	1	0	0.3918495	0	0

Support-vector machine (svmPoly)

C	degree	scale	AUC/ROC	Sensitivity	Specificty	AUC/ROC SD	Sens SD	Spec SD
0.01	1	1	0.5225	0.858	0.30	0.4348245	.20293677	0.4701623
0.10	1	1	0.5025	0.864	0.25	0.4354293	.22616366	0.4442617
1.00	1	1	0.4925	0.840	0.25	0.4353689	.20412415	0.4442617
10.00	1	1	0.5025	0.852	0.25	0.4354293	.20024984	0.4442617
0.01	1	2	0.4475	0.804	0.30	0.4375395	.23625551	0.4701623
0.10	1	2	0.4525	0.862	0.25	0.4327011	.21712132	0.4442617
1.00	1	2	0.4975	0.860	0.35	0.4408350	.20665995	0.4893605
10.00	1	2	0.5125	0.858	0.25	0.4352480	.23124662	0.4442617
0.01	1	3	0.4775	0.882	0.25	0.4348245	.15803481	0.4442617
0.10	1	3	0.5625	0.872	0.30	0.4306895	.20970217	0.4701623
1.00	1	3	0.4475	0.838	0.20	0.4375395	.23991318	0.4103913
10.00	1	3	0.5125	0.866	0.30	0.4352480	.20902552	0.4701623
0.01	2	1	0.7400	0.952	0.25	0.3952348	.13266499	0.4442617
0.10	2	1	0.6650	0.976	0.25	0.4380279	.06633250	0.4442617
1.00	2	1	0.6500	0.960	0.20	0.4394973	.10000000	0.4103913
10.00	2	1	0.6600	0.944	0.25	0.4357691	.14742230	0.4442617
0.01	2	2	0.7350	0.952	0.25	0.4029823	.13266499	0.4442617
0.10	2	2	0.6900	0.936	0.25	0.4228973	.14966630	0.4442617
1.00	2	2	0.7750	0.984	0.25	0.3753945	.05537749	0.4442617
10.00	2	2	0.6600	0.952	0.30	0.4357691	.13266499	0.4701623
0.01	2	3	0.6900	0.936	0.20	0.4228973	.14966630	0.4103913
0.10	2	3	0.7350	0.952	0.30	0.4029823	.13266499	0.4701623
1.00	2	3	0.6900	0.936	0.20	0.4228973	.14966630	0.4103913
10.00	2	3	0.7500	0.976	0.30	0.3886549	.06633250	0.4701623
0.01	3	1	0.6275	0.880	0.35	0.4247213	.15942605	0.4893605
0.10	3	1	0.5275	0.884	0.30	0.4435132	.16312061	0.4701623
1.00	3	1	0.4825	0.906	0.30	0.4398789	.13868429	0.4701623
10.00	3	1	0.6775	0.894	0.40	0.4053832	.15160255	0.5026247
0.01	3	2	0.5825	0.888	0.30	0.4320316	.17576025	0.4701623
0.10	3	2	0.5875	0.910	0.35	0.4352480	.12747549	0.4893605
1.00	3	2	0.5075	0.882	0.25	0.4443431	.14922020	0.4442617
10.00	3	2	0.4975	0.840	0.30	0.4444023	.18085445	0.4701623
0.01	3	3	0.5275	0.860	0.25	0.4435132	.16894279	0.4442617
0.10	3	3	0.5575	0.890	0.25	0.4404767	.17078251	0.4442617
1.00	3	3	0.5075	0.912	0.25	0.4443431	.16026021	0.4442617
10.00	3	3	0.5975	0.886	0.35	0.4330051	.13958271	0.4893605