A machine learning approach using a handheld near-infrared (NIR) device to predict the effect of storage conditions on tomato biomarkers

Journal:	*ACS Food Science & Technology*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Emsley, Natalia; Lancaster Girls' Grammar School, Chemistry Holden, Claire; Lancaster University Faculty of Science and Technology Guo, Yi; Lancaster Girls' Grammar School, Chemistry Bevan, Rhiann; Lancaster Girls' Grammar School, Chemistry Rees, Chris; Lancaster Girls' Grammar School, Chemistry McAinsh, Martin; Lancaster University, Centre for Biophotonics Martin, Francis; Biocel Ltd, Biocel Analytics Morais, Camilo; University of Central Lancashire, School of Pharmacy and Biomedical Sciences

A machine learning approach using a handheld near-infrared (NIR) device to predict the effect of storage conditions on tomato biomarkers

Natalia E M Emsley[1,¶], Claire A. Holden[2,¶], Sarah Guo[1], Rhiann S Bevan[1], Christopher Rees[3], Martin R. McAinsh[2], Francis L Martin[4,*], Camilo L M Morais[5,*]

Sixth Form, Lancaster Girls' Grammar School, Lancaster, Lancashire, UK

[2] *Lancaster Environment Centre, Library Avenue, Lancaster University, Lancaster LA1 4YQ, UK*

[3] *Chemistry Department, Lancaster Girls' Grammar School, Lancaster, Lancashire, UK*

[4] *Biocel Ltd, Hull HU10 7TS, UK*

[5] *School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston, Lancashire, UK*

***Corresponding authors:** Francis L Martin (flm13@biocel.uk); Camilo L M Morais (camilomorais1@gmail.com)

¶, Joint first authors

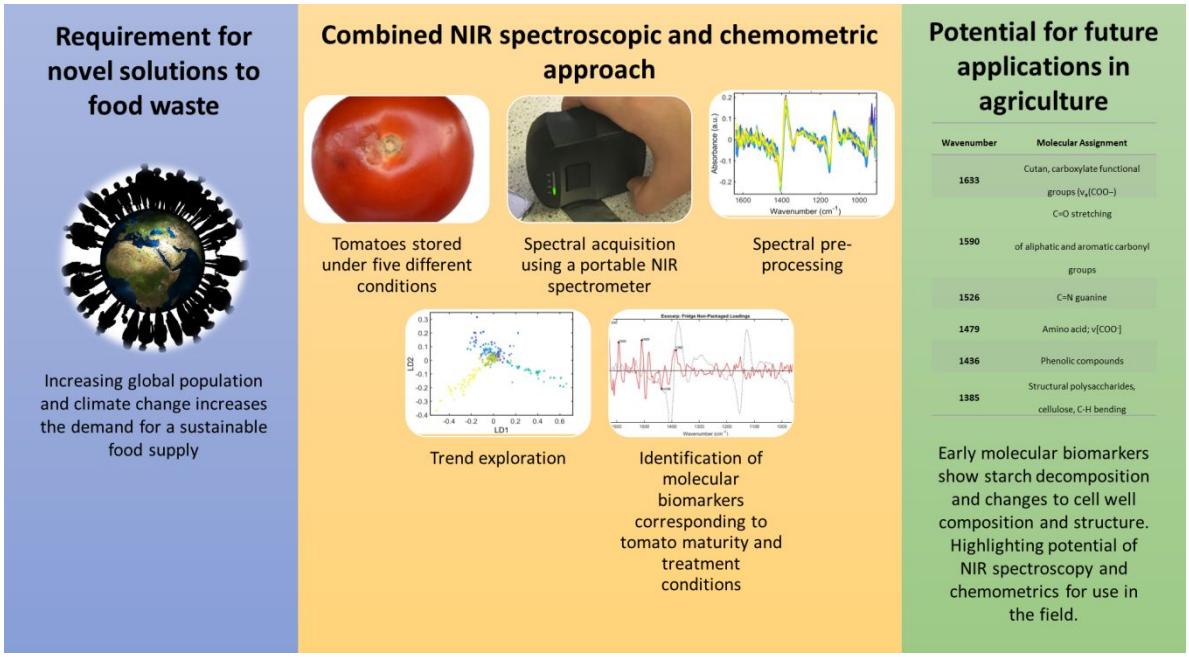
Abstract

Minimising food waste critical to future global food security. This study aimed to assess the potential of near-infrared (NIR) spectroscopy combined with machine learning to monitor the stability of tomato fruit during storage. Freshly harvested UK-grown tomatoes (*n*=135) were divided into five equally sized groups, each stored in different conditions. Absorbance spectra were obtained from both the tomato exocarp and locular gel using a portable NIR spectrometer, capable of connecting to a mobile phone, before subsequent chemometric analysis. Results show that support vector machines can predict the storage conditions and time-after-harvest of tomatoes. Molecular biomarkers highlighting key wavenumber and molecular changes due to time and storage conditions were also identified. This method shows potential for development of this approach for use in the field to help mitigate the environmental and economic impacts of food waste.

Keywords

Tomato, infrared spectroscopy, food security, machine learning, chemometrics, food storage

Table of Contents Graphic



Introduction

Sustainable food supply chains are required to provide adequate nutrition for the rising global population, which is projected to reach approximately 10 billion by 2050 [1]. Achieving food security and ending malnutrition are priorities in the United Nations sustainable development agenda for 2030 [2]. To achieve these goals using the same amount of available land, it is important that to maximise the efficiency of current supply chains; this includes the reduction of food waste. In the UK, food waste within the home is largely avoidable with 2.9 million metric tons, equating to £6.3 billion, of food spoiling before consumption [3]. Tomatoes are a popular crop globally, representing a versatile and nutrient-dense superfood, offering an excellent source of lycopene and other antioxidant molecules which benefit human health [4]. In 2017, 9.2 kg of fresh tomatoes and 33.2 kg of processed tomato products were consumed per capita in the United States of America [5]. However, fresh ripe tomatoes are prone to a high rate of food waste, with a total loss of 53.8% collectively from production to consumption [6]. This results from a 20% loss in agricultural production, 7% loss in processing and packaging, 10% distribution and retail loss and 31% loss by consumers [6]. The importance of reducing this food waste is emphasised during the current climate crisis, particularly as the production of tomatoes comes with a high environmental impact. They have been recognized as one of the most carbon-intensive food products due to electricity and fertiliser usage [7], and greenhouse grown tomatoes are estimated to have a carbon footprint of up to 10.1 kg CO_2-eq/kg [8]. Improving the efficiency of the supply chain for tomato, and agri-food in general, would reduce the economic and environmental impacts of food waste, and help us to meet the challenge of nourishing a growing population. Technical innovations to increase shelf-life have the potential to reduce the unnecessary disposal of home produce and hence reduce food waste [9]. Factors that affect the marketability and shelf-life of fresh ripe tomatoes include the duration that the fruits are safe for consumption, but also aesthetic qualities such as colour and firmness. As a climacteric fruit, the quality of fresh tomatoes can be improved after harvest. The ripening process involves changes to cell wall composition and thickness [10], and the conversion of storage

carbohydrates into monosaccharides [11]. These processes increase the softness and sweetness of the fruit, which in turn increases its appeal to consumers. However, if damaged or left to ripen for too long, tomatoes are prone to infections such as sour-rot [12]. Early biomarker determination to rapidly predict fruit ripening, and hence its potential shelf-life before the fruit are no longer saleable, could help to achieve the optimum balance in post-harvest ripening under different climacteric conditions.

Near-infrared (NIR; wavelengths from 800 to 2,500 nm) spectroscopy has been used extensively in various recent studies to quantitatively determine fruit quality or maturity in tomatoes [12–18]. The portability of new miniaturised spectrometers has opened up new applications for spectroscopic studies, such as agriculture, see [19–22]. Near infrared spectroscopy uses changes in molecular vibrations upon absorption of infrared light to gain information about the chemical composition of a sample. Biological materials preferentially absorb light in the fingerprint region (1800-900 cm^{-1}), which gives information about key biomolecules [23]. These include: lipids (C=O symmetric stretching at ~1,750 cm–1 and CH2 bending at ~1,470 cm–1), proteins (amide I at ~1,650 cm–1, amide II at ~1,550 cm–1 and amide III at ~1,260 cm–1), carbohydrates (CO-O-C symmetric stretching at ~1,155 cm–1), nucleic acid (asymmetric phosphate stretching at ~1,225 cm–1 and symmetric phosphate stretching at ~1,080 cm–1), glycogen (C-O stretching at ~1,030 cm–1) and protein phosphorylation (~970 cm–1) [23]. The current study utilised a handheld NIR spectroscopy device to determine tomato maturity and explore the effects of five different storage conditions over twenty days, varying in temperature and packaging. Spectral measurements were taken from both the exocarp and the locular gel within, allowing comparison of results from destructive and conservative techniques. A novel approach combining NIR using a handheld spectrometer and machine learning was used for classification and identification of key biomarkers. Chemometric methods included principal component analysis (PCA), PCA coupled with linear discriminant analysis (PCA-LDA) and support vector machines (SVM). Our results highlight key wavenumber changes associated with ripening and the effects of post-harvest climatic conditions, raising opportunity of developing this combined method for use in the field.

Materials and Methods

Samples and data acquisition

This study measured tomatoes over a period of twenty days under five different storage conditions (ambient packaged, ambient non-packaged, fridge packaged, fridge non-packaged, and incubated non-packaged). A total of 135 tomatoes were used in this study, twenty-seven per storage condition. All tomatoes were F1 hybrids, a first generation cross between two varieties, grafted onto a strong disease resistant rootstock and grown in Rockwool, UK (supplied by John Lane). They were harvested at ten weeks old on Friday 13th September 2019. The tomatoes used in this study were selected for uniformity. At the start of the study, they were all graded to be of similar size, colour, and ripeness, were undamaged, and had a consistent colour across the whole surface. They were then randomly sorted into treatment groups. Three tomatoes from each treatment condition were selected for exocarp sampling and labelled X, Y and Z. Each treatment group was stored in different conditions over the course of the experiment, including ambient packaged, ambient non-packaged, fridge packaged, fridge non-packaged and incubated non-packaged. The packaging used consisted of a sealed plastic freezer bag to simulate modified atmosphere packaging (MAP) although without the introduction of a protective gas mix into the bag. The ambient tomatoes were stored at room temperature (18°C), the fridge tomatoes at 3°C and the incubated tomatoes at 25°C.

Spectral absorbances were measured using a hand-held NIR spectrometer NIR-S-G1 (Allied Scientific Pro, Gatineau, Quebec, Canada) using ISC NIRScan software (raw spectra are shown in the Supporting Information, Figure S1). Spectra were measured from two types of sample; exocarp (the tomato surface), and locular gel (a gel that develops prior to ripening of the pericarp and exhibits a liquid-like consistency towards the terminal stage of ripening, see [24]). The spectrometer crystal was cleaned between measurements using isopropyl alcohol wipes (Bruker Optics, Coventry, UK), and each time background spectra were taken to account for ambient atmospheric conditions. For locular gel samples spectral acquisition took place on days 3, 5, 7, 10, 12, 14, 17, and 20. Each day three different tomatoes were destructively analysed to extract the locular gel, and two spectral

replicates per tomato were taken. This resulted in six locular gel spectra per day, forty-eight per treatment group, and 240 spectra in total. The locular gel was extracted by incision at two approximately equidistant locations around the equator. Fifty μL of this gel was collected using a Gilson pipette and transferred onto glass slides covered in aluminium foil. Before spectral acquisition of samples an aluminium foil standard spectrum was measured. For the exocarp measurements, the same three tomatoes were analysed throughout the whole time-course and subsequently returned to their original storage conditions. For exocarp samples, measurements were taken on days 1, 3, 5, 7, 10, 12, 14, 17, and 20. Ten spectral replicates of each tomato were taken, resulting 270 spectra per treatment group and 1350 spectra in total.

Data analysis and validation

The reflectance spectral data were imported and processed within MATLAB R2014b (MathWorks, Inc., Natick, MA, USA). Pre-processing and data analysis were performed using the PLS Toolbox version 7.9.3 (Eigenvector Research, Inc., Manson, WA, USA). The raw spectra were pre-processed by Savitzky-Golay smoothing (window of 7 points, 2nd order polynomial fitting) to improve the signal-to-noise ratio and standard normal variate (SNV) to correct for light scattering. The pre-processed data were also mean-centred before multivariate analysis.

Principal component analysis (PCA) was applied to the pre-processed spectral data for exploratory analysis. PCA is an exploratory analysis technique that reduces the spectral dataset into a small number of principal components (PCs) responsible for the majority of the original data variance [25]. Each PC is orthogonal to each other, and they are formed in a decreasing order of explained variance, so PC1 covers more variance than PC2, and so on. Each PC is composed of scores and loadings. The scores represent the variance on sample direction, thus being used to identify patterns of similarity between the samples, and the loadings represent the variance on wavelength direction, thus being used to identify possible spectral markers responsible for the scores pattern. Although being a robust exploratory analysis technique, PCA was not able to classify samples in an objective fashion, therefore, a supervised classifier was added and the samples were analysed by principal

component analysis with linear discriminant analysis [26]. For more rigorous classification, support vector machines (SVM) algorithm was applied to estimate the tomatoes time-after-harvest. SVM is a linear classifier with a non-linear step called the kernel transformation [27]. The kernel function transforms the data space to a feature space where samples can be better discriminated. Herein, the radial basis function (RBF) kernel was used and optimised via cross-validation. The pre-processed spectra were then randomly split into training (70%) and test (30%) sets, and a supervised classification model was constructed using support vector machines (SVM) to systemically predict the time-after-harvest regardless the storage condition. The training set was used to build the SVM training model, and the test set to evaluate its predictive ability. The SVM model was optimised by cross-validation venetian blinds with 10 data splits.

Metrics such as accuracy, sensitivity and specificity were calculated for the test set. For more than two-classes, these metrics are calculated individually per class; herein, the average for all classes is reported. The accuracy (AC) represents the total number of samples correctly classified considering true and false negatives; the sensitivity (SENS) represents the proportion of positives that are correctly classified; and the specificity (SPEC) represents the proportion of negatives that are correctly identified [28]. These metrics are calculated as follow:

$$\mathrm{AC}(\%) = \left(\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}\right) \times 100 \quad (1)$$

$$\mathrm{SENS}(\%) = \left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\right) \times 100 \quad (2)$$

$$\mathrm{SPEC}(\%) = \left(\frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right) \times 100 \quad (3)$$

where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives.

Results

Figure 1 shows the fingerprint spectra for a) exocarp and b) locular gel absorbances, and pre-processed spectra for c) exocarp and d) locular gel. The pre-processed spectra for the exocarp measurements in Figure 1c contain bands around 1000 nm (C-H stretching 3rd overtone), 1200 nm (C-H stretching 2nd overtone in fibre parameters such as cellulose and lignin), 1360 nm (small arm, R-O-H stretching 1st overtone in alcohol), 1450 nm (O-H stretching 1st overtone in water), and spectral differences at around 1700 nm (C-H stretching 1st overtone in glucose/lignin) [29,30]. The locular gel fingerprint spectra, Figure 1d, have their bands compressed by the strong water peak at 1450 nm [23].

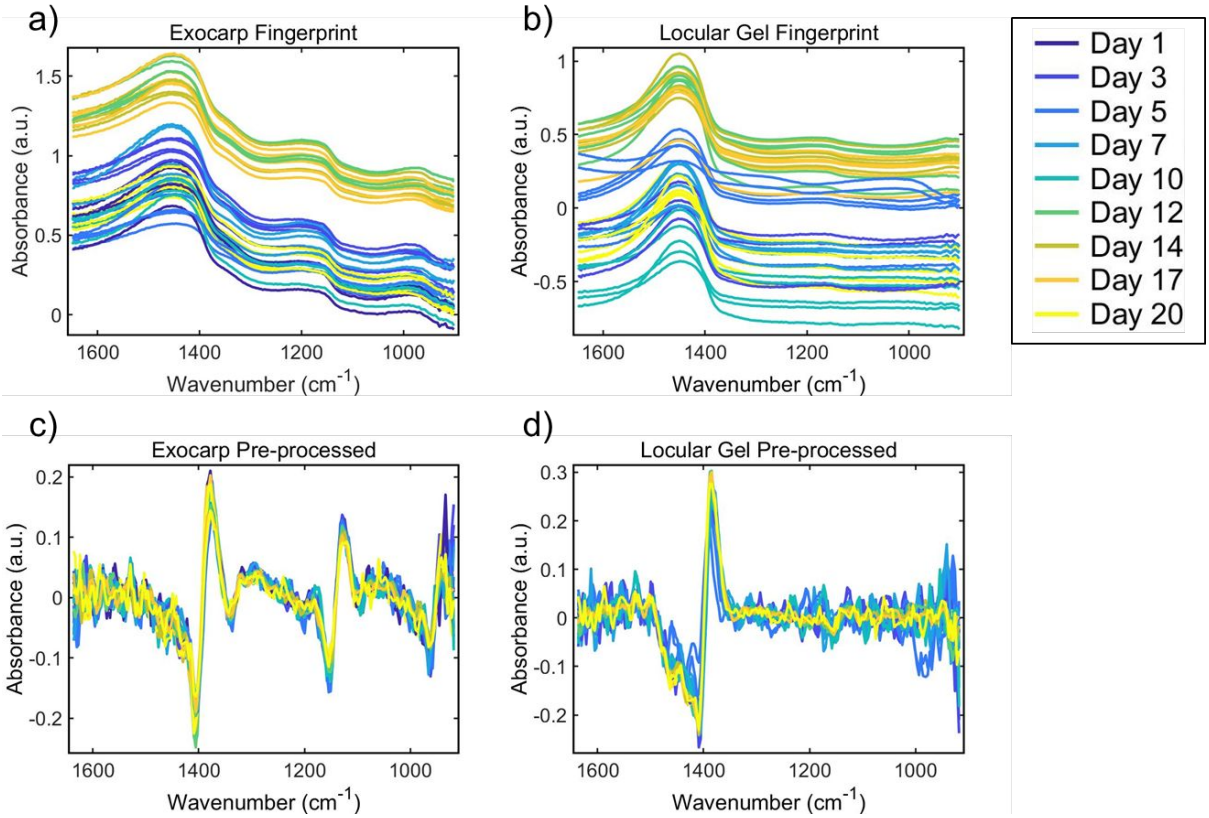


Figure 1: Fingerprint spectra for a) exocarp and b) locular gel absorbances, and pre-processed spectra for c) exocarp and d) locular gel absorbances. Each line is a class mean of spectra from a specific day and treatment condition. The colours represent the day the spectra were taken.

The pre-processed spectral data initially underwent unsupervised exploratory analysis by PCA, where overall segregation trends were observed in the data. Figure 2 shows PCA scatter plots for a)

exocarp and b) locular gel spectra. The colours represent the time in days and the marker shapes represent the treatment and numbers inside parenthesis represent the explained variance for each PC. PCA did not provide separation between the samples (Figures 2a and 2b). For PCA scores of each storage condition individually, see supporting information Figure S2 and S3, for exocarp and locular gel measurements respectively.

Following the initial PCA, a supervised method, linear discriminant analysis, was also applied. PCA-LDA scatter plots are shown in Figure 2c) for exocarp and 2d) for locular gel spectra. Although there is some evidence of a time evolution trend for the exocarp measurements (see Figure 2c), the samples cannot be differentiated into clear clusters. Some clustering was however achieved for locular gel samples. Figure 2d shows that days 1, 10 and 20 are separated out best in the locular gel PCA-LDA scatter plot along the axes LD1 and LD2.

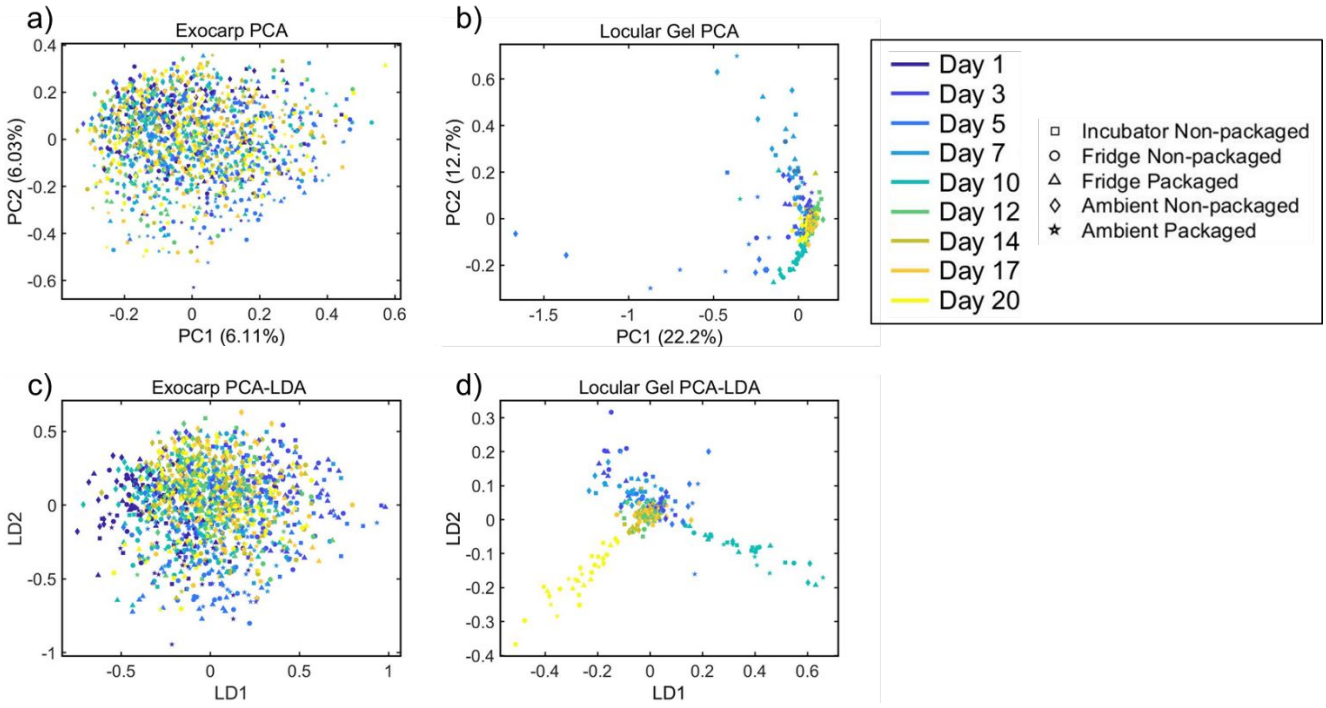


Figure 2: PCA scatter plots for a) exocarp and b) locular gel spectra. Numbers inside parenthesis represent the explained variance for each PC. PCA-LDA scatter plots for c) exocarp and d) locular gel spectra. The colours represent the time in days and the marker shapes represent the treatment.

Figure 3 shows the SVM model prediction results for each time and treatment sample type grouped separately. Overall, exocarp achieved higher accuracy sensitivity and specificity than locular gel spectra. For all forty-five exocarp and forty-one local gel categories, the SVM model achieved poor results at 51% and 54% accuracy, respectively. The SVM model to predict the time-after-harvest regardless the storage condition for the exocarp spectra achieved 92% accuracy, 86% sensitivity and 98% specificity in test sets. For locular gel spectra, the SVM model achieved 84% accuracy, 74% sensitivity and 95% specificity in test sets. Sensitivity of the exocarp spectra was - high, with many samples correctly classified. For locular gel samples, packaged fridge-stored samples at days 3 and 20, and ambient stored non-packaged samples at day 3 achieved higher true positive rates than the other groups. To give an overall picture of the trend over time, despite treatment, spectra from different storage conditions were also grouped together. The SVM results for this grouping can be viewed in the supporting information: see Figure S4 for optimisation parameters, Table S1 for classification metrics, and Table S2 for test set confusion matrices.

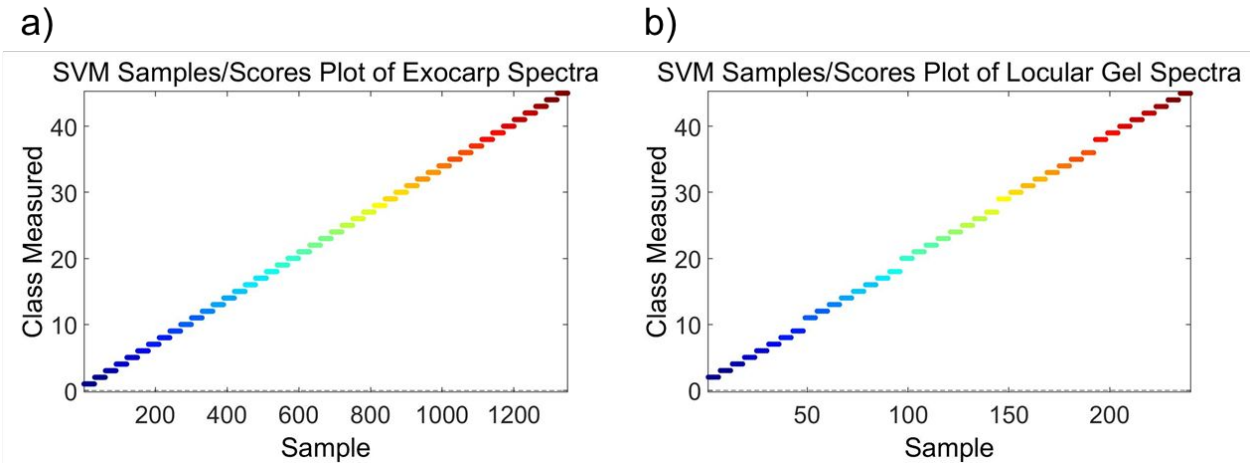


Figure 3: SVM predicted results for the test set for a) exocarp and b) locular gel. For all forty-five exocarp categories, the SVM model achieved 92% accuracy, 86% sensitivity and 98% specificity in test sets. For all forty-one locular gel categories, the SVM model achieved 84% accuracy, 74% sensitivity and 95% specificity in test sets. Each predicted category is shown in a different colour.

PCA loadings were used to determine the key wavenumbers associated with biochemical changes over time for each treatment, in both exocarp and locular gel spectra (see Tables 1 and 2, respectively). These spectral changes related to their biological origin by seeking molecular assignments from existing literature. Spectra from each treatment group were considered separately so that the changes highlighted in Tables 1 and 2 relate only to time.

Table 1: Loadings showing the key wavenumber changes for locular gel spectra associated with biochemical changes over time for each treatment

Sample Type	Treatment Type	Loadings	Assignment	Reference
Locular Gel	INP	1636	CHO stretching of carbonyl group, typical saccharide absorption	30
		1456	CH_3 bending vibration (lipids and proteins)	30
		1170	C-O bands from glycomaterials and proteins	30
		1095	Stretching PO_2 symmetric	30
		981	Phosphodiester region	30
		918	Polysaccharides	31
Locular Gel	FP	1605	v_{as} (COO^-) (polysaccharides, pectin)	30
		1571	C=N adenine	30
		1381	Amide II	32
		1075	Symmetric phosphate stretching modes or v(PO_2^-) sym.	30
		1012	Starch	32
		945	D-(-)-Arabinose	33
Locular Gel	ANP	1636	C=O stretching of carbonyl group, typical saccharide absorption	30
		1596	Methylated nucleotides	30
		1443	δ(CH) (polysaccharides, pectin), $\delta(CH_2)$, lipids, fatty acids	30
		1170	C-O bands from glycomaterials and proteins	30
		993	Arabinoxylans	34
		930	Polysaccharides	31
Locular Gel	AP	1636	C=O stretching of carbonyl group, typical saccharide absorption	30
		1493	Protein	35
		1422	Protein and lipids	35
		1124	Polysaccharides	36
		977	C-O-C stretching at the β-(1 → 4)-glycosidic linkages of amorphous cellulose	37
		918	Polysaccharides	31

Locular Gel	FNP	1609	Adenine vibration in DNA	[30]
		1574	C=N adenine	[30]
		1500	In-plane CH bending vibration from the phenyl rings, or Amide II (an N-H bending vibration coupled to C-N stretching)	[30]
		1236	Amide III and asymmetric phosphodiester stretching mode, $\nu_{as}(PO_2^-)$ mainly from the nucleic acids	[30]
		993	Arabinoxylans	[34]
		926	Polysaccharides	[31]

216

217

Table 2: Loadings showing the key wavenumber changes for exocarp spectra associated with biochemical changes over time for each treatment

Sample Type	Treatment Type	Loadings	Assignment	Reference
Exocarp	INP	1636	C=O stretching of carbonyl group, typical saccharide absorption	30
		1590	C=O stretching of aliphatic and aromatic carbonyl groups	30
		1542	Amide II	30
		1483	Protein and lipids	35
		962	Polysaccharides	31
		918	Polysaccharides	31
Exocarp	FP	1627	Phenolic compounds	38
		1584	Amide II	30
		1526	C=N guanine	30
		1456	CH_3 bending vibration (lipids and proteins)	30
		981	Phosphodiester region	30
		918	Polysaccharides	31
Exocarp	ANP	1636	C=O stretching of carbonyl group, typical saccharide absorption	30
		1574	C=N adenine	30
		1532	Stretching C=N, C=C	30
		1479	Amino acid; $\nu[COO^-]$	32
		1381	Amide II	32
		918	Polysaccharides	31
Exocarp	AP	1627	Phenolic compounds	38
		1581	Ring C-C stretch of phenyl	30
		1509	Hemicellulose, C-C and C=C	39
		1473	Glycerolipids, wax hydrocarbons, $\delta(CH_2)$ scissoring	14
		1409	Succinic acid	40
		933	Z type DNA	30
Exocarp	FNP	1630	Amide I	30
		1593	Protein	36
		1509	Hemicellulose, C-C and C=C	39
		1436	Phenolic compounds	38
		1385	Structural polysaccharides, cellulose, C-H bending	41
		926	Polysaccharides	31

To compare the effects of different storage treatments on spectral absorbances, data from the middle of the time course prior to any visible change in tomato quality were selected. Spectra from

day 10 were used to derive loadings showing which wavenumbers were associated with differing treatment conditions; these were subsequently connected with their corresponding biomarkers. To view loading graphs from which the information in Tables 1-3 were derived, see the supporting information Figure S5 for locular gel, Figure S6 for exocarp and Figure S7 for day ten.

Table 3: Loadings showing the key wavenumber differences between treatments at day 10 for locular gel and exocarp spectra

Sample Type	Day	Wavenumber	Molecular Assignment	Reference
Exocarp	Day 10	1633	Cutan, carboxylate functional groups ($\nu_a(COO^-)$)	38
		1590	C=O stretching of aliphatic and aromatic carbonyl groups	30
		1526	C=N guanine	30
		1479	Amino acid; $\nu[COO^-]$	32
		1436	Phenolic compounds	38
		1385	Structural polysaccharides, cellulose, C-H bending	41
Locular Gel	Day 10	1602	Pectin, phenolic compounds; C-C aromatic stretching; C-O-O^- asymmetric stretching	42
		1486	Proteins	43
		1440	Phenolic compounds; ν(C-C) aromatic (conjugated with C=C)	38
		1064	C–O stretching, C–C stretching (mannose-containing hemicellulose)	33
		993	Arabinoxylans	34
		933	Z type DNA	30

The different treatment conditions resulted in tomatoes of different qualities after 20 days (Figure 4). Visually, fridge non-packaged and ambient packaged treatments best preserved the quality of the tomatoes, whereas tomatoes from the incubator non-packaged and ambient non-packaged treatments showed signs of infection, and fridge packaged tomatoes showed signs of a possible chilling stress.

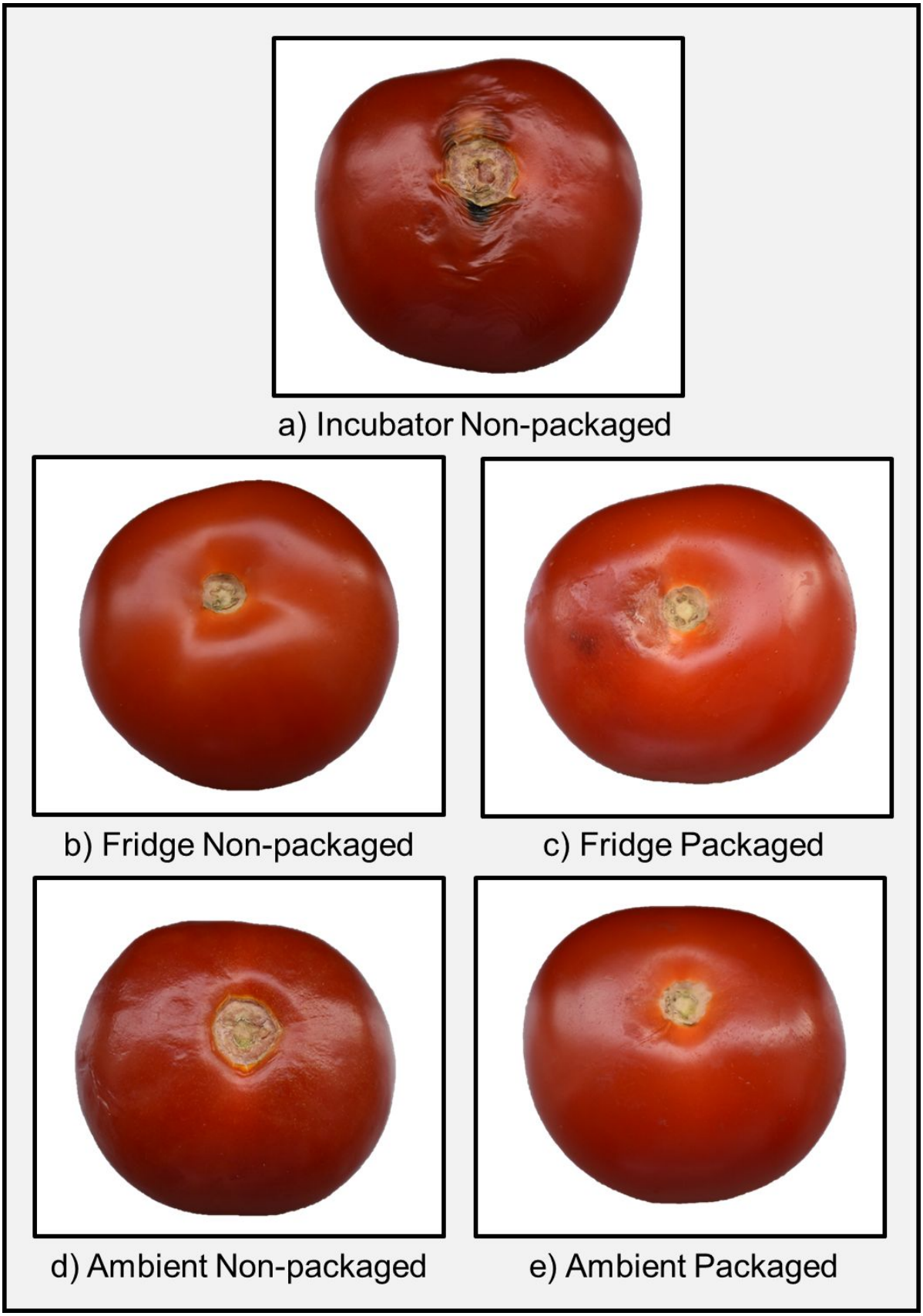


Figure 4: Images illustrating relative tomato quality between treatments on day 20

Discussion

SVM achieved outstanding classification accuracy across exocarp (92%) and locular gel (84%) samples to estimate the time-after-harvest despite the storage condition. The accuracy to predict all forty-five and forty-one categories of samples (including differences between storage conditions) could be improved if the model was built to differentiate between fewer categories, or a larger dataset was used. When more than ten classes are involved in the classification model its performance tends to reduce substantially since the dataset size is relatively reduced by a fraction of the number of classes when building the classifier on a one-against-the-others categories basis [ref.] Therefore, a very large dataset would be required to perform this type of classification. The models were better at identifying where not to place a spectrum but often failed to find the correct category to sort it into, with high specificities (60% and 91%) but low sensitivities (3% and 20%). When spectra from all storage condition treatments were grouped together, classification by time achieved excellent results with SVM; 92% average accuracy, 86% average sensitivity, and 98% average specificity in the test set for exocarp measurements, and 84% average accuracy, 74% average sensitivity, and 95% average specificity for the test set of locular gel samples, see supporting information Table S2.

PCA loadings identified key wavenumbers which allow discrimination between spectral absorbances from tomatoes of different storage conditions and maturities, see Tables 1, 2 and 3. These were connected to biomarkers using existing literature databases, allowing insight into the biochemical changes taking place. Many of the identified changes related to the progression of starch degradation, where the starch stored during development and is converted to soluble sugars, a key process during post-harvest fruit ripening [11]. Wavenumber 1012 cm^{-1} in locular gel fridge-stored spectra identified starch specifically as an important indicator of tomato age [32]. All treatment conditions for both locular gel and exocarp spectra identified that peaks for polysaccharides [31] and/ or saccharides [30] were used for the differentiation between tomatoes of different ages. This

indicates that all fruits underwent some level of post-harvest ripening, irrespective of storage conditions.

Other key peaks identified across different conditions were those relating to the structural and compositional development of the cuticle and cell wall. Compositional changes in key compounds such as pectin, cellulose, and other polysaccharides, as well as changes in cell wall thickness, are part of the ripening process [10]. Tomato maturity was indicated in the spectra of ambient packaged exocarp samples at peak 1473 cm^{-1}, which relates to the $\delta(CH_2)$ scissoring of cuticle glycerolipids and wax hydrocarbons [14]. Pectin was identified as a measure of tomato maturity in ambient non-packaged and fridge packaged samples from locular gel spectra [30], and also as a differentiator between storage conditions [42]. Cellulose was an indicator of tomato maturity in the locular gel of ambient packaged fruit [37]. However cellulose and hemicellulose was more commonly associated with exocarp spectra where these compounds differentiated between maturity in ambient packaged and fridge non-packaged samples [39]. Cellulose was a key differentiator between treatment conditions in exocarp spectra [41].

Dissolution of the cell wall is key to another ripening process, softening. Arabinogalactan proteins, which act as a cross-linker for pectin and arabinoxylan, are thought to be important in the alteration of cell wall assembly [44]. In this study, locular gel arabinoxylan levels were a key differentiator between treatment conditions and allowed maturity determination in fridge and ambient non-packaged fruit [34]. These results suggest that the packaging is altering the environment experienced by the tomatoes, and having an impact on arabinoxylan, and consequently fruit-softness. The texture changes can be seen in Figure 4, where the ambient non-packaged fruit appears softer than its packaged equivalent. Expression of arabinogalactan proteins in tomato fruit has also been linked to possible involvement in stress adaptations [44], which may be why the packaged tomatoes suffered from a chilling stress in the fridge whilst their unpacked equivalents appeared to be in better condition, see Figure 4. Fleshy fruit such as tomatoes are particularly vulnerable to chilling injury, is a

type of oxidative stress that occurs during storage below 10 °C [45]. Despite this chilling, fridge non-packaged tomatoes were amongst the best-preserved tomatoes alongside ambient packaged tomatoes, see Figure 4. The least well-preserved tomatoes were those not protected with packaging, stored under ambient room temperature or in the incubator, which showed signs of infection. The lack of packaging likely exposed the tomatoes to any pathogens present, and the high temperatures in the incubator conditions promoted growth. Bacteria and fungi grow best at temperatures between 25-30°C [46,47], and the incubated tomatoes were stored at 25°C.

This study has identified biomarkers, including those indicating tomato ripening, which are modified by storage and packaging conditions and that have the potential to be used to target reductions in the unnecessary disposal of tomatoes through spoilage in the supply chain. The rapid and non-destructive nature of exocarp scanning and the portability of the spectrometer, capable of connection to a mobile phone, renders this method particularly suitable for use within the food industry. With further development, this user-friendly and non-destructive technology displays potential for a wider range of applications within and beyond the food industry.

Acknowledgements

We would like to acknowledge John Lane, who supplied the tomatoes and Lancaster Girls' Grammar School for their support for the project.

Author Contributions

NEME, FLM and CLMM conceived the study; NEME, SG, RSB, FLM and CLMM jointly designed the experiment; NEME, SG, RSB conducted the experiment and collected the data; CLMM and CAH undertook the chemometric analyses; NEME, CAH, SG, RSB, CLMM jointly drafted the manuscript, MRM advised on the plant biology and all authors approved the final version; CR provided support and supervision for the conduct of the experiment; FLM was principal investigator.

Declaration of Interest

FLM is a major shareholder in Biocel UK Ltd, a company with interests in developing tools described herein for commercial gain.

References

1. Food and Agriculture Organization of the United Nations. High-level expert forum—How to feed the world in 2050. (2009).
2. Nino, F. S. Sustainable Development Goals—United Nations. *United Nations Sustain. Dev.* (2015).
3. Parfitt, J., Barthel, M. & MacNaughton, S. Food waste within food supply chains: Quantification and potential for change to 2050. *Philos. Trans. R. Soc. B Biol. Sci.* **365**, 3065–3081 (2010).
4. Story, E. N., Kopec, R. E., Schwartz, S. J. & Harris, G. K. An Update on the Health Effects of Tomato Lycopene. *http://dx.doi.org/10.1146/annurev.food.102308.124120* **1**, 189–210 (2010).
5. Wu, X., Yu, L. & Pehrsson, P. R. Are Processed Tomato Products as Nutritious as Fresh Tomatoes? Scoping Review on the Effects of Industrial Processing on Nutrients and Bioactive Compounds in Tomatoes. *Adv. Nutr.* (2021) doi:10.1093/ADVANCES/NMAB109.
6. Boz, Z. & Sand, C. K. A systematic analysis of the overall nutritional contribution of food loss and waste in tomatoes, spinach, and kidney beans as a function of processing. *J. Food Process Eng.* **43**, e13509 (2020).
7. Xue, L. *et al.* Mapping the EU tomato supply chain from farm to fork for greenhouse gas emission mitigation strategies. *J. Ind. Ecol.* **25**, 377–389 (2021).
8. Ntinas, G. K., Neumair, M., Tsadilas, C. D. & Meyer, J. Carbon footprint and cumulative energy demand of greenhouse and open-field tomato cultivation systems under Southern and Central European climatic conditions. *J. Clean. Prod.* **142**, 3617–3626 (2017).
9. Quested, T., Ingle, R. & Parry, A. Household food and drink waste in the United Kingdom 2012, Waste & Resources Action Programme (WRAP). *Banbury, UK* (2013).
10. Segado, P., Domínguez, E. & Heredia, A. Ultrastructure of the Epidermal Cell Wall and Cuticle of Tomato Fruit (Solanum lycopersicum L.) during Development. *Plant Physiol.* **170**, 935

(2016).

11. Fan, Z.-Q. *et al.* A banana R2R3-MYB transcription factor MaMYB3 is involved in fruit ripening through modulation of starch degradation by repressing starch degradation-related genes and MabHLH6. *Plant J.* **96**, 1191–1205 (2018).

12. Skolik, P., McAinsh, M. R. & Martin, F. L. ATR-FTIR spectroscopy non-destructively detects damage-induced sour rot infection in whole tomato fruit. *Planta* **249**, 925–939 (2019).

13. Clément, A., Dorais, M. & Vernon, M. Multivariate Approach to the Measurement of Tomato Maturity and Gustatory Attributes and Their Rapid Assessment by Vis−NIR Spectroscopy. *J. Agric. Food Chem.* **56**, 1538–1544 (2008).

14. Skolik, P., Morais, C. L. M., Martin, F. L. & McAinsh, M. R. Determination of developmental and ripening stages of whole tomato fruit using portable infrared spectroscopy and Chemometrics. *BMC Plant Biol.* **19**, 236 (2019).

15. Tiwari, G., Slaughter, D. C. & Cantwell, M. Nondestructive maturity determination in green tomatoes using a handheld visible and near infrared instrument. *Postharvest Biol. Technol.* **86**, 221–229 (2013).

16. Huang, Y., Lu, R. & Chen, K. Prediction of firmness parameters of tomatoes by portable visible and near-infrared spectroscopy. *J. Food Eng.* **222**, 185–198 (2018).

17. Sirisomboon, P., Tanaka, M., Kojima, T. & Williams, P. Nondestructive estimation of maturity and textural properties on tomato 'Momotaro' by near infrared spectroscopy. *J. Food Eng.* **112**, 218–226 (2012).

18. Saad, A., Gawad Saad, A., Jaiswal, P. & Narayan Jha, S. Non-destructive quality evaluation of intact tomato using VIS-NIR spectroscopy Related papers Non-destructive quality evaluation of intact tomato using VIS-NIR spectroscopy. *Int. J. Adv. Res.* **2**, 632–639 (2014).

19. Nolasco Perez, I. *et al.* Classification of Chicken Parts Using a Portable Near-Infrared (NIR) Spectrophotometer and Machine Learning. *Appl. Spectrosc.* **72**, 1774–1780 (2018).

20. Kaufmann, K. *et al.* Portable NIR Spectrometer for Prediction of Palm Oil Acidity. *J. Food Sci.*

84, 406–411 (2019).

21. Oliveira, M. M., Cruz-Tirado, J. P., Roque, J. V, Teófilo, R. F. & Barbin, D. F. Portable near-infrared spectroscopy for rapid authentication of adulterated paprika powder. *J. Food Compos. Anal.* **87**, 103403 (2020).

22. Beć, K. B., Grabska, J. & Huck, C. W. Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. *Chem. – A Eur. J.* **27**, 1514–1532 (2021).

23. Morais, C. L. M., Lima, K. M. G., Singh, M. & Martin, F. L. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nature Protocols* vol. 15 2143–2162 (2020).

24. Atta-Aly, M. A., Brecht, J. K. & Huber, D. J. Ripening of tomato fruit locule gel tissue in response to ethylene. *Postharvest Biol. Technol.* **19**, 239–244 (2000).

25. Bro, R. & Smilde, A. K. Principal component analysis. *Anal. methods* **6**, 2812–2831 (2014).

26. Morais, C. L. M. & Lima, K. M. G. Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. *Artic. J. Braz. Chem. Soc* **29**, 472–481 (2018).

27. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).

28. Morais, C. L. M. & Lima, K. M. G. Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data. *Chemom. Intell. Lab. Syst.* **170**, 1–12 (2017).

29. Jerome, J. & Workman, J. Interpretive Spectroscopy for Near Infrared. *http://dx.doi.org/10.1080/05704929608000571* **31**, 251–320 (2006).

30. Talari, A. C. S., Martinez, M. A. G., Movasaghi, Z., Rehman, S. & Rehman, I. U. Advances in Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Appl. Spectrosc. Rev.* **52**, 456–506 (2017).

31. Di Giambattista, L. *et al.* New marker of tumor cell death revealed by ATR-FTIR spectroscopy. *Anal. Bioanal. Chem.* **399**, 2771–2778 (2011).

32. Luo, Y. *et al.* Diagnostic Segregation of Human Breast Tumours Using Fourier-transform

Infrared Spectroscopy Coupled with Multivariate Analysis: Classifying Cancer Subtypes. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **255**, 119694 (2021).

33. Liu, X., Renard, C. M. G. C., Bureau, S. & Le Bourvellec, C. Revisiting the contribution of ATR-FTIR spectroscopy to characterize plant cell wall polysaccharides. *Carbohydr. Polym.* **262**, 117935 (2021).

34. Demir, P., Onde, S. & Severcan, F. Phylogeny of cultivated and wild wheat species using ATR-FTIR spectroscopy. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **135**, 757–763 (2015).

35. Meinen, C. & Rauber, R. Root discrimination of closely related crop and weed species using FT MIR-ATR spectroscopy. *Front. Plant Sci.* **6**, 765 (2015).

36. Butler, H. J., McAinsh, M. R., Adams, S. & Martin, F. L. Application of vibrational spectroscopy techniques to non-destructively monitor plant health and development. *Anal. Methods* **7**, 4059–4070 (2015).

37. Bekiaris, G. *et al.* Rapid estimation of sugar release from winter wheat straw during bioethanol production using FTIR-photoacoustic spectroscopy. *Biotechnol. Biofuels 2015 81* **8**, 1–12 (2015).

38. Heredia-Guerrero, J. A. *et al.* Infrared and Raman spectroscopic features of plant cuticles: a review. *Front. Plant Sci.* **0**, 305 (2014).

39. Zhou, C., Jiang, W., Via, B. K., Fasina, O. & Han, G. Prediction of mixed hardwood lignin and carbohydrate content using ATR-FTIR and FT-NIR. *Carbohydr. Polym.* **121**, 336–341 (2015).

40. Kang, S., Amarasiriwardena, D. & Xing, B. Effect of dehydration on dicarboxylic acid coordination at goethite/water interface. *Colloids Surfaces A Physicochem. Eng. Asp.* **318**, 275–284 (2008).

41. Rana, R. *et al.* Leaf Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) biochemical profile of grassland plant species related to land-use intensity. *Ecol. Indic.* **84**, 803–810 (2018).

42. Moskal, P., Wesełucha-Birczyńska, A., Łabanowska, M. & Filek, M. Adaxial and abaxial pattern

of Urtica dioica leaves analyzed by 2DCOS ATR-FTIR as a function of their growth time and impact of environmental pollution. *Vib. Spectrosc.* **104**, 102948 (2019).

43. Jin, N. *et al.* Spectrochemical analyses of growth phase-related bacterial responses to low (environmentally-relevant) concentrations of tetracycline and nanoparticulate silver. *Analyst* **143**, 768–776 (2018).

44. Leszczuk, A., Kalaitzis, P., Blazakis, K. N. & Zdunek, A. The role of arabinogalactan proteins (AGPs) in fruit ripening—a review. *Hortic. Res. 2020 71* **7**, 1–12 (2020).

45. Stevens, R. *et al.* Tomato fruit ascorbic acid content is linked with monodehydroascorbate reductase activity and tolerance to chilling stress. *Plant. Cell Environ.* **31**, 1086–1096 (2008).

46. Jay, J. M., Loessner, M. J. & Golden, D. A. Low-temperature food preservation and characteristics of psychrotrophic microorganisms. in *Modern food microbiology* (Springer Science & Business Media, 2008).

47. Adams, M. R. & Moss, M. O. The microbiology of food preservation. (2020).