Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees

Weijie Lan, Vincent Baeten, Benoit Jaillais, Catherine Renard, Quentin Arnould, Songchao Chen, Alexandre Leca, Sylvie Bureau

▶ To cite this version:

Weijie Lan, Vincent Baeten, Benoit Jaillais, Catherine Renard, Quentin Arnould, et al.. Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees. Journal of Food Engineering, 2022, 323, pp.111002. 10.1016/j.jfoodeng.2022.111002 . hal-03604054

Comparison of near-infrared, mid-infrared, Raman spectroscopy and near-infrared hyperspectral imaging to determine chemical, structural and rheological properties of apple purees

Weijie Lan[a], Vincent Baeten[b], Benoit Jaillais[c], Catherine M.G.C. Renard[a,d], Quentin Arnould[b], Songchao Chen[e], Alexandre Leca[a], Sylvie Bureau[a]*

[a] INRAE, Avignon University, UMR408 Sécurité et Qualité des Produits d'Origine Végétale, F-84000 Avignon, France.

[b] Walloon Agricultural Research Centre (CRA-W). Quality and Authentication of Products Unit, Valorisation of Agricultural Products Department. 24. BE-5030 Gembloux, Belgium.

[c] INRAE, ONIRIS, Unité Statistiques, Sensométrie, Chimiométrie (StatSC), F-44322 Nantes, France.

[d] INRAE, TRANSFORM Division, F-44000 Nantes, France.

[e] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311200, China.

Corresponding author*

Sylvie Bureau (E-mail: sylvie.bureau@inrae.fr).

INRAE, UMR408 SQPOV « Sécurité et Qualité des Produits d'Origine Végétale »
228 route de l'Aérodrome CS 40509

F-84914 Avignon cedex 9

Tel: +33 432722509

Other authors

Catherine M. G. C Renard: catherine.renard@inrae.fr

Vincent Baeten: v.baeten@cra.wallonie.be

Benoit Jaillais: benoit.jaillais@inrae.fr

Quentin Arnould: q.arnould@cra.wallonie.be

Weijie Lan: Weijie.Lan@inrae.fr

Alexandre Leca: Alexandre.Leca@inrae.fr

Songchao Chen: chensongchao@zju.edu.cn

Highlights

MIR provided a better discrimination of puree variability than other techniques.

MIR gave the best prediction of puree textural and rheological properties.

HSI technique had a better ability to assess puree quality and variability than NIR.

Raman spectroscopy could not provide sufficient assessment of puree quality.

Abstract

Near-infrared (NIR), mid-infrared (MIR), Raman spectroscopy and hyperspectral imaging (HSI) were comprehensively compared for their capacity to evaluate the composition and texture characteristics of apple purees issued from a large variability (cultivar, fruit thinning, post-harvest mealy texture and processing). NIR, MIR and HSI techniques had a good ability to estimate puree composition such as soluble solids (RPD > 2.5), titratable acidity (RPD > 2.4) and dry matter (RPD > 2.3). Raman spectroscopy was less accurate to determine puree biochemical (RPD < 1.8) and textural parameters (RPD < 1.4) than the other techniques. MIR was the best tool to identify aforementioned factors (> 91.7 % of correct classification) and to satisfactory predict the puree average particle size (RPD = 2.9), viscosity (RPD ≥ 2.1) and viscoelasticity (RPD > 2.3). Consequently, NIR, MIR and HSI should be prioritized as process analytical technologies to detect the variability of purees and assess their texture and taste.

Keywords:

Malus x domestica Borkh.; Infrared spectroscopy; Raman; Process analytical technique; Puree quality.

1. Introduction

Apple puree is the basic ingredient of many fruit-based products, such as jams, preserves or compotes, yogurts and pie fillings for food industry (Defernez, Kemsley, & Wilson, 1995). It appears to be particularly suitable to test candidate process analytical techniques (PATs) as there are clear levers to introduce controlled variability in a sample set either from raw material or from process conditions (Lan, Jaillais, Leca, Renard, & Bureau, 2020). Today, apple purees are predominantly analyzed by chromatography and specific rheometers to determine their biochemical (Keenan, Brunton, Butler, Wouters, & Gormley, 2011) and rheological properties (Buergy, Rolland-Sabaté, Leca, & Renard, 2020; Espinosa-Muñoz, Renard, Symoneaux, Biau, & Cuvelier, 2013). These methods provide accurate quantifications, but they are time-consuming, expensive and not suitable for fast and numerous characterizations.

Developing highly efficient, economic and reliable PATs is a key point for food quality control in industrial and scientific works. Spectroscopic and imaging techniques have been considered to be some of the representative PATs for the rapid qualification of agricultural commodities and processed food (Cullen, O'Donnell, & Fagan, 2014). In particular, near-infrared (NIR), mid-infrared (MIR), Raman and hyperspectral imaging (HSI) offer the advantages of a minimal sample preparation and a rapid data acquisition.

NIR technique has been widely applied for the safety inspection and quality assessment of apple fruits at the wavelength range from 780-2500 nm (Nicolai, et al., 2007; Pissard, et al., 2013). The broad bands of NIR contain the overlapping absorption bands corresponding mainly to overtones and combinations of vibrational mode C-H and O-H bonds of fruit components (Osborne, 2006). Several internal attributes of apple purees, such as soluble solids content (SSC), dry matter content (DMC) and titratable acidity (TA), can thus be evaluated from a single spectrum, with acceptable precisions (Lan, Jaillais, Leca, Renard, & Bureau, 2020).

MIR spectroscopy on fresh and processed apples gives a good estimation of SSC, DMC, TA, malic acid and some individual sugars (Bureau, et al., 2013; Lan, Renard, Jaillais, Leca, & Bureau, 2020). Compared to the low structural selectivity in the broad bands of NIR spectra, more resolved fundamental bands of MIR spectra allow to better elucidating the chemical and structural information of samples. However, the lower energy of MIR radiations and the strong water interactions in fruit suspensions prevent the sensitive evaluation of chemical compositions and structural properties (Lan, Renard, et al., 2020).

Raman spectroscopy can provide a complementary interpretation of molecule vibration changes in polarizability, which is distinct from the vibration used in MIR by the changes in dipole moment (Pistorius, 1996). For highly hydrated products, such as fresh and processed fruits, Raman presents two advantages in comparison with infrared: a weak scattering of the polar O-H group and more intense bands of homo-nuclear molecular bonds (C-C, C=C, etc.). To date, no detailed study has compared the differences and limitations of Raman and infrared spectroscopy (NIR and MIR) to determine the structural and rheological properties of fruit purees.

Hyperspectral imaging (HSI) is an emerging platform technique that integrates imaging and spectroscopy to provide both spatial and spectral information (Baeten & Dardenne, 2005). Several applications of HSI were carried out on fresh fruits to estimate their external and internal quality (Baeten & Dardenne, 2005; Mendoza et al., 2011; Ma et al., 2018). For fruit processed purees, no work has been done on HSI to detect their biochemical composition, structural and rheological properties.

To date, the comprehensive comparison of these techniques to determine chemical (SSC, TA, DMC, individual sugars and malic acid), structural (particle sizes) and rheological (viscosity and viscoelasticity) characteristics of fruit puree products stays limited. Therefore, identifying the most efficient spectroscopic method to assess quality of processed fruit purees is a crucial point to prioritize further developments.

In this work, four different spectroscopic and imaging techniques, namely NIR, MIR, Raman and HSI, were applied on the same set of diverse (cultivar, fruit thinning practice, fruit texture, processing) apple puree samples in order to: i) evaluate their potential to detect the puree variability; ii) compare their performance to predict chemical, structural and rheological characteristics of purees and then iii) identify signals specific of the puree properties.

2. Materials and methods

2.1 Apple purees

2.1.1 Apples

A large variability of apples has been introduced in this work, in order to explore the potential of different spectroscopic techniques to detect the variability of the processed apple purees. Around 50 kg of Apples of four cultivars: 'Golden Delicious' (GD), 'Granny Smith' (GS), 'Royal Gala' (GA) and 'Braeburn' (BR, BM) were harvested at a commercial maturity in 2018 from the La Pugère Experimental Orchard (Chambre d'Agriculture des Bouches du Rhône) (Mallemort, Bouches du Rhône, France).

Fruit thinning generates significant differences of apple cell numbers during growth (Link, 2000), and results in intensive variations of puree structural and chemical properties (Buergy, Rolland-Sabaté, Leca, & Renard, 2020; Lan, Jaillais, et al., 2020). In this study, GS, GA, BR, BM and half of GD apples were grown under a standard chemical fruit thinning practice (Th+) with 50-100 fruits / tree. The other half of GD apples was non-thinned (Th-) with 150-200 fruits / tree.

After harvesting, different storage conditions (temperature, time, humidity etc.) can strongly influence apple physical, structural, and biochemical properties (Tu et al., 2000). Four apple groups (GD Th+, GD Th-, GS, GA) were stored at 4 °C in normal atmosphere to ensure starch regression (customised phytotron, Froid et Mesures, Beaucouzé, France). As post-harvest storage is known to particularly affect the texture of Braeburn apples (Tu et al., 2000), two different storage conditions were applied specifically on Braeburn apples, resulting either in crunchy Braeburn apples (BR; stored at 4 °C in normal atmosphere), or mealy Braeburn apples (BM; kept for 11 days at 23 °C and at around 90% relative humidity).

Totally, six apple groups (GD Th-, GD Th+, GS, GA, BR and BM) were used for puree processing (**Fig. 1**).

2.1.2 Purees processing

For all apple groups, three replicates of apple purees were processed from 3 kg of apples each. After sorting and washing, apples (3 kg) were cored, and sliced into 12 portions, then processed under vacuum by a multi-functional processing system (RoboQbo Qb8-3, Bentivoglio, Italy), following two different processing recipes:

- I) ground at 3000 rpm for 202 s during the increase of temperature and heated at 70 °C for 15 min, then pasteurized at 95 °C for 2 min;

-II) ground at 3000 rpm for 360 s during the temperature increase step, followed by 400 rpm at 95 °C for 17 min.

Afterwards, half of each processed puree was refined at 0.5 mm using a Robot Coupe C80 automatic refiner (Robot Coupe C80, Robot Coupe SNC, Vincennes, France) and the other was not refined. Finally, all processed apple purees were conditioned in hermetically sealed cans, then cooled at 23 °C before the measurements performed the day after. In total, 72 puree samples (6 apple groups × 2 processing recipes × 2 refining levels × 3 processing replicates) were obtained (**Fig. 1**).

2.2 Determination of quality traits

2.2.1 Rheological and structural analyses

The puree rheological measurements, consisting in rotational (flow curve) and oscillatory (amplitude sweep) tests, were carried out using a Physica MCR-301 controlled stress rheometer (Anton Paar, Graz, Austria) equipped with a vane measuring system with a 3.46 mm gap (CC27/S cup and FL100/6W bob, Anton Paar), at 22.5 °C.

The flow curves were performed after a pre-shearing period of 1 minute at 50 s^{-1} followed by 5 minutes at rest. The viscosity was then measured at a controlled shear rate range of [10; 250] s^{-1} on a logarithmic ramp, at a rate of 1 point every 15 seconds. The complete flow curves were fitted with a power law model according to the previous works (Lan, Jaillais, et al., 2020), as described by Eq. (1).

$$\eta = K \ \dot{\gamma}^{n-1} \qquad (Eq1)$$

where η is the apparent viscosity (Pa.s), $\dot{\gamma}$ the shear rate (s^{-1}), K the consistency parameter, and *n-1* the flow parameter.

Amplitude sweep tests were performed at an angular frequency of 10 rad.s^{-1} in the deformation range of [0.01; 100]%, in order to determine the linear viscoelastic range of the purees and the yield stress, defined as the crossing point between the storage modulus (G’) and the loss modulus (G”) curves. The damping factor $\tan\delta = G"/G'$ of purees was calculated.

The particle sizes were measured according to our previous work (Lan, Jaillais, et al., 2020). Puree samples were diluted in distilled water to separate particles and stained with calcofluor white at 0.1 g/L and highlighted with a 365 nm UV lamp. A high-resolution digital video camera (Baumer VCXU31C, Baumer SAS, Fillinges, France) with a macro lens (VSTech 0513, VS Technology Corporation, Tokyo, Japan.) was used to visualize the distribution and dispersion of puree particles. The particle sizes averaged over volume d(4:3) (de Brouckere mean) and over surface area d(3:2) (Sauter mean) were measured with a laser granulometer (Mastersizer 2000, Malvern Instruments, Malvern, UK).

2.2.2 Biochemical analyses

Several biochemical analyses were performed on apple purees based on the previous works (Bureau et al., 2013; Lan, Jaillais, et al., 2021). SSC was determined with a digital refractometer (PR-101 ATAGO, Norfolk, VA, USA) and expressed in °Brix at 23 °C. TA was determined by titration up to pH 8.1 with 0.1 mol/L NaOH and expressed in mmol H^+/kg of fresh weight (FW) using an autotitrator (Methrom, Herisau, Switzerland). Individual sugars and malic acid were quantified using colorimetric enzymatic kits, according to the manufacturer's instructions (R-biopharm, Darmstadt, Germany). The content of glucose, fructose, sucrose and malic acid were expressed in g/kg of FW. These measurements were performed with a SAFAS flx-Xenius XM spectrofluorimeter (SAFAS, Monaco) at 570 nm for sugars and 450 nm for malic acid. DMC was estimated from the weight of freeze-dried samples upon reaching a constant weight (freeze-drying for 5 days). Cell wall materials (AIS) of purees were isolated using the alcohol insoluble solids method proposed by Renard (2005) and the cell wall contents (AIS contents) were expressed in FW.

2.3 Spectral and image data acquisition

2.3.1 NIR spectroscopy

NIR spectra were collected with a multi-purpose analyzer (MPA) spectrometer (Bruker Optics®, Wissembourg, France) at 23 °C. Puree samples were transferred into 10 mL glass vials (5 cm height × 18 mm diameter) which were placed on the automated sample wheel of the spectrophotometer. Logarithmic transformed reflectance spectra (log (1/R)) were acquired with a spectral resolution of 8 cm^{-1} from 12500 to 4000 cm^{-1} (corresponding to wavelengths from 800 to 2500 nm). Each spectrum corresponded to the average of 32 scans. The spectral acquisition and instrument adjustments were controlled by OPUS software Version 5.0 (Bruker Optics®, Ettlingen, Germany). A reference background measurement was automatically acquired before each data set acquisition using an internal Spectralon reference. Each puree sample was measured randomly three times on different aliquots. The mean of three replicate scans for each sample was calculated, and finally 72 NIR spectra of different apple purees (6 apple groups × 2 processing recipes × 2 refining levels × 3 processing replicates) were used in subsequent chemometric analysis.

2.3.2 MIR spectroscopy

MIR spectra of purees were acquired at 23 °C using a Tensor 27 FTIR spectrometer (Bruker Optics®, Wissembourg, France) equipped with a horizontal attenuated total reflectance (ATR) sampling accessory and a deuterated triglycine sulphate (DTGS) detector. The purees were placed at the surface of a zinc selenide (ATR-ZnSe) crystal with six internal reflections. Spectra with 32 scans each were collected from 4000 cm^{-1} to 800 cm^{-1} with a 4 cm^{-1} resolution and were corrected against the background spectrum of air. Three replications of spectral measurement were performed randomly on each puree, and these averaged MIR spectra of the 72 samples were used for further analysis.

2.3.3 Raman spectroscopy

Raman spectra were acquired on a Confocal Raman Microscope Senterra II spectrometer (Bruker Optics, Ettlingen, Germany) with a 785 nm diode laser and a thermoelectrically cooled CCD detector, operating at -65 °C. For spectra collection, each puree sample was manually placed and compacted in 36 holes (those in the middle) of a 96 well aluminium plate (12 × 8) with an inner diameter of 6 mm each. After removing the water of purees by evaporation at the ambient temperature (~20 °C), spectra were accumulated with a bleaching of 20 s, an integration time of 2 s and 7 coadditions using a 100 mW laser. Raman intensity were recorded from 50 to 3650 cm^{-1} with a spectral resolution of 4 cm^{-1} intervals. OPUS 7.8 Software (Bruker Optics, Ettlingen, Germany) was used for spectral data acquisition. Each sample was independently and randomly scanned six times. The final spectrum of each puree was the average of these 6 replicates, resulting in 72 Raman spectra.

2.3.4 HSI acquisition

The hyperspectral images of apple purees were acquired on a pushbroom (a line-scanning type) near infrared hyperspectral imaging system (SPECIM, Oulu, Finland), which consisted of a SWIR camera (SWIR-CL-400-N25E, SPECIM) covering the spectral range of 900-2500 nm with a spectral resolution of about 12 nm, an OLES 56 camera lens (SPECIM), an illumination source (halogen lamps) and a translating scanner. Before measurements, the reflectance calibration was performed based on our previous work (Lan, Jaillais, et al., 2021). All the image acquisition parameters (exposure time of camera, scanning speed etc.) were controlled by the LUMO® software from SPECIM. Each puree sample was placed on a hole (with an inner diameter of 3 cm) of the standard white plate (nine holes totally). All images were acquired in the reflectance mode and the final image size for each kernel is 387 × 127 × 288, the two first values representing pixel dimensions in the x and y directions (field of view of 9.5 × 3.1 cm, with a spatial resolution of 245 μm) and the third value accounting for the number of spectral channels. As the beginning and ending wavelengths contained noise caused by the instrument itself, the 258 bands from 990 to 2450 nm were selected for further spectral analysis. The averaged HSI spectrum of each puree sample was calculated and finally 72 HSI spectra were used for further discrimination and regression analyses.

2.4 Statistical analyses of reference data

After checking for normal distribution with a Shapiro-Wilk test (α=0.05), the reference data of processed purees are presented as mean values and the data dispersion within our experimental dataset expressed as standard deviation values (SD) (**Table S1**). Analysis of variance (ANOVA) was carried out to determine the significant differences due to the different apple cultivars, process recipes and mechanical refining treatments (**Table S1**) using XLSTAT (version 2018.5.52037, Addinsoft SARL, Paris, France) data analysis toolbox. Principal component analysis (PCA) was carried out on all reference data of processed purees to evaluate their discriminant contributions using Matlab 7.5 software using the SAISIR package (Cordella & Bertrand, 2014).

2.5 Chemometric analysis

NIR, MIR, Raman and HSI spectra were pre-processed with Matlab 7.5 software using the SAISIR package (Cordella & Bertrand, 2014). The discriminant analysis and multivariate regression were performed with several packages of the R software (version 4.0.2) (R Core Team, 2019), as detailed in our previous work (Lan, Bureau, et al., 2021).

Several different preprocessing methods have been performed on NIR, MIR, Raman and HSI spectral metrics. Particularly, smoothing (Savitzky Golay algorithm with a window size of 3, 13, 23 variables) referred to the numerical operations on puree spectra in order to reduce the noise. Standard normal variate (SNV) performed a normalization of all puree spectra that consists in subtracting each spectrum by its own mean and dividing it by its own standard deviation. Savitzky-Golay derivates (the first or second derivatives with gap sizes of 11, 21, 31, 41) were used to resolve overlapping puree spectral signals and enhance signal properties. All these methods and their combinations (smoothing + SNV, SNV + first derivation, SNV + second derivation, smoothing + SNV + first derivation, smoothing + SNV + second derivation, as well as the direct processing of the raw spectra) were used to pretreat the spectra for discrimination and regression, to compare and obtain the best results.

After several pretests, smoothing (Savitzky Golay algorithm with a window size of 13 variables) with SNV transformed NIR data in 800-2500 nm; the SNV pre-processed MIR spectra in 1800-900 cm^{-1}; the smoothing with SNV (a window size of 13 variables) of Raman in 1800-800 cm^{-1} and the SNV with 3 windows (a window size of 3 variables) smoothed HSI data in 990-2450 nm had the best performances to classify and assess the puree quality and were retained for further analysis.

Partial least squares (PLS) regression, a typical linear algorithm, combines principal component analysis and canonical correlation analysis (Geladi & Kowalski, 1986). In short, PLS models maximize the covariance between Y- matrix (reference datasets) and X- matrix (spectra dataset) in a way to have better predictions of Y- matrix by maximizing the variance of X-matrix. It has been successfully used to determine the global quality parameters of apple purees using NIRS information (Lan, Jaillais, Leca, Renard, & Bureau, 2020).

Random forest (RF) is an ensemble of learning methods for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time (Ho, 1995). For classification tasks, the output of RF is to identify a class selected by most trees. For regression tasks, the mean or average prediction of individual trees is returned.

Support vector machine (SVM) has been introduced for predicting numerical property values. SVM can efficiently perform a non-linear classification using what is called the ‘kernel trick’, implicitly mapping their inputs into high-dimensional feature spaces. Besides, SVM regression models can resolve nonlinear relationships in original feature spaces through dimensionality extension (Noble, 2006).

These two machine learning approaches (RF and SVM) have been specially constructed to address large and complex nonlinear systems (Liu, Wang, Wang & Li, 2013) and have provided satisfactory estimation of puree rheological properties (Lan, Bureau, et al., 2021). In this study, PLS, SVM and RF algorithms were used to discriminate purees (**Part 3.2**) and predict their quality traits (**Part 3.3**). The 10-fold full cross-validation was applied to the 72 spectra of NIR, MIR, Raman and HSI datasets, respectively.

For discrimination models (PLS-DA, SVM-DA and RF-DA), the discrimination accuracy (acc) was used to describe the discriminating ability of the different spectroscopic techniques (**Table 2** and **Table 3**). The ability of the four different techniques coupled with PLS-DA, SVM-DA and RF-DA was compared to classify different factors: (a) cultivars (48 purees from (Th+) GD, GA, GS and BR apples), (b) process recipes (72 samples of processes I and II), (c) refining treatments (72 NR and Ra), (d) fruit thinning practices (24 GD purees from Th+ and Th-) and (e) fruit stress treatments (24 Braeburn purees with crunchy BR and mealy BM). The main vibrational bands observed in NIR, MIR, Raman and HSI datasets, which contributed to the best discrimination models are shown for all factors (a-e) (**Table 3)**.

Prediction of puree rheological (K, n, G', G'', yield stress and tan δ), structural (d4:3 and d3:2) and biochemical properties (SSC, DMC, TA, malic acid, fructose, glucose, sucrose, AIS) were compared according to the four spectroscopic techniques (NIR, MIR, Raman and HSI) (**Tables 4** and **5**). For regression models (PLS-R, SVM-R and RF-R), the prediction performances were assessed by the determination coefficient of cross-validation (R_{cv}^2), the root mean square error of cross-validation ($RMSE_{cv}$) and the residual predictive deviation (RPD). Particularly, the RPD values from 2 to 2.5 indicate the possibility for approximate qualitative predictions, whereas from 2.5 to 3 or above correspond to good and excellent prediction accuracy (Nicolai, et al., 2007). The optimal numbers of latent variables (LVs) were obtained from developed PLS-DA and PLS-R models. Besides, the main attributed vibrational bands were selected based on the beta-coefficients of PLS models (Lan, Bureau, et al., 2021), and the variable importance (VIP) of SVM and RF models using the ‘varImp’ function by ‘caret’ package in R software (Kuhn, 2015). Particularly, the VIP method here was based on the mean square error of developed models using all the spectral variables (MSE_0) and the mean square error of new models (MSE_n) by permuting each spectral variable. Afterwards, the VIP score for each spectral variable was calculated by the increase of mean square error (IncMSE), following Eq (1):

$$IncMSE = \left(\frac{MSE_n - MSE_0}{MSE_0}\right) * 100\% \qquad (1)$$

A larger I*ncMSE* indicates a greater importance for a spectral variable. The main correlated spectral signals of the best developed models are shown in **Tables 3, 4** and **5**.

3 Results and discussion

3.1 Characteristics of apple purees

After puree processing, the different purees provided a large variability of chemical, textural and rheological properties (**Table S1**). In the PCA, the first principal component (PC1) and the second principal component (PC2) explained respectively 48.6% and 19.5% of the total variance. This PCA allowed to mainly represent the strong differences due to apple cultivars taking into account all the characterized parameters of the total 72 different purees after processing (**Fig. 2**).

‘Granny Smith’ (GS) purees (C) were clearly discriminated from the other puree groups along the PC1. The GS purees presented a significantly (p < 0.001) higher viscosity (*K* and *n*) and elasticity (yield stress, G’ and G”), particle size d(4:3) and volume d (3:2), TA, malic acid and AIS content than the others (**Fig. 2b** and **Table S1**). Remarkable higher values (p < 0.001) for SSC and DMC allowed the separation of

‘Golden Delicious’ (A and B) and ‘Royal Gala’ purees (D) along the second principal component (**Fig. 2a** and **2b**). Thinning practice (Th+) on GD apples (B) resulted in a less viscous purees than non-thinned GD purees (A) (**Table S1**), which is in line with our previous research (Lan, Renard, et al., 2020). For all non-refined (NR) purees, ‘Royal Gala’ had the lowest viscoelastic moduli (G’ < 934.0 ± 35.4 Pa, G”< 194.3 ± 7.2 Pa), titratable acidity (TA < 3.8 ± 0.2 meq/kg) and cell wall contents (AIS < 128.4 ± 9.5 mg/g). However, the overlapping of the two kinds of ‘Braeburn’ purees (E and F) (**Fig. 2a**) revealed the difficulty to produce different purees after processing and refining of either, crunchy (puncture linear distance of 14.0 ± 1.2 Newton) and mealy (puncture linear distance of 11.7 ± 0.7 Newton) apples.

The two different processing recipes used here (Processes I and II) led to significant ($p < 0.01$) changes of puree rheological behaviors (*K*, *n*, G’, G”, yield stress and tan δ) and particle distributions (d4:3 and d3:2), but not of chemical attributes (SSC, DMC and AIS; $p > 0.05$) (**Table S1**). Particularly, purees processed at 95 °C and 400 rpm (Process II) had a soft solid-like behavior. They were more viscous (*K* and *n*) with higher G’ and G” and larger particles (d4:3 and d3:2) than the purees processed at 70 °C and 3000 rpm (Process I).

Moreover, as expected, the refining treatment generated a significant ($p < 0.01$) decrease of puree viscosity and elasticity (*K*, *n*, G’, G” and yield point), particle sizes (d4:3 and d3:2) and cell wall contents, but did not impact ($p > 0.05$) chemical attributes.

3.2 Discrimination of variability of apple purees

Generally, PLS-DA models developed using NIR, MIR, Raman and HSI spectra of purees had the best performances to discriminate the cultivars (a), processes (b), fruit thinning (d) and stress treatments (e) (**Table 2**). However, specifically for refined purees, MIR technique coupled with machine learning (RF-DA and SVM-DA) gave a higher discrimination accuracy (acc > 90.3%) of purees than PLS-DA (acc = 84.7%) (**Table 2**).

NIR technique coupled with PLS-DA models gave a correct discrimination of the four cultivars (acc = 88.8%, 4 LVs), the two GD fruit thinning purees (acc = 86.7%, 2 LVs) and the two Braeburn storage impacts (acc = 95.8%, 3 LVs). The specific NIR spectral regions at 818-850, 1849, 1880 and 2145-2155 nm mainly contributed to cultivar discrimination (**Table 3**). Particularly, the spectral area at 800-1000 nm, which is known as the absorption of apple carbohydrates and water variations (Giovanelli, Sinelli, Beghi, Guidetti, & Casiraghi, 2014; Zude, Herold, Roger, Bellon-Maurel, & Landahl, 2006), has been used for the apple cultivar classification (Bobelyn, et al., 2010). The absorption bands around 1880 nm are explained by the O-H combinations of water contents in apples (Camps, Guillermin, Mauget, & Bertrand, 2017). The broad band at 2100-2200 nm corresponds to the first combination band of C-H bonds of sugars and acids, and has already been highlighted in our previous work (Lan, Jaillais, et al., 2020). Besides, the wavelengths around 1400 nm (1345, 1392 and 1379-1384 nm), related to the soluble solids variations in apple juices (Kaur, Künnemeyer, & McGlone, 2020), were one of the major contributors for the discriminations of apple thinning (Th+ and Th-) and stress treatments (crunchy BR and mealy BM). However, NIR technique was not able to well classify (acc < 55.6%) the processing recipes and refining levels, which nevertheless induced intensive structural and rheological variations of purees (**Table S1** and **Table 3**).

MIR technique provided a better discrimination of all studied factors (**Table S-2**) than NIR. Particularly, three different discrimination models (PLS-DA, SVM-DA, RF-DA) (**Table 2**) allowed to classify the four puree cultivars with the acc values of 100%. The specific spectral wavenumbers at 1723-1718, 1107, 1061 and 1022 cm^{-1} (**Table 3**), attributed to the stretching bonds of C=O of malic acid, and the C-O and C-C of glucose, fructose and sucrose (Bureau, Cozzolino, & Clark, 2019), were consistent with the measured differences of purees coming from different cultivars (**Fig. 2 and Table S-2**). Compared to NIR results, the satisfactory classifications by MIR of processing recipe (acc = 100 %) and refining (acc = 91.7%) were mainly based on the overlapped region between 1750 and 1650 cm^{-1} (1749 cm^{-1}, 1730-1715 cm^{-1} and 1640-1628 cm^{-1} in **Table 3**), related to the organic acids, soluble polysaccharides, pectins, phenolics and absorbed water (Lan, Renard, et al., 2020). MIR was able to highlight the physicochemical modifications of apple purees generated by different processing strategies (heating temperature and grinding speed) and mechanical refining treatments. Besides the aforementioned spectral signals, the excellent PLS discriminations of apple thinning (acc = 100%) and stress treatments (acc = 100%) were linked to three specific wavenumbers at 1084, 1056 and 998 cm^{-1}, corresponding to the variations of glucose and sucrose in fruits (Bureau, et al., 2019).

For Raman spectroscopy, PLS-DA models developed over the range of 800-1800 cm^{-1} had a lower discrimination accuracy and more LVs to discriminate puree cultivars (acc = 81.3%, 7 LVs), thinning practices (acc = 75.0%, 6 LVs) and stress treatments (acc = 70.8%, 6 LVs) than the models obtained with NIR and MIR (**Table 3**). The main vibrational bands responsible for these discriminations were related to the variations of major sugars and acids in apple purees, which have been highlighted in honey products (Pompeu, et al., 2018) and soft drinks (Ilaslan, Boyaci, & Topcu, 2015). In particular, were observed the C-C stretching and C-H deformation vibrations of glucose at 840-842 cm^{-1} (Özbalci, Boyaci, Topcu, Kadılar, & Tamer, 2013); the stretching of C-O-C at 872 cm^{-1} and the deformation of C-OH of fructose at 872, 939, 944 and 1054 cm^{-1} (Cerchiaro, Sant'Ana, Temperini, & da Costa Ferreira, 2005; Mathlouthi & Luu, 1980; Özbalci, et al., 2013); the C-O and C-OH vibrations of sucrose at 1126 cm^{-1} (Ilaslan, et al., 2015; Pierna, Abbas, Dardenne, & Baeten, 2011) and the C=O stretching of malic acid at 1734 cm^{-1} (Barańska, Kuduk-Jaworska, Szostak, & Romaniewska, 2003). Interestingly, Raman spectra discriminated different puree processing conditions with the acc value of 82.3%. Besides the aforementioned wavenumbers, the specific Raman bands at 845 and 1433-1436 cm^{-1} were observed to discriminate puree processing changes. These wavelengths are known to represent the C-O-C and COO-antisymmetric stretching of pectins during the clarification of apple juice (Camerlingo, et al., 2007).

HSI technique coupled with PLS-DA showed a relatively higher discrimination accuracy of puree cultivars (acc = 100%, 7 LVs), processing recipes (acc = 86.1%, 10 LVs), fruit thinning practices (acc = 91.6%, 6 LVs) and stress treatments (acc = 100%, 4 LVs) than the conventional NIR spectroscopy, but using a higher number of latent variables. Besides the similar aforementioned wavenumber regions as in NIR around 1400, 1880 and 2100-2300 nm, specific spectral areas at 1048-1088 and 1106-1145 nm were observed, corresponding to the SSC and DMC variations in fruits (Lan, Jaillais, et al., 2021; Wang, Peng, Xie, Bao, & He, 2015). Comparing to NIR, PLS-DA on the averaged HSI puree spectra gave an impressive improvement of the discrimination of puree processing recipes, from 51.4% to 86.1%. However, both NIR and HSI spectra had a limited ability to discriminate the different refining levels (< 58.3% correct identification). These two techniques had the potential to detect puree variability (cultivar, fruit thinning, process) involving significant differences in composition (**Table S-1**), but not to estimate puree textural changes (refining) (**Table S-1**).

3.3 Prediction of apple puree quality traits

According to the RPD values described by Nicolai et al. (2007), NIR showed a poor prediction ($R_{cv}^2 < 0.52$, RPD < 1.4) of puree rheological (*K*, *n*, G', G'', yield stress and tan δ) and structural parameters (d4:3 and d3:2) (**Table 4**). However, it gave a good prediction of puree composition, such as DMC ($R_{cv}^2 = 0.82$, RPD = 2.3), SSC ($R_{cv}^2 = 0.83$, RPD = 2.5), TA ($R_{cv}^2 = 0.83$, RPD = 2.4) and pH ($R_{cv}^2 = 0.85$, RPD = 2.6). Particularly, the specific wavebands in the intervals 937-1050, 1180-1210 and 1290-1330 nm, corresponding to O-H and C-H vibrations of water and carbohydrates (Giovanelli, et al., 2014; Zude, et al., 2006), highly contributed to the DMC and SSC models,. Besides the aforementioned absorbance regions, NIR wavenumbers between 2208 and 2254 nm, corresponding to the combination bands of C-H and O-H (Wang, et al., 2015), were also considered in the puree DMC prediction. The wavelengths located around 1600 nm (1534-1607 nm for TA models) and 1850 nm (1835-1873 nm for TA and pH models) were used to estimate puree acidity, already described to correspond to the C-O vibration of COOH and O-H combinations (Camps, et al., 2017; Wang, et al., 2015). The prediction of puree individual compounds was acceptable only for malic acid ($R_{cv}^2 = 0.80$, RPD = 2.1). Generally, NIR spectra coupled with PLS gave a better estimation of puree quality than SVM and RF regression.

MIR technique was potentially able to estimate the rheological parameters (*K*, *n*, G', G'' and tan δ) with acceptable R_{cv}^2 (> 0.81) and RPD (> 2.0) values (**Table 4**). Particularly, PLS and RF models obtained acceptable predictions of the consistency (*K*) ($R_{cv}^2 > 0.81$, RPD > 2.1) and flow (n) ($R_{cv}^2 > 0.80$, RPD > 2.0) parameters of the power-law viscosity model of apple purees. PLS models gave the best predictions ($R_{cv}^2 > 0.82$, RPD > 2.3) of the viscoelastic parameters G' and G'' of purees but were less accurate for the yield stress ($R_{cv}^2 = 0.77$, RPD = 1.7). Impressively, MIRS coupled with PLS showed an excellent prediction of tan δ ($R_{cv}^2 = 0.96$, RPD = 5.1), corresponding to the integrative assessment of both elastic and viscous contributions of apple purees (Espinosa-Muñoz, et al., 2013). The spectral region at 1500-1750 cm^{-1} was highly relevant to estimate puree viscosity and viscoelasticity. It corresponds to the C=O and C-O stretching of carboxylic acids at 1745-1740 cm^{-1} and the C=O vibration of pectic methyl ester at 1628-1634 cm^{-1} (Liu, Renard, Rolland-Sabaté, Bureau, & Le Bourvellec, 2020). Concerning the puree structural properties, RF model was the best to predict particle sizes over volume d(4:3) ($R_{cv}^2 = 0.88$, RPD = 2.9) and over surface area d(3:2)

(R_{cv}^2 = 0.82, RPD = 2.2). For composition, acceptable to good PLS predictions were obtained for SSC, DMC, TA, pH, malic acid and sucrose, giving RPD from 2.2 to 3.9 (**Table 5**). The specific spectral signals related to the acids at 1736-1718 cm^{-1} and to the fructose and sucrose at 1065-1055 cm^{-1} and 1024-1016 cm^{-1} (Bureau, et al., 2019), were the major contributors of SSC and DMC models. The excellent predictions of TA and pH, with RPD values of 3.6 and 3.9, respectively, depended on the particularly strong absorptions bands between 1736-1715 cm^{-1}. However, a lower RPD (RPD = 2.2) and a higher LVs were obtained for malic acid than for TA. For individual sugars, an acceptable PLS prediction was obtained for fructose (R_{cv}^2 = 0.85, RPD = 2.6) based on its typical fingerprints at 1155, 1056 and 980 cm^{-1} (Bureau, et al., 2019; Lan, Renard, et al., 2020), but neither for sucrose (R_{cv}^2 < 0.78, RPD <1.9) nor for glucose (R_{cv}^2 < 0.49, RPD <1.4).

Raman spectroscopy showed a limited ability to estimate the rheological and structural properties of apple purees with low R_{cv}^2 (< 0.48) and RPD (< 1.4) values (**Table 4**). These results were in line with the lower ability of the aforementioned Raman model to distinguish between non-refined and refined purees (acc = 56.9%) (**Part 3.2**). Moreover, none of the developed Raman models gave acceptable predictions of the global (SSC, DMC, TA and pH) and individual biochemical compositions (sugars, acids and cell wall contents) of apple purees. The best Raman model had a R_{cv}^2 of 0.71 and a RPD value of 1.8, indicating a possible application only to distinguish puree samples presenting a large variation of titratable acidity (TA).

The models based on HSI data could not predict rheological (*K*, *n* , G', G'', yield stress, tan δ) ($R_{cv}^2 < 0.48$, RPD < 1.4) and structural (d4:3 and d3:2) ($R_{cv}^2 < 0.47$, RPD < 1.4) properties. Acceptable PLS predictions were obtained for SSC ($R_{cv}^2 = 0.86$, RPD = 2.7), DMC ($R_{cv}^2 = 0.84$, RPD = 2.4), TA ($R_{cv}^2 = 0.83$, RPD = 2.4) and pH ($R_{cv}^2 = 0.85$, RPD = 2.6). Particularly, the most contributing wavelengths, located at around 1180-1219, 1282-1327 and 2179-2207 nm, were the same as described with the NIR spectroscopy (**Table 5**). However, none of the models could predict individual sugars (fructose, glucose and sucrose) ($R_{cv}^2 < 0.74$, RPD < 1.8) and AIS contents ($R_{cv}^2 < 0.42$, RPD < 1.3).

3.4 Comparison of NIR, MIR, Raman and HSI performances

NIR spectroscopy, the easiest to apply and cheapest spectroscopic techniques in this work, showed an acceptable ability (2.3 < RPD < 2.6) to predict puree major chemical composition, including SSC, DMC, TA and pH. Such good NIR predictions will probably contribute to the development of the rapid routine evaluation of the composition of fruit-based products. For individual components, a good estimation was only obtained for malic acid, depending on its positive correlation with TA ($R^2 = 0.78$) and pH ($R^2 = 0.76$). However, NIR could not provide acceptable estimations of puree textural changes, in line with our previous conclusions (Lan, Jaillais, et al., 2020).

Compared to NIR, MIR technique had the potential to assess puree rheological properties, including both, viscosity and viscoelasticity. However, the predictions shown in this paper were less accurate (RPD > 2.0) than our previous ones (RPD > 2.4), which concerned purees presenting a larger range of rheological behaviors (Lan, Jaillais, et al., 2020). Interestingly, among puree viscoelastic parameters, tan δ was the best estimated by MIR (RPD = 5.1). Compared to machine learning models (SVM and RF), PLS regressions generally showed a better ability to predict puree rheological and biochemical properties. However, for the puree particle structure (size and volume), RF regression provided the best predictions. The informative wavenumber regions at 1500-1750 and 900-1200 cm^{-1} should be considered for rheological and structural assessments of apple purees, which was in line with previous works (Ayvaz, et al., 2016; Lan, Renard, et al., 2020). MIR coupled with PLS regression provided the best prediction of the global quality traits of purees (SSC, DMC, TA and pH) with the possibility to evaluate some individual components (malic acid and sucrose). The lower prediction of malic acid than of TA was probably due to its relatively low concentration (3.0 - 7.5 g/kg of malic acid and 3.5 – 11.1 g/kg of TA) and limited variations (SD = 1.0 g/kg of malic acid, SD = 2.2 g/kg of TA). For individual sugars, the higher internal correlations between fructose and SSC (R^2 = 0. 78) than between sucrose and SSC (R^2 = 0. 51) probably explained the better prediction of fructose than of sucrose.

In this study, Raman spectroscopy showed a potential to discriminate different purees, according to cultivar and processing recipe (acc > 81.3%), but it was not able to predict puree rheological, structural and chemical parameters. However, Raman gives excellent biochemical predictions on homogeneous samples, such as commercial tomato purees (Baranska, et al., 2006) and honey products (Özbalci, et al., 2013; Pierna, et al., 2011). It has also been used to detect the rheological changes of monotonous mixed food matrices (Nawrocka, Miś, & Szymańska-Chargot, 2016; Ngarize, Adams,

& Howell, 2004). In this work, the unsatisfactory predictions using Raman spectroscopy could be due to i) the very weak spectral signals corresponding to the biochemical variations in apple purees (even after water evaporation before spectrum acquisition) and ii) the variable heterogeneity according to the puree refining and grinding, which make a barrier against an efficient light diffusion.

The models based on the averaged NIR-HSI spectra of apple purees provided a significant improvement of puree discrimination (**Table 3**) and a slight increase in quality prediction (**Table 4 and Table 5**) in comparison with the results issued from a measurement of a limited sample area (~2 cm^2) by NIR spectroscopy. The averaged NIR-HSI spectra, which contained a richer spectral information of puree heterogeneity than the local NIR spectra, might explain the better model performance and relative higher numbers of LVs (Cheng & Sun, 2017). However, both NIR spectroscopy and HSI technique had a limited ability to detect puree differences after refining and to predict their rheological and structural properties. Strangely, the PLS-DA models using the full number of HSI spectra of each puree had a relatively lower discriminating accuracy than their corresponding averaged spectra. Previous works noticed the heterogeinity of tested samples usually affected the NIR and HSI determination precisions (Prieto, Roehe, Lavín, Batten, & Andrés, 2009). The large heterogeneity, including irregular particle size and shape and the high water content on puree surface, could introduce a strong diffuse reflection and spectral noise during the HSI image acquisition. Although NIR-HSI on purees slightly improved the prediction of SSC and DMC over the NIR results, the much larger volume of dataset and the longer time needed for image pre-processing limited its use in comparison with NIR local measurements.

Further, the AIS, which contributes to the rheological properties of processed puree products, was not well evaluated in this study whichever the spectroscopic technique or chemometric method used directly on puree samples.

4. Conclusion

This study provided a first comprehensive assessment to choose the best technique among NIR, MIR, Raman spectroscopies and HSI for evaluating apple puree variability and quality. MIR had the best performance to provide an accurate identification of puree properties due to apple variability (cultivar, fruit thinning and postharvest stress) and processing conditions (heating, grinding and refining). It gave also a reliable evaluation of puree rheological and structural characteristics and composition (RPD values from 2.1 to 5.1). NIR and HSI techniques can be more easily adapted to routine characterization of the more global parameters in purees (soluble solids, titratable acidity and dry matter), but not of their textural changes. Raman spectroscopy offered an insufficient information to evaluate apple puree variability and quality. Clearly, Raman spectroscopy should not be prioritized in further studies on the characterization of apple purees.

The current study also enables considering future applications with NIR, NIR-HSI and MIR according to the industrial or research needs (speed of data acquisition and presentation of the sample). These techniques are very suitable for the development of

Process Analytical Technology in order to trace samples and optimize conditions during processing.

Acknowledgements

The authors thank Patrice Reling, Barbara Gouble, Marielle Boge, Caroline Garcia Line Touloumet and Gisèle Riqueau (INRAE, SQPOV unit) for their technical help. The 'Interfaces' project is an Agropolis Fondation project publicly funded through the ANR (French Research Agency) under "Investissements d'Avenir" programme (ANR-10-LABX-01-001 Labex Agro, coordinated by Agropolis Fondation). Studies conducted with the phytotron were supported by the various CPER Platform 3A funders: European Union, European Regional Development Fund, the French Government, the Sud Provence-Alpes-Côte d'Azur Region, the Departmental Council of Vaucluse and the Urban Community of Greater Avignon. Weijie Lan was supported by a doctoral grant from Chinese Scholarship Council.

References

Ayvaz, H., Sierra-Cadavid, A., Aykas, D. P., Mulqueeney, B., Sullivan, S., & Rodriguez-Saona, L. E. (2016). Monitoring multicomponent quality traits in tomato juice using portable mid-infrared (MIR) spectroscopy and multivariate analysis. *Food Control, 66*, 79-86. https://doi.org/10.1016/j.foodcont.2016.01.031

Baeten, V., & Dardenne, P. (2005). Applications of near-infrared imaging for monitoring agricultural food and feed products. In *Spectrochemical Analysis Using Infrared Multichannel Detectors* (pp. 283-301). https://doi.org/10.1002/9780470988541.ch13

Barańska, H., Kuduk-Jaworska, J., Szostak, R., & Romaniewska, A. (2003). Vibrational spectra of racemic and enantiomeric malic acids. *Journal of Raman Spectroscopy, 34*(1), 68-76. https://doi.org/10.1002/jrs.953

Baranska, M., Schütze, W., & Schulz, H. (2006). Determination of lycopene and β-carotene content in tomato fruits and related products: comparison of FT-Raman, ATR-IR, and NIR spectroscopy. *Analytical Chemistry, 78*(24), 8456-8461. https://doi.org/10.1021/ac061220j

Bobelyn, E., Serban, A.-S., Nicu, M., Lammertyn, J., Nicolai, B. M., & Saeys, W. (2010). Postharvest quality of apple predicted by NIR-spectroscopy: Study of the effect of biological variability on spectra and model performance. *Postharvest Biology and Technology, 55*(3), 133-143. http://dx.doi.org/10.1016/j.postharvbio.2009.09.006

Buergy, A., Rolland-Sabaté, A., Leca, A., & Renard, C. M. G. C. (2020). Pectin modifications in raw fruits alter texture of plant cell dispersions. *Food Hydrocolloids, 107*, 105962. https://doi.org/10.1016/j.foodhyd.2020.105962

Bureau, S., Cozzolino, D., & Clark, C. J. (2019). Contributions of Fourier-transform mid infrared (FT-

MIR) spectroscopy to the study of fruit and vegetables: A review. *Postharvest Biology and Technology, 148*, 1-14. https://doi.org/10.1016/j.postharvbio.2018.10.003

Bureau, S., Quilot-Turion, B., Signoret, V., Renaud, C., Maucourt, M., Bancel, D., & Renard, C. M. G. C. (2013). Determination of the composition in sugars and organic acids in peach using mid infrared spectroscopy: comparison of prediction results according to data sets and different reference methods. *Analytical Chemistry, 85*(23), 11312-11318. https://doi.org/10.1021/ac402428s

Camerlingo, C., Zenone, F., Delfino, I., Diano, N., Mita, D. G., & Lepore, M. (2007). Investigation on clarified fruit juice composition by using visible light micro-Raman spectroscopy. *Sensors, 7*(10), 2049-2061. https://doi.org/10.3390/s7102049

Camps, C., Guillermin, P., Mauget, J. C., & Bertrand, D. (2017). Discrimination of storage duration of apples stored in a cooled room and shelf-life by visible-near infrared spectroscopy. *Journal of Near Infrared Spectroscopy, 15*(3), 169-177. https://doi.org/10.1255/jnirs.726

Cerchiaro, G., Sant'Ana, A. C., Temperini, M. L. A., & da Costa Ferreira, A. M. (2005). Investigations of different carbohydrate anomers in copper(II) complexes with d-glucose, d-fructose, and d-galactose by Raman and EPR spectroscopy. *Carbohydrate Research, 340*(15), 2352-2359. https://doi.org/10.1016/j.carres.2005.08.002

Cheng, J., & Sun, D. (2017). Partial Least Squares Regression (PLSR) Applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food Engineering Reviews, 9*(1), 36-49. https://doi.org/10.1007/s12393-016-9147-1

Cordella, C. B. Y., & Bertrand, D. (2014). SAISIR: A new general chemometric toolbox. *TRAC Trends in Analytical Chemistry, 54*, 75-82. https://doi.org/10.1016/j.trac.2013.10.009

Cullen, P. J., O'Donnell, C. P., & Fagan, C. C. (2014). Benefits and challenges of adopting PAT for the food industry. In *Process analytical technology for the food industry* (pp. 1-5): Springer. https://doi.org/10.1007/978-1-4939-0311-5_1

Defernez, M., Kemsley, E. K., & Wilson, R. H. (1995). Use of infrared spectroscopy and chemometrics for the authentication of fruit purees. *Journal of Agricultural and Food Chemistry, 43*(1), 109-113. https://doi.org/10.1021/jf00049a021

Espinosa-Muñoz, L., Renard, C. M. G. C., Symoneaux, R., Biau, N., & Cuvelier, G. (2013). Structural parameters that determine the rheological properties of apple puree. *Journal of Food Engineering, 119*(3), 619-626. https://doi.org/10.1016/j.jfoodeng.2013.06.014

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185, 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

Giovanelli, G., Sinelli, N., Beghi, R., Guidetti, R., & Casiraghi, E. (2014). NIR spectroscopy for the optimization of postharvest apple management. *Postharvest Biology and Technology, 87*, 13-20. https://doi.org/10.1016/j.postharvbio.2013.07.041

Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Ilaslan, K., Boyaci, I. H., & Topcu, A. (2015). Rapid analysis of glucose, fructose and sucrose contents of commercial soft drinks using Raman spectroscopy. *Food Control, 48*, 56-61. https://doi.org/10.1016/j.foodcont.2014.01.001

Kaur, H., Künnemeyer, R., & McGlone, A. (2020). Investigating aquaphotomics for temperature-independent prediction of soluble solids content of pure apple juice. *Journal of Near Infrared*

Spectroscopy, 28(2), 103-112. https://doi.org/10.1364/JNIRS.28.000103

Keenan, D. F., Brunton, N., Butler, F., Wouters, R., & Gormley, R. (2011). Evaluation of thermal and high hydrostatic pressure processed apple purees enriched with prebiotic inclusions. *Innovative Food Science & Emerging Technologies, 12*(3), 261-268. https://doi.org/10.1016/j.ifset.2011.04.003

Kuhn, M. (2015). Caret: classification and regression training. R package version 6.0-85. https://CRAN.R-project.org/package=caret

Lan, W., Bureau, S., Chen, S., Leca, A., Renard, C. M. G. C., & Jaillais, B. (2021). Visible, near- and mid-infrared spectroscopy coupled with an innovative chemometric strategy to control apple puree quality. *Food Control, 120*, 107546. https://doi.org/10.1016/j.foodcont.2020.107546

Lan, W., Jaillais, B., Leca, A., Renard, C. M. G. C., & Bureau, S. (2020). A new application of NIR spectroscopy to describe and predict purees quality from the non-destructive apple measurements. *Food Chemistry, 310*, 125944. https://doi.org/10.1016/j.foodchem.2019.125944

Lan, W., Jaillais, B., Renard, C. M. G. C., Leca, A., Chen, S., Le Bourvellec, C., & Bureau, S. (2021). A method using near infrared hyperspectral imaging to highlight the internal quality of apple fruit slices. *Postharvest Biology and Technology, 175*, 111497. https://doi.org/10.1016/j.postharvbio.2021.111497

Lan, W., Renard, C. M. G. C., Jaillais, B., Leca, A., & Bureau, S. (2020). Fresh, freeze-dried or cell wall samples: Which is the most appropriate to determine chemical, structural and rheological variations during apple processing using ATR-FTIR spectroscopy? *Food Chemistry, 330*, 127357. https://doi.org/10.1016/j.foodchem.2020.127357

Liu, X., Renard, C. M. G. C., Rolland-Sabaté, A., Bureau, S., & Le Bourvellec, C. (2020). Modification of apple, beet and kiwifruit cell walls by boiling in acid conditions: Common and specific responses. *Food Hydrocolloids*, 106266. https://doi.org/10.1016/j.foodhyd.2020.106266

Liu, M., Wang, M., Wang, J., & Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical, 177*, 970-980. https://doi.org/10.1016/j.snb.2012.11.071

Link, H. (2000). Significance of flower and fruit thinning on fruit quality. *Plant growth regulation, 31*(1), 17-26. https://doi.org/10.1023/A:1006334110068

Ma, T., Li, X., Inagaki, T., Yang, H., & Tsuchikawa, S. (2018). Noncontact evaluation of soluble solids content in apples by near-infrared hyperspectral imaging. *Journal of Food Engineering, 224,* 53-61. https://doi.org/10.1016/j.jfoodeng.2017.12.028

Mathlouthi, M., & Luu, D. V. (1980). Laser-raman spectra of d-fructose in aqueous solution. *Carbohydrate Research, 78*(2), 225-233. https://doi.org/10.1016/0008-6215(80)90002-6

Mendoza, F., Lu, R., Ariana, D., Cen, H., & Bailey, B. (2011). Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content. *Postharvest Biology and Technology, 62*(2), 149-160. https://doi.org/10.1016/j.postharvbio.2011.05.009

Nawrocka, A., Miś, A., & Szymańska-Chargot, M. (2016). Characteristics of relationships between structure of gluten proteins and dough rheology – influence of dietary fibres studied by FT-Raman Spectroscopy. *Food Biophysics, 11*(1), 81-90. https://doi.org/10.1007/s11483-015-9419-y

Ngarize, S., Adams, A., & Howell, N. K. (2004). Studies on egg albumen and whey protein interactions by FT-Raman spectroscopy and rheology. *Food Hydrocolloids, 18*(1), 49-59. https://doi.org/10.1016/S0268-005X(03)00041-9

Nicolai, B. M., Beullens, K., Bobelyn, E., Peirs, A., Saeys, W., Theron, K. I., & Lammertyn, J. (2007). Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review. *Postharvest Biology and Technology, 46*(2), 99-118. https://doi.org/10.1016/j.postharvbio.2007.06.024

Noble, W. S. (2006). What is a support vector machine?. *Nature Biotechnology*, 24(12), 1565-1567. https://doi.org/10.1038/nbt1206-1565

Osborne, B. G. (2006). Near-infrared spectroscopy in food analysis. *Encyclopedia of analytical chemistry: applications, theory and instrumentation*. : https://doi.org/10.1002/9780470027318.a1

Özbalci, B., Boyaci, İ. H., Topcu, A., Kadılar, C., & Tamer, U. (2013). Rapid analysis of sugars in honey by processing Raman spectrum using chemometric methods and artificial neural networks. *Food Chemistry, 136*(3), 1444-1452. https://doi.org/10.1016/j.foodchem.2012.09.064

Pierna, J. A. F., Abbas, O., Dardenne, P., & Baeten, V. (2011). Discrimination of Corsican honey by FT-Raman spectroscopy and chemometrics. *BASE*. https://popups.uliege.be/1780-4507/index.php?id=6895.

Pissard, A., Fernández Pierna, J. A., Baeten, V., Sinnaeve, G., Lognay, G., Mouteau, A., Dupont, P., Rondia, A., & Lateur, M. (2013). Non-destructive measurement of vitamin C, total polyphenol and sugar content in apples using near-infrared spectroscopy. *Journal of the Science of Food and Agriculture, 93*(2), 238-244. https://doi.org/10.1002/jsfa.5779

Pistorius, A. M. A. (1996). *Biochemical applications of FT-IR spectroscopy*: [Sl: sn].

http://hdl.handle.net/2066/18822

Pompeu, D. R., Larondelle, Y., Rogez, H., Abbas, O., Pierna, J. A. F., & Baeten, V. (2018). Characterization and discrimination of phenolic compounds using Fourier transform Raman spectroscopy and chemometric tools. *BASE*. https://popups.uliege.be/1780-4507/index.php?id=16270

Prieto, N., Roehe, R., Lavín, P., Batten, G., & Andrés, S. (2009). Application of near infrared reflectance spectroscopy to predict meat and meat products quality: A review. *Meat Science, 83*(2), 175-186. https://doi.org/10.1016/j.meatsci.2009.04.016

R Core Team, R. C. (2019). R: A language and environment for statistical computing.

Renard, C. M. G. C. (2005). Variability in cell wall preparations: quantification and comparison of common methods. *Carbohydrate Polymers, 60*(4), 515-522. https://doi.org/10.1016/j.carbpol.2005.03.002

Tu, K., Nicolaï, B., & De Baerdemaeker, J. (2000). Effects of relative humidity on apple quality under simulated shelf temperature storage. *Scientia Horticulturae, 85*(3), 217-229. https://doi.org/10.1016/S0304-4238(99)00148-X

Wang, H. L., Peng, J. Y., Xie, C. Q., Bao, Y. D., & He, Y. (2015). Fruit quality evaluation using spectroscopy technology: a review. *Sensors, 15*(5), 11889-11927. https://doi.org/10.3390/s150511889

Zude, M., Herold, B., Roger, J. M., Bellon-Maurel, V., & Landahl, S. (2006). Non-destructive tests on the prediction of apple fruit flesh firmness and soluble solids content on tree and in shelf life. *Journal of Food Engineering, 77*(2), 254-260. https://doi.org/10.1016/j.jfoodeng.2005.06.027

Figure captions

Fig. 1. Experimental scheme of apple puree processing, quality characterization and spectral acquisition**.**

Fig. 2. Principal component analysis on chemical, structural and rheological parameters of six puree groups (A: GD Th-; B: GD Th+; C: GS; D: GA; E: BR; F:BM): **(a)** the scores plot of the two first components (PC1 and PC2); **(b)** the correlation plot of the PC1 and PC2.

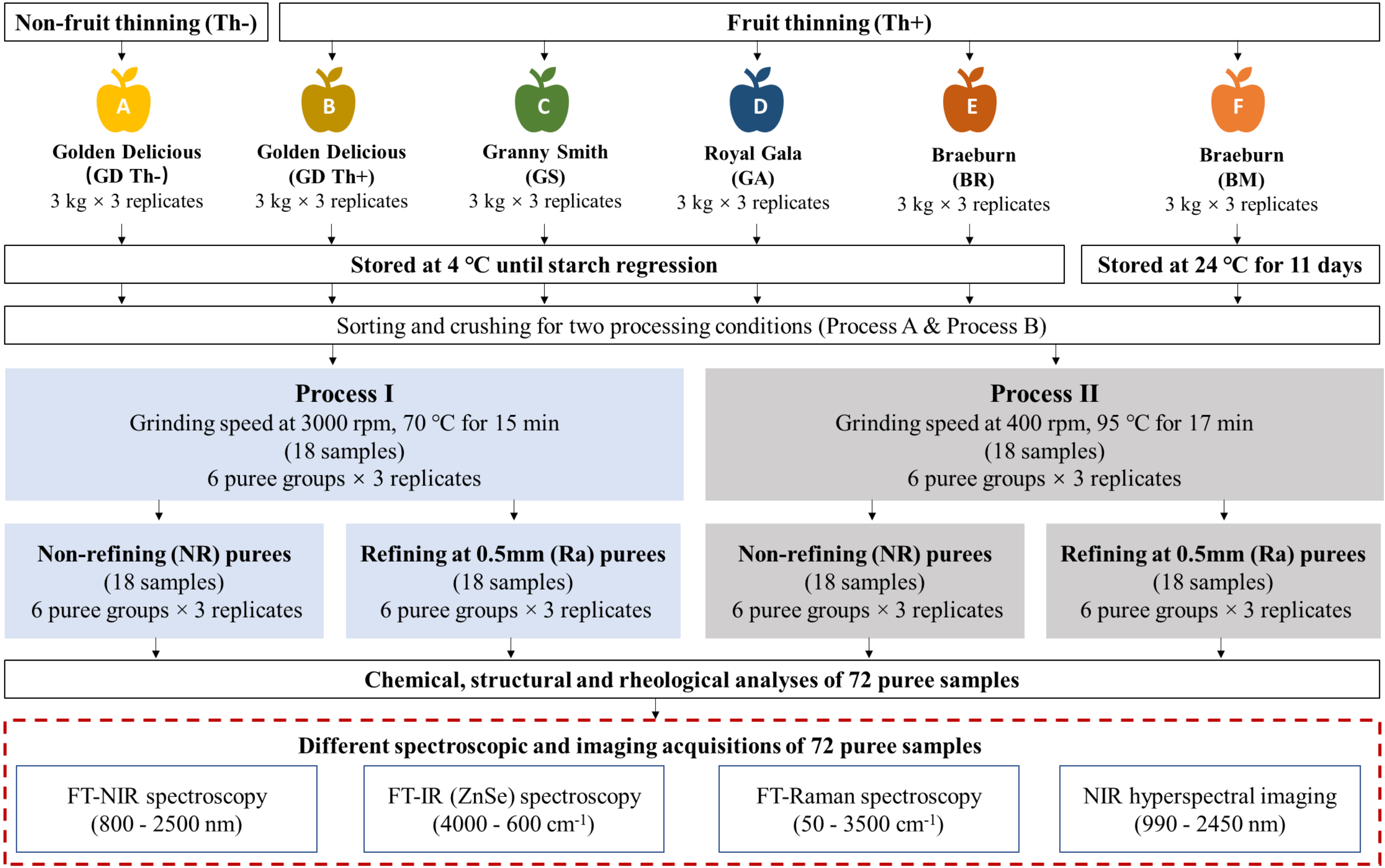
Non-fruit thinning (Th-)
Fruit thinning (Th+)
A
B
C
D
E
F
Golden Delicious (GD Th-)
3 kg × 3 replicates
Golden Delicious (GD Th+)
3 kg × 3 replicates
Granny Smith (GS)
3 kg × 3 replicates
Royal Gala (GA)
3 kg × 3 replicates
Braeburn (BR)
3 kg × 3 replicates
Braeburn (BM)
3 kg × 3 replicates
Stored at 4 °C until starch regression
Stored at 24 °C for 11 days
Sorting and crushing for two processing conditions (Process A & Process B)
Process I
Grinding speed at 3000 rpm, 70 °C for 15 min
(18 samples)
6 puree groups × 3 replicates
Process II
Grinding speed at 400 rpm, 95 °C for 17 min
(18 samples)
6 puree groups × 3 replicates
Non-refining (NR) purees
(18 samples)
6 puree groups × 3 replicates
Refining at 0.5mm (Ra) purees
(18 samples)
6 puree groups × 3 replicates
Non-refining (NR) purees
(18 samples)
6 puree groups × 3 replicates
Refining at 0.5mm (Ra) purees
(18 samples)
6 puree groups × 3 replicates
Chemical, structural and rheological analyses of 72 puree samples
Different spectroscopic and imaging acquisitions of 72 puree samples
FT-NIR spectroscopy (800 - 2500 nm)
FT-IR (ZnSe) spectroscopy (4000 - 600 cm-1)
FT-Raman spectroscopy (50 - 3500 cm-1)
NIR hyperspectral imaging (990 - 2450 nm)

Fig. 1

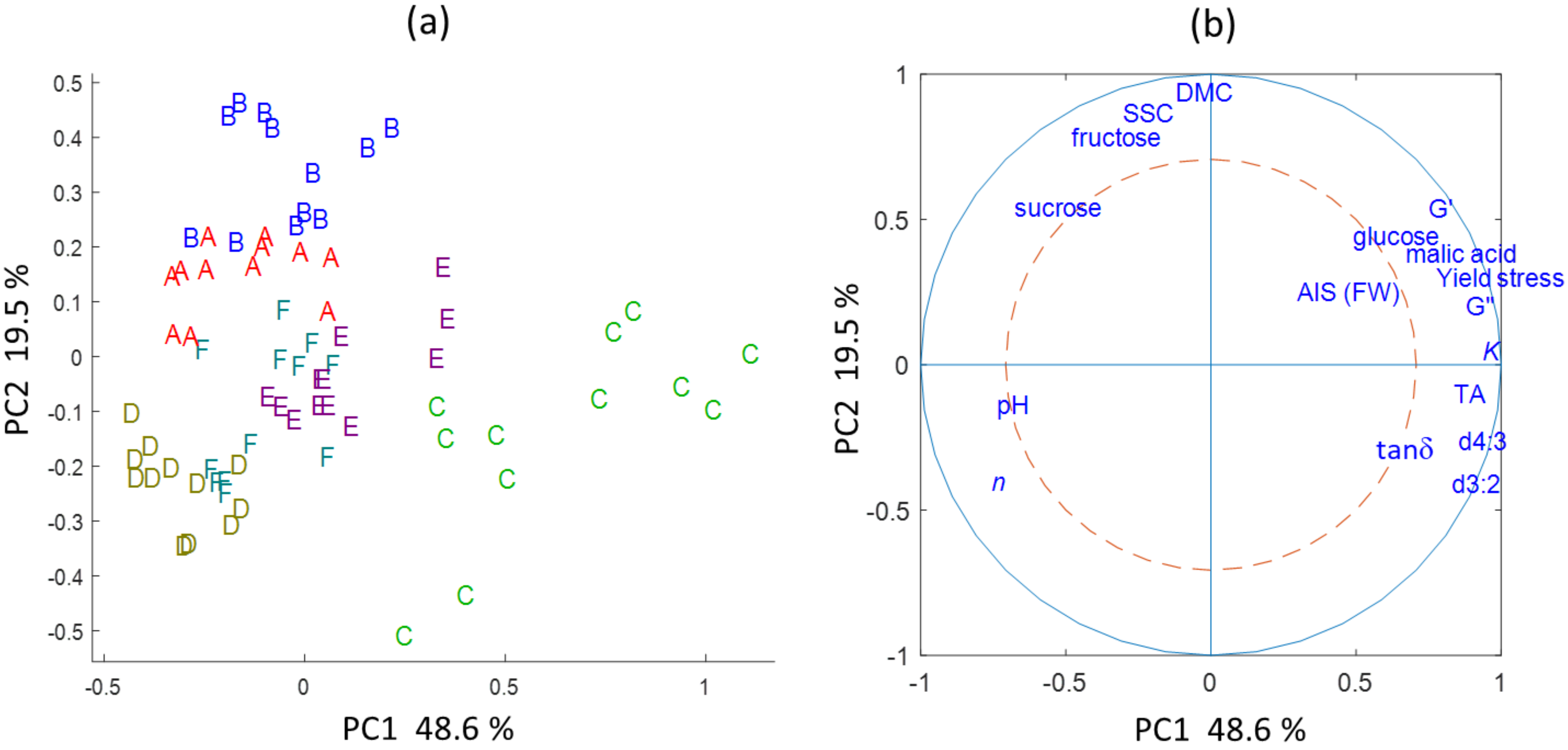
(a)
PC2 19.5 %
PC1 48.6 %
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
0
0.5
1
A
B
C
D
E
F
(b)
PC2 19.5 %
PC1 48.6 %
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
DMC
SSC
fructose
sucrose
G'
glucose
malic acid
Yield stress
AIS (FW)
G"
K
TA
pH
tanδ
d4:3
d3:2
n

Fig. 2

Table 1. The common names and their abbreviations used in this study

Common names	Abbreviations
process analytical techniques	PATs
near infrared spectroscopy	NIR
mid infrared spectroscopy	MIR
Raman spectroscopy	Raman
hyperspectral imaging	HSI
‘Golden Delicious’	GD
‘Granny Smith’	GS
‘Royal Gala’	GA
Crunchy ‘Braeburn’ stored at 4 °C	BR
Mealy ‘Braeburn’ stored at 23 °C	BM
fruit thinned / non-thinned apples	Th+ / Th-
non-refined / refined	NR / Ra
partial least square	PLS
random forest	RF
support vector machine	SVM
the storage modulus of purees	G’
the loss modulus of purees	G”
G''/G' of purees	tan δ
puree particle sizes averaged over volume	d4:3
puree particle sizes averaged over surface areas	d3:2
dry matter content	DMC
soluble solid content	SSC
titratable acidity	TA
alcohol insoluble solids	AIS
standard deviation value	SD
principal component analysis	PCA
fresh weight	FW
standard normal variate	SNV
determination coefficient of cross validation	R_{cv}^2
root mean square error of cross validation	RMSEP
the number of latent variables	LVs
residual predictive deviation	RPD

Table 2. Discrimination using 10-fold full cross-validation PLS-DA, SVM-DA and RF-DA models of apple purees according to (a) cultivars, (b) processes, (c) refining levels, (d) fruit thinning practices of Golden Delicious apples, (e) stress treatments of Braeburn apples, using NIR, MIR, Raman and HSI data.

Spectral techniques	**NIR**			**MIR**			**Raman**			**HSI**		
Spectral ranges	800- 2500 nm			900- 1800 cm^{-1}			800- 1800 cm^{-1}			990-2450 nm		
Models	PLS-DA	SVM-DA	RF-DA	PLS-DA	SVM-DA	RF-DA	PLS-DA	SVM-DA	RF-DA	PLS-DA	SVM-DA	RF-DA
(a) Cultivar (GD/GS/BR/GA)												
No. of samples	48	48	48	48	48	48	48	48	48	48	48	48
Correct discrimination rate	88.8 %	81.25 %	84.6 %	100.0 %	100.0 %	100.0 %	81.3 %	50.0 %	60.4 %	100 %	72.9 %	72.9 %.
LVs	4	-	-	3	-	-	7	-	-	7		
(b) Process (I/ II)												
No. of samples	72	72	72	72	72	72	72	72	72	72	72	72
Correct discrimination rate	51.4 %	31.9 %	44.4 %	100 %	97.2 %	93.1 %	82.3 %	67.7 %	67.7 %	86.1 %	41.7 %	47.2 %
LVs	4	-	-	5	-	-	8	-	-	10		
(c) Refining levels (NR/ Ra)												
No. of samples	72	72	72	72	72	72	72	72	72	72	72	72
Correct discrimination rate	51.4 %	38.9 %	55.6 %	84.7 %	90.3 %	91.7 %	56.9 %	40.3 %	45.8 %	55.1 %	51.4 %	58.3 %
LVs	5	-	-	4	-	-	7	-	-	6		
(d) Fruit thinning (Th+/ Th-)												
No. of samples	24	24	24	24	24	24	24	24	24	24	24	24
Correct discrimination rate	86.7 %	53.3 %	82.5 %	100.0 %	100.0 %	100.0 %	75.0 %	16.7 %	45.8 %	91.6 %	79.2 %	87.5 %
LVs	3	-	-	3	-	-	6	-	-	6		
(e) stress treatments (BR/ BM)												
No. of samples	24	24	24	24	24	24	24	24	24	24	24	24
Correct discrimination rate	95.8 %	63.3 %	87.5 %	100.0 %	100.0 %	100.0 %	70.8 %	25.0 %	54.2 %	100.0 %	58.3 %	87.5 %
LVs	3	-	-	3	-	-	6	-	-	4		

Note: ‘Cultivar’: (four varieties of ‘Golden Delicious’, ‘Braeburn’, ‘Granny Smith’ and ‘Royal Gala’); ‘fruit thinning’: different fruit thinning practices for Golden Delicious apples (50 - 100 fruits/ tree or 150-200 fruits/ tree); ‘stress’: two different textures of Braeburn apples (11 days at 24 °C or 2 months at 4 °C); ‘processing’: two processing recipes (70 °C for 15 mins with 3000 rpm grinding or 95 °C for 17 mins with 400 rpm grinding); ‘refining’: two refining conditions after puree processing (refined at 0.5 mm or not refined).

Table 3. The main attributions for vibrational bands of the best overall discrimination models developed for puree samples.

Spectra	Spectral ranges	Factors	No. samples	Model	LVs	acc (%)	Key frequencies NIR (nm), MIR (cm^{-1}), Raman (cm^{-1}), HSI (nm)
NIR	800-2500 nm	cultivar	48	PLS-DA	5	88.8	818-850, 1849, 1880, 2145-2155
		process	72	PLS-DA	4	51.4	/
		refining	72	RF-DA	-	55.6	/
		fruit thinning	24	PLS-DA	2	86.7	904, 1392, 1864
		stress	24	PLS-DA	3	95.8	913, 1345, 1379-1384
MIR	1800- 900 cm^{-1}	cultivar	48	PLS-DA	4	100.0	1723-1718, 1107, 1061, 1022
		process	72	PLS-DA	5	100.0	1730-1715, 1640-1628, 1138, 1084, 1001-998
		refining	72	RF-DA	-	91.7	1749, 1636, 1061, 1018, 995
		fruit thinning	24	PLS-DA	3	100.0	1772, 1593, 1084, 1022, 998
		stress	24	PLS-DA	3	100.0	1658-1608, 1056, 1018, 1001
Raman	800-1800 cm^{-1}	cultivar	48	PLS-DA	7	81.3	842, 873, 1064, 1126, 1266, 1433, 1610
		process	72	PLS-DA	8	82.3	816-818, 845, 939, 972, 1362-1367, 1433-1436, 1734
		refining	72	PLS-DA	7	56.9	/
		fruit thinning	24	PLS-DA	6	75.0	842, 1054, 1077, 1427, 1608, 1675
		stress	24	PLS-DA	6	70.8	840, 904, 944, 1059-1063, 1334, 1734
HSI	990-2450 nm	cultivar	48	PLS-DA	7	100.0	1106-1145, 1259, 1338, 1406, 1869-1874, 1931-1964
		process	72	PLS-DA	10	86.1	1048-1088, 1191, 1242, 2117, 2274-2387, 2437
		refining	72	RF-DA	/	58.3	/
		fruit thinning	24	PLS-DA	6	91.6	1065-1088, 1338-1367, 2145, 2331-2342, 2376-2398, 2426
		stress	24	PLS-DA	4	100.0	1048, 1134, 1389, 1947, 2409

Note: acc: discrimination accuracy; PLS-DA: partial least square discrimination; RF-DA: random forest discrimination. ‘Cultivar’: four apple varieties of ‘Golden Delicious’, ‘Braeburn’, ‘Granny Smith’ and ‘Royal Gala’ ; ‘fruit thinning’: different fruit thinning practices for Golden Delicious apples (50 - 100 fruits/ tree or 150-200 fruits/ tree); ‘stress’: two stress treatments of Braeburn apples (11 days at 24 °C or 2 months at 4 °C); ‘processing’: two processing recipes (70 °C for 15 mins with 3000 rpm grinding or 95 °C for 17 mins with 400 rpm grinding); ‘refining’: two refining conditions after puree processing (refined at 0.5 mm or not refined).

Table 4. Prediction of rheological and structural properties of apple purees using the full cross-validation PLS, SVM and RF regression based on their NIR, MIR, Raman and HSI spectra.

Parameter	**Spectra**	**Ranges**	**SD**	**PLS-R**				**SVM-R**			**RF-R**			**Key frequencies**
				R_{cv}^2	**$RMSE_{CV}$**	**RPD**	**LVs**	**R_{cv}^2**	**$RMSE_{CV}$**	**RPD**	**R_{cv}^2**	**$RMSE_{CV}$**	**RPD**	**NIR (nm), MIR (cm^{-1}), Raman (cm^{-1}), HSI (nm)**
Viscosity- *K*	NIR	6.6 - 46.8	8.7	0.41	6.8	1.3	6	0.31	8.2	1.1	0.32	7.7	1.1	/
	MIR			0.81	4.1	2.1	7	0.71	5.5	1.6	0.81	4.1	2.1	1712, 1682 - 1668, 1539, 1152, 1094, 1061, 998
	Raman			0.37	6.95	1.3	6	0.31	7.4	1.3	0.31	7.2	1.2	/
	HSI			0.54	6.1	1.4	10	0.27	7.5	1.2	0.36	6.6	1.3	/
Viscosity- *n*	NIR	0.19 - 0.34	0.04	0.52	0.03	1.4	6	0.30	0.04	1.0	0.35	0.03	1.2	/
	MIR			0.81	0.02	2.2	8	0.80	0.02	2.1	0.80	0.02	2.1	1745 - 1740, 1712 - 1710, 1539, 1140, 1081, 1065-1059,1036, 980
	Raman			0.48	0.03	1.4	8	0.36	0.03	1.4	0.44	0.03	1.3	/
	HSI			0.42	0.03	1.3	7	0.25	0.03	1.1	0.33	0.03	1.2	/
G' (Pa)	NIR	617 - 1962	322	0.32	270	1.2	6	0.11	320	1.0	0.27	282	1.1	/
	MIR			0.82	140	2.3	8	0.80	156	2.1	0.83	139	2.3	1745-1740, 1707, 1634, 1558 - 1537, 1140, 1078, 1063, 1036, 980
	Raman			0.10	326	1.0	6	0.11	303	1.0	0.25	276	1.2	/
	HSI			0.38	263	1.2	9	0.21	298	1.1	0.26	273	1.2	/
G'' (Pa)	NIR	114 - 593	92	0.36	77	1.2	6	0.21	92	1.0	0.26	84	1.1	/
	MIR			0.84	36	2.5	6	0.77	62	1.5	0.81	42	2.2	1745-1740, 1709, 1634-1628, 1558 - 1537, 1139, 1065, 1034, 980
	Raman			0.12	100	0.9	6	0.22	82	0.9	0.20	82	1.1	/
	HSI			0.41	74	1.3	10	0.16	84	1.1	0.19	85	1.1	/
yield stress	NIR	6.4 - 27.7	5.2	0.36	4.2	1.2	6	0.21	5.2	1.0	0.34	4.5	1.2	/
	MIR			0.77	3.0	1.7	7	0.73	2.8	1.8	0.67	3.0	1.8	/
	Raman			0.33	4.3	1.2	8	0.26	4.5	1.2	0.27	4.4	1.2	/
	HSI			0.47	4.0	1.3	11	0.27	4.4	1.2	0.36	4.3	1.2	/
tan δ	NIR	0.18 - 0.30	0.03	0.22	0.03	1.1	5	0.16	0.03	1.0	0.15	0.03	1.0	/
	MIR			0.96	0.01	5.1	7	0.95	0.01	3.7	0.96	0.01	4.5	1749, 1537, 1109 - 1105, 1040 - 1038, 1018 - 1016, 980
	Raman			0.44	0.02	1.3	5	0.45	0.02	1.3	0.42	0.02	1.3	/
	HSI			0.24	0.03	1.1	6	0.14	0.03	1.1	0.15	0.03	1.1	/
d4:3	NIR	239 - 777	130	0.47	95	1.4	6	0.21	130	1.0	0.32	106	1.2	/

	MIR			0.85	50	2.6	8	0.81	60	2.2	0.88	45	2.9	1745, 1626 - 1620, 1539- 1510, 1151, 1099 - 1092, 1061, 1001, 922
	Raman			0.47	93	1.4	6	0.17	117	1.4	0.19	117	1.1	/
	HSI			0.59	85	1.5	9	0.22	114	1.1	0.30	107	1.2	/
d3:2	NIR	170 - 402	53	0.42	41	1.3	6	0.22	53	1.0	0.29	47	1.1	/
	MIR			0.66	31	1.7	8	0.70	30	1.8	0.81	24	2.2	1745, 1699, 1626-1620, 1151, 1099 - 1092, 1061, 1001, 975, 922
	Raman			0.43	41	1.3	6	0.14	49	1.3	0.14	50	1.1	/
	HSI			0.50	40	1.3	9	0.26	46	1.2	0.29	44	1.2	/

Notes: Puree spectra and reference data from four varieties (‘Golden Delicious’, ‘Braeburn’, ‘Granny Smith’ and ‘Royal Gala’) with different fruit thinning practices for Golden Delicious apples (50 - 100 fruits/ tree or 150-200 fruits/ tree), stress treatments for Braeburn apples (11 days at 24 °C or 2 months at 4 °C), two processing recipes (70 °C for 15 mins with 3000 rpm grinding or 95 °C for 17 mins with 400 rpm grinding) and two refining conditions (refined at 0.5 mm or not refined). All results corresponded to 10-fold full-crossed validation tests. R_{cv}^2: determination coefficient of the full-crossed validation test; $RMSE_{cv}$: root mean square error of full-cross validation test; RPD: the residual predictive deviation of full-crossed validation test, LVs: the optimal numbers of latent variables. PLS-R: partial least square regression; RF-R: random forest regression; SVM-R: support vector machine regression.

Table 5. Prediction of biochemical properties of apple purees using the full cross-validation PLS, SVM and RF regression based on their NIR, MIR, Raman and HSI spectra.

Parameter	Spectra	Ranges	SD	PLS-R				SVM-R			RF-R			Key frequencies
				R_{cv}^2	$RMSE_{CV}$	RPD	LVs	R_{cv}^2	$RMSE_{CV}$	RPD	R_{cv}^2	$RMSE_{CV}$	RPD	NIR (nm), MIR (cm^{-1}), Raman (cm^{-1}), HSI (nm)
DMC (g/g)	NIR	0.16 - 0.23	0.01	0.82	0.01	2.3	7	0.73	0.01	1.9	0.78	0.01	2.1	937, 946, 1139, 1180 - 1210, 1307 - 1330, 2208 - 2254
	MIR			0.85	0.01	2.7	5	0.76	0.01	1.8	0.78	0.01	1.9	1734 - 1718, 1655 - 1637, 1084, 1061, 1024 – 1016
	Raman			0.20	0.01	1.0	8	0.02	0.01	1.0	0.01	0.01	1.0	/
	HSI			0.84	0.01	2.4	7	0.70	0.01	1.6	0.79	0.01	2.1	1037-1065, 1145, 1180-1219,1305-1338, 2286, 2421
SSC (°Brix)	NIR	11.6 - 15.8	1.1	0.83	0.4	2.5	6	0.50	0.8	1.4	0.57	0.7	1.5	944 - 946, 992, 1180- 1210, 1239, 1290 - 1330
	MIR			0.88	0.4	2.9	3	0.78	0.5	2.2	0.82	0.4	2.4	1736 - 1718, 1065 – 1055, 1022 - 1016
	Raman			0.39	0.9	1.2	9	0.18	1.0	1.2	0.15	1.0	1.1	/
	HSI			0.86	0.4	2.7	8	0.66	0.7	1.5	0.76	0.6	1.9	1048-1071, 1140-1151, 1180-1219,1290-1338
TA (meq/kg)	NIR	3.5 - 11.1	2.2	0.83	0.9	2.4	7	0.43	1.7	1.3	0.72	1.2	1.8	1017, 1049, 1167, 1374, 1534 - 1607, 1835 - 1873
	MIR			0.92	0.6	3.6	5	0.92	0.6	3.6	0.91	0.6	3.4	1736 - 1718, 1605 - 1601, 1042 - 1030, 1001 - 995
	Raman			0.71	1.2	1.8	9	0.58	1.6	1.8	0.58	1.4	1.6	/
	HSI			0.83	0.9	2.4	7	0.66	1.3	1.7	0.76	1.1	2.0	1054-1071, 1085-1214, 1293-1316, 2179-2207
pH	NIR	3.4 - 4.3	0.2	0.85	0.09	2.6	7	0.71	0.13	1.8	0.73	0.13	1.9	912, 1018, 1178, 1280 - 1305, 1835 - 1875
	MIR			0.93	0.06	3.9	5	0.91	0.07	3.6	0.91	0.07	3.6	1718 - 1715, 1094, 1065, 1034, 998, 968
	Raman			0.59	0.2	1.5	9	0.43	0.2	1.5	0.37	0.2	1.3	/
	HSI			0.85	0.1	2.6	7	0.66	0.1	1.7	0.73	0.1	1.9	1054-1065, 1185-1280,1282-1327, 2179-2207
malic (g/kg)	NIR	3.0 - 7.5	1.0	0.80	0.5	2.1	8	0.61	0.7	1.4	0.66	0.7	1.5	912, 1018, 1178, 1365, 1384, 1843 - 1860, 1908
	MIR			0.81	0.5	2.2	6	0.79	0.6	1.6	0.78	0.6	1.8	1730 - 1715, 1095 - 1082, 1001 - 995, 968 - 962
	Raman			0.27	0.9	1.2	9	0.15	0.9	1.2	0.13	1.0	1.1	/
	HSI			0.80	0.5	2.0	7	0.65	0.7	1.5	0.70	0.6	1.7	1134, 1185-1280, 1338-1367, 1843-1860, 2196-2246
fructose (g/kg)	NIR	18.7 - 84.4	13.6	0.73	7.2	1.9	8	0.51	9.7	1.4	0.52	9.8	1.4	/
	MIR			0.85	5.2	2.6	6	0.79	7.2	1.9	0.84	5.4	2.5	1155, 1094, 1065, 1056, 1034, 980
	Raman			0.66	8.5	1.6	7	0.25	8.5	1.6	0.39	10.5	1.3	/
	HSI			0.74	7.1	1.9	7	0.43	9.6	1.4	0.57	9.2	1.5	/
sucrose (g/kg)	NIR	11.0 - 81.9	17.8	0.53	11.9	1.5	7	0.40	12.3	1.4	0.41	10	1.3	/
	MIR			0.78	9.4	1.9	8	0.76	8.9	1.7	0.75	9.8	1.8	/
	Raman			0.47	12.7	1.4	5	0.33	15.5	1.1	0.35	14.8	1.3	/

	HSI			0.61	11.1	1.6	7	0.15	16.3	1.1	0.27	15.1	1.2	/
glucose (g/kg)	NIR	10.0 - 22.5	2.9	0.35	2.3	1.2	4	0.31	2.6	1.1	0.39	2.2	1.3	/
	MIR			0.44	2.2	1.3	7	0.43	2.2	1.3	0.49	2.0	1.4	/
	Raman			0.11	2.9	1.0	8	0.03	2.8	1.0	0.09	2.8	1.0	/
	HSI			0.41	2.2	1.3	4	0.27	2.4	1.2	0.37	2.3	1.3	/
AIS (FW)	NIR	16.0 - 26.7	2.7	0.34	2.3	1.2	5	0.36	2.2	1.3	0.31	2.5	1.1	/
	MIR			0.42	2.2	1.2	10	0.57	1.8	1.5	0.51	1.9	1.4	/
	Raman			0.10	2.9	1.0	6	0.10	2.7	1.0	0.11	2.90	0.9	/
	HSI			0.35	2.3	1.2	6	0.21	2.5	1.1	0.30	2.4	1.1	/

Notes: Puree spectra and reference data from four varieties (‘Golden Delicious’, ‘Braeburn’, ‘Granny Smith’ and ‘Royal Gala’) with different fruit thinning practices for Golden Delicious apples (50 - 100 fruits/ tree or 150-200 fruits/ tree), stress treatments for Braeburn apples (11 days at 24 °C or 2 months at 4 °C), two processing recipes (70 °C for 15 mins with 3000 rpm grinding or 95 °C for 17 mins with 400 rpm grinding) and two refining conditions (refined at 0.5 mm or not refined). All results corresponded to 10-fold full-crossed validation tests. R_{cv}^2: determination coefficient of the full-crossed validation test; $RMSE_{cv}$: root mean square error of full-cross validation test; RPD: the residual predictive deviation of full-crossed validation test, LVs: the optimal numbers of latent variables. PLS-R: partial least square regression; RF-R: random forest regression; SVM-R: support vector machine regression.

Table S1. Chemical, structural and rheological characteristics of studied apple purees.

Groups	Process	Refining	Viscosity		G'	G"	Yield stress	tan δ	d4:3	d3:2	SSC	DMC	pH	TA	malic acid	fructose	sucrose	glucose	AIS
			K	*n*	Pa	Pa		-	-	-	(°Brix)	(g/g)		(meq/kg)	(g/kg)	(g/kg)	(g/kg)	(g/kg)	mg/g
GD Th-	I	NR	15.4	0.24	1245.9	247.4	12.8	0.20	267.4	174.5	14.4	0.21	3.8	5.4	4.9	67.7	73.3	17.5	138.3
		Ra	14.6	0.24	1158.8	222.5	12.6	0.19	264.5	172.4	14.1	0.20	3.7	5.5	4.6	67.2	66.4	17.1	117.8
	II	NR	23.1	0.21	1601.3	341.6	20.1	0.21	397.0	230.7	14.4	0.21	3.8	4.9	4.0	60.5	55.2	17.0	141.8
		Ra	20.1	0.21	1351.1	273.9	17.4	0.20	384.0	226.2	14.1	0.20	3.8	5.0	4.6	70.6	65.3	17.5	115.5
GD Th+	I	NR	11.4	0.27	984.4	190.7	9.7	0.19	262.9	177.4	15.3	0.22	3.7	5.9	5.5	78.7	73.0	16.3	145.2
		Ra	10.4	0.27	922.1	172.3	9.4	0.19	256.5	174.4	14.8	0.22	3.7	6.2	5.7	81.9	72.3	16.2	119.4
	II	NR	20.0	0.22	1390.7	290.9	16.9	0.21	382.1	228.5	14.7	0.21	3.7	5.5	5.1	76.2	60.8	17.5	141.1
		Ra	17.8	0.22	1200.4	241.5	15.0	0.20	371.4	223.8	14.5	0.21	3.7	5.7	5.2	71.6	63.1	17.6	118.3
GS	I	NR	32.0	0.21	1835.5	385.9	25.4	0.21	598.5	314.2	12.3	0.19	3.4	10.7	7.2	44.1	29.7	20.5	182.5
		Ra	22.3	0.22	1131.7	227.5	15.6	0.20	545.1	287.4	11.7	0.19	3.4	10.6	6.7	41.6	28.9	19.2	147.8
	II	NR	44.8	0.20	1794.5	543.2	25.8	0.30	774.4	399.5	12.2	0.19	3.4	10.4	6.4	42.8	25.5	19.9	169.7
		Ra	23.8	0.22	944.1	280.0	14.4	0.30	488.2	256.4	12.2	0.18	3.4	10.4	4.7	25.9	13.9	13.4	145.5
GA	I	NR	7.3	0.33	720.2	137.8	7.6	0.19	383.1	226.6	12.6	0.19	4.0	3.8	4.1	60.7	72.3	13.5	128.4
		Ra	7.1	0.32	675.6	125.6	7.4	0.19	372.6	223.0	12.4	0.19	4.1	3.9	4.2	65.8	71.6	13.6	119.7
	II	NR	12.4	0.27	934.0	194.3	10.5	0.21	440.3	261.1	12.5	0.18	4.3	3.7	3.6	53.8	57.0	12.5	124.5
		Ra	11.3	0.27	810.1	160.6	9.7	0.20	431.2	256.8	12.2	0.18	4.3	3.7	3.5	47.8	50.8	11.7	122.7
BR	I	NR	11.2	0.28	1080.3	215.7	12.5	0.20	421.7	227.5	12.9	0.19	3.6	6.7	5.6	54.7	43.0	17.1	156.1
		Ra	9.8	0.29	987.8	192.6	11.6	0.20	412.5	223.8	12.7	0.19	3.6	7.8	5.7	61.4	39.5	18.2	132.8
	II	NR	22.6	0.23	1508.8	323.3	19.8	0.21	537.8	283.5	13.3	0.20	3.5	6.7	5.7	61.9	37.4	18.7	154.5
		Ra	15.9	0.24	1054.8	210.1	13.7	0.20	499.8	267.8	13.1	0.19	3.5	6.9	5.7	61.8	41.2	18.4	122.7
BM	I	NR	8.0	0.29	965.1	200.2	7.7	0.21	241.7	172.1	12.6	0.19	3.7	5.8	4.1	51.5	36.0	16.5	145.4
		Ra	8.0	0.28	957.8	195.8	8.0	0.20	240.1	170.7	12.4	0.19	3.7	5.7	4.4	55.6	38.0	17.9	125.7
	II	NR	13.9	0.24	1373.1	309.9	12.3	0.23	292.6	212.1	13.2	0.18	3.8	5.5	4.7	59.0	39.4	21.1	142.7
		Ra	13.7	0.23	1288.3	278.7	12.0	0.22	286.7	199.8	12.7	0.18	3.7	5.5	4.9	69.2	39.8	20.8	117.8
SD			*8.7*	*0.04*	*321.7*	*92.5*	*5.2*	*0.03*	*129.7*	*53.2*	*1.1*	*0.01*	*0.2*	*2.2*	*1.0*	*13.6*	*17.8*	*2.9*	*18.3*
F-value and significance		***Cultivar***	*192.0*	*120.5*	*50.9*	*73.3*	*74.9*	*1071.5*	*394.5*	*386.2*	*117.1*	*58.8*	*1285.8*	*215.0*	*43.4*	*154.9*	*218.1*	*30.4*	*2.8*
			***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	*

Process	110.1	98.6	13.3	52.1	35.4	1609.1	218.6	303.4	0.02	0.4	47.2	52.6	16.1	21.4	54.9	1.8	1.6
	***	***	**	***	***	***	***	***	ns	ns	***	***	***	***	***	ns	ns
Refining	70.7	4.3	66.8	77.2	54.3	107.3	82.5	114.1	5.8	1.2	5.7	1.7	2.4	2.3	2.5	2.2	120.9
	***	*	***	***	***	***	***	***	ns	ns	ns	ns	ns	ns	ns	ns	***

Note: GD Th-: non-thinned Golden Delicious; GD Th+: thinned Golden Delicious; GS: Granny Smith; GA: Royal Gala; BR: crunchy Braeburn, stored at 4°C; BM: mealy Braeburn, stored at 24 °C. G', G'': storage and loss modulus, at an angular frequency of 10 rad/s; AIS: Alcohol insoluble solids. Data expressed in Fresh weight (FW) values correspond to the mean of 3 lots x 10 apples. Two processing strategies: Process I of 70 °C, 3000 rpm and Process II of 95 °C, 400 rpm. Processed purees with non-refining (NR) or refined at 0.5 mm. In grey, ANOVA results of puree cultivar, process and refining conditions. ns, *, **, ***: Non-significant or significant at $P < 0.05$, 0.01, 0.001 respectively.