Visible, near- and mid-infrared spectroscopy coupled with an innovative chemometric strategy to control apple puree quality

Weijie Lan, Sylvie Bureau, Songchao Chen, Alexandre Leca, Catherine M.G.C. Renard, Benoit Jaillais

Visible, Near- and Mid- infrared spectroscopy coupled with an innovative chemometric strategy to control apple puree quality

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

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

[b] Unité InfoSol, INRAE, Orléans 45075, France.

[c] TRANSFORM Division, INRAE, 44000 Nantes, France.

[d] INRAE-ONIRIS Unité Statistiques, Sensométrie, Chimiométrie (Stat SC), F-44322 Nantes, France.

Corresponding authors*

Sylvie Bureau (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

Fax: +33 432722492

Others authors

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

Benoit Jaillais: benoit.jaillais@inrae.fr

Alexandre Leca: Alexandre.Leca@inrae.fr

Songchao Chen: songchao.chen@inrae.fr

Weijie Lan: Weijie.Lan@inrae.fr

Highlights

- Vis-NIRS and MIRS coupled with PLS can detect the cultivar composition of mixed purees.
- MIRS evaluated with a high confidence the quality characteristics of formulated purees by PLS.
- Spectra of individual puree cultivars can be used to control the quality of formulated apple purees.
- MIRS coupled with MCR-ALS can reconstruct the puree mixture using the concentration profiles.

Abstract

Vis-NIRS, MIRS, and a combination of both coupled with PLS and machine learning were applied to i) trace the composed proportions of different apple varieties in formulated purees and ii) predict the quality characteristics of formulated purees from spectral information of initial puree cultivars. The PLS models could estimate proportions of each apple cultivar in puree mixtures using MIR spectra (RMSEP<8.1%, RPD> 3.6), especially for Granny Smith (RMSEP=2.7%, RPD=11.4). The concentration profiles from multivariate curve resolution-alternative least squares (MCR-ALS) made possible to reconstruct spectra of formulated purees. MIRS technique was evidenced to predict the final puree quality, such as viscosity (RPD>4.0), contents of soluble solids (RPD=4.1), malic acid (RPD=4.7) and glucose (RPD=4.3), based only on the spectral data of composed puree cultivars. Infrared technique should be a powerful tool for puree traceability, even for multicriteria optimization of final products from the characteristics of composed puree cultivars before formulation.

Key word: *Malus domestica* Borkh., Vis-NIR, MIR, machine learning, MCR-ALS

1. Introduction

Apple puree is an ideal source of healthy constituents such as polyphenols and fibers (Le Bourvellec et al., 2011) and ~~antioxidants such as polyphenols with their major polymeric form, procyanidins (Loncaric, Dugalic, Mihaljevic, Jakobek, & Pilizota, 2014; Oszmiański, Wolniak, Wojdyło, & Wawer, 2008; Rembiałkowska, Hallmann, & Rusaczonek, 2007). It~~ can be used as an intermediate for smoothies, fruit sauce, pie fillings and fruit-based baby food (Opatová, Voldřich, Dobiáš, & Čurda, 1992). ~~The industrial production of apple purees consists typically in cooking at 93 – 98°C for about 4 – 5 min, refining to remove seeds and skin pieces and then pasteurization at 90°C around 20 min to obtain a shelf life of 6 months at room temperature (Oszmiański, Wolniak, Wojdyło, & Wawer, 2008).~~ Puree quality characteristics vary with fruit genetics (Rembiałkowska, Hallmann, & Rusaczonek, 2007), storage (Loncaric, Dugalic, Mihaljevic, Jakobek, & Pilizota, 2014), cooking parameters (Picouet, Landl, Abadias, Castellari, & Viñas, 2009), grinding intensity (Espinosa et al., 2011) and refining (Lan, Jaillais, Leca, Renard, & Bureau, 2020). In order to reach an apple puree with anticipated and constant taste and texture, a mixture of proportions of different apple varieties is generally done, presenting also the most economic and efficient strategies for manufacturers (O'sullivan, 2016). ~~Most papers dealing with the apple processing have not considered this practice insofar as they have been focused only on one apple cultivar (Espinosa et al., 2011; Picouet, Landl, Abadias, Castellari, & Viñas, 2009; Keenan, Brunton, Butler, Wouters, & Gormley,~~

2011). Thus, developing rapid and reliable approaches to determine the puree formulation, including fruit cultivars and the proportions of each one, could be highly beneficial for fruit processed products and traceability control.

Infrared spectroscopy (visible-near and mid infrared) known as a rapid, relatively cheap, easy-to-use, non-destructive and automatable technique, has been applied for the quality analysis of apple based-products, such as juices (Kelly & Downey, 2005; León, Kelly, & Downey, 2005; Reid, Woodcock, O'Donnell, Kelly, & Downey, 2005) and wine (Peng, Ge, Cui, & Zhao, 2016). For fruit purees, the studies have mainly aimed at detecting adulterations in mixed purees of different fruit species (Contal, León, & Downey, 2002; Defernez, Kemsley, & Wilson, 1995; Kemsley, Holland, Defernez, & Wilson, 1996). Particularly, the MIR technique combined with partial least squares discrimination analysis (PLS-DA) detects the presence of apple starting at 20% in apple-raspberry mixed purees (Kemsley, Holland, Defernez, & Wilson, 1996). Similar detectable limits are obtained using Vis-NIRS coupled with a principal component analysis (PCA) and a linear discriminant analysis (LDA) in apple-strawberry mixed purees (Contal, León, & Downey, 2002). The infrared spectroscopy (Vis-NIR and MIR) appears as a potential tool to access the composition of purees prepared with several fruit species. However, so far, there has been no attempt to use such approaches for more advanced works on purees of apples only, but resulting from various proportions of different cultivars.

Further, for fruit processors, the ever-increasing variability of raw fruits may mean that their empirical knowhow may not be sufficient to produce expected and constant

102 final purees. The challenge is therefore to provide specific guidance for formulation of
103 final purees based on information of individual batches of single cultivar puree.
104 Multivariate curve resolution-alternative least square (MCR-ALS) has been widely
105 used to simultaneously elucidate the pure spectra of different species present in
106 processed products and their concentration profiles (de Juan & Tauler, 2006), such as
107 edible oils from different vegetable sources (Le Dréau, Dupuy, Artaud, Ollivier, &
108 Kister, 2009) and fruit juices with various organic acids (Silva, Lourenço, & de
109 Araujo, 2018). The interest of this approach is to reconstruct the spectra of final
110 processed products (in our case, formulated purees) according to the relative spectra
111 of individual components (single cultivar purees) by MCR-ALS. If so, the predictive
112 models of processed puree quality traits (physical and chemical) using the
113 reconstructed spectra dataset could open the possibility to provide a multicriteria
114 optimization of puree formulation based on the prior information of single cultivar
115 purees.
116 Partial least squares (PLS), a typical linear algorithm, has been used to successfully
117 determine the global quality parameters of apple purees using NIRS information, such
118 as titratable acidity, dry matter and soluble solids (Lan, Jaillais, Leca, Renard, &
119 Bureau, 2020). However, the overlapping of absorption bands linked to non-linear
120 rheological variations gave poor prediction of puree's texture by PLS regression.
121 Machine learning approaches, such as random forest (RF) and Cubist, have been
122 specially constructed to address large and complex nonlinear systems. Indeed, RF
123 algorithm allows a better detection of adulteration in formulated oils than PLS (de

Santana, Borges Neto, & Poppi, 2019). Cubist regression working as decision tree models, gives a higher prediction accuracy than RF and PLS regression in palm-based cooking oil (Goh et al., 2019).

Accordingly, Vis-NIRS, MIRS and the combination of both (CB) infrared spectra coupled with machine learning (RF and Cubist) and PLS regressions were applied in our work on apples to: i) assess the possibility to detect the proportions of specific cultivar purees in the formulated purees and evaluate the limits of the detection; ii) build models to evaluate the quality parameters of formulated purees obtained from different proportions of single cultivar purees; and then iii) use information of single cultivar purees to reconstruct spectra of formulated purees by MCR-ALS and investigate the possibility to develop regression models to guidance the quality of final purees.

2. Material and methods

2.1 Apple purees

2.1.1 Purees processing

Apples of four varieties: 'Golden Delicious'(GD), 'Granny Smith'(GS), 'Braeburn'(BR) and 'Royal Gala'(GA) (all abbreviations are shown in **Table 1**) were harvested at a commercial maturity from La Pugère experimental orchard (Mallemort, Bouches du Rhône, France) in 2019, and stored for up to 2 months at 4 °C and around 90% relative humidity to ensure starch regression. After sorting and washing, on three consecutive weeks, a batch of each apple cultivar (2 kg) was processed into purees in a multi-functional processing system (Roboqbo, Qb8-3, Bentivoglio, Italy) following

a Hot Break recipe: cooked at 95°C for 5 min at a 1500 rpm grinding speed, then cooled down to 65°C while maintaining the grinding speed. Finally, processed purees were conditioned in two hermetically sealed cans: one was cooled in a cold room (4°C) before formulation, while the other was stored at -20°C for biochemical measurement of individual sugars (fructose, sucrose and glucose) and malic acid.

2.1.2 Puree formulations

After processing the single-cultivar purees, a total of 6 experimental groups (named A to E) were prepared, each, with two apple cultivars (**Figure 1**). Each group (A-F) included 9 samples with different formulated proportions of weight, was divided into two subsets: the first including 6 proportions (10%-90%, 25%-75%, 50%-50%, 75%-25%, 90%-10%, 95%-5%) for the modeling set, while the second with 3 proportions (80%-20%, 33%-67%, 14%-86%) for the external prediction set. Finally, spectral measurements (Vis-NIR and MIR), chemical (soluble solids, titratable acidity, pH, dry matter) and physical (color and rheological tests) characterizations were performed on each sample (single and formulated purees).

2.2 Determination of quality traits

2.2.1 Physical characterizations

The puree color was determined three times through a dedicated glass cuvette using a CR-400 chromameter (Minolta, Osaka, Japan) and expressed in the CIE 1976 L*a*b* color space (illuminant D65, 0° view angle, illumination area diameter 8 mm). The puree rheological measurements, as flow curves, were carried out using a Physica MCR-301 controlled stress rheometer (Anton Paar, Graz, Austria) and a 6-vane

geometry (FL100/6W) with a gap of 3.46 mm, at 22.5°C. The flow curves were performed after a pre-shearing period of 1 minute at a shear rate of 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. The values of viscosity at 50 s^{-1} and 100 s^{-1} (η_{50} and η_{100} respectively) were kept as final indicators of the puree texture linked to sensory characteristics during consumption (Chen & Engelen, 2012).

2.2.2 Biochemical analyses

Soluble solids content (SSC) was determined with a digital refractometer (PR-101 ATAGO, Norfolk, VA, USA) and expressed in °Brix at 20°C. Titratable acidity (TA) was determined by titration up to pH 8.1 with 0.1 mol/L NaOH and expressed in mmol H^+ kg^{-1} of fresh weight (FW) using an autotitrator (Methrom, Herisau, Switzerland). Individual sugars and malic acid were quantified using colorimetric enzymatic kits (glucose: No. 10716251035; fructose: No. 10139106035, sucrose: No. 10716260035, malic acid: No. 10139068035) according to the manufacturer's instructions (R-biopharm, Darmstadt, Germany), respectively. The content of glucose, fructose, sucrose and malic acid were expressed in g kg^{-1} FW. These measurements were performed with a SAFAS flx-Xenius XM spectrofluorimeter (SAFAS, Monaco) at 570 nm for the sugars and 450 nm for malic acid. The dry matter content (DMC) was estimated from the weight of freeze-dried samples upon reaching a constant weight (freeze-drier, 5 days). The individual sugars (fructose, glucose, sucrose) and malic acid contents of formulated puree samples were calculated based on the measured values of processed single cultivar purees.

2.3 Spectrum acquisition

The Vis-NIR spectral data of purees was acquired with a multi-purpose analyzer spectrometer (Bruker Optics®, Wissembourg, France) at 23°C, which provides diffuse reflectance measurements with a spectral resolution of 8 cm^{-1} from 12500 to 4000 cm^{-1} (wavelength from 400 to 2500 nm). For each spectrum, 32 scans were recorded and averaged. The spectral acquisition and instrument adjustments were controlled by OPUS software Version 5.0 (Bruker Optics®). Puree were transferred into 10 mL glass vials (5 cm height x 18 mm diameter) which were placed on the automated sample wheel of the spectrophotometer. Each puree sample was measured three times on different aliquots. A reference background measurement was automatically activated before each data set acquisition using an internal Spectralon reference.

The MIR spectra of purees was collected 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. Three replications of spectral measurement were performed on different aliquots. The purees were placed at the surface of a zinc selenide (ATR-ZnSe) crystal with six internal reflections. Spectra with 32 scans for ATR-ZnSe were collected from 4000 cm^{-1} to 650 cm^{-1} with a 4 cm^{-1} resolution and were corrected against the background spectrum of air.

The whole spectral dataset of Vis-NIR or MIR included 36 spectra (3 replicates × 3 processing weeks × 4 varieties) of single-cultivar purees, 324 spectra of formulated

purees spectra for the modelling set (3 replicates × 3 processing weeks × 6 formulated puree groups × 6 proportions) and 162 spectra for the external prediction set (3 replicates × 3 processing weeks × 6 formulated puree groups × 3 proportions) described in 2.2.1 and **Figure 1**.

2.4 Statistical analyses of reference data

After ensuring normal distribution with a Shapiro-Wilk test (α=0.05), the reference data of processed purees were presented as mean values and the data dispersion within our experimental dataset expressed as standard deviation values (SD). Analysis of variance (ANOVA) was carried out to determine the significant differences due to the different single apple varieties (**Table S-1**) or formulated puree groups (**Table S-2**) using XLSTAT (version 2018.5.52037, Addinsoft SARL, Paris, France) data analysis toolbox. And the pairwise comparison between means was performed using Tukey's test. Principal component analysis (PCA) was carried out on all reference data of single-cultivar purees or of formulated purees to evaluate their discriminant contributions using Matlab 7.5 (Mathworks Inc. Natick, MA, USA) software.

2.5 MCR-ALS and spectra reconstruction

MCR-ALS (multivariate curve resolution-alternative least square) is an effective multivariate self-modelling curve resolution method developed by Tauler (de Juan & Tauler, 2006). The relative contributions given by MCR-ALS were obtained for both, the Vis-NIR (400-2500 nm) and MIR (900-1800 cm^{-1}) spectral information, using the formulated purees and their corresponding single-cultivar purees (**Figure 2**). For the formulated samples, one matrix D ($n \times \lambda$) was made up with the number of samples (n)

and the intensity at each wavenumbers or wavelengths (λ). The S^T matrix ($s\times \lambda$) is the spectroscopic matrix describing the 'pure' infrared spectra (λ) of all single-cultivar purees (s). The D matrix can be mathematically decomposed into the individual contributions related to the spectral information of 'pure' purees in matrix S^T according to Eq. (1) and is interactively transformed using an alternative least square (ALS) procedure as Eq (2).

$$D = CS^T + E \quad (1)$$

$$C = R(S^T)^+ \quad (2)$$

Matrix C ($n\times q$) is the concentration matrix describing the contribution of every single-cultivar purees (q) in reconstructed purees (n). E is the error matrix that provides the data variation not explained by their contributions. The matrix $(S^T)^+$ is the pseudo-inverse matrix of S^T. A general constraint used in curve resolution method is the non-negativity on the concentration profiles.

Once the concentration profiles (matrix C) for each single-cultivar spectrum, including Golden Delicious (C_{GD}), Granny Smith (C_{GS}), Braeburn (C_{BR}) and Royal Gala (C_{GA}), were obtained, they were used to reconstruct a new spectroscopic matrix R ($n \times k$) for monitoring all formulated purees. Each row R_i. (i=1,…n) was made up of a reconstructed spectrum. And each column R.j (j=1,…k) gave the reconstructed spectral intensity at a wavenumber of MIRS or a wavelength of Vis-NIRS based on the corresponding pure puree spectra of Golden Delicious (λ_{GD}), Granny Smith (λ_{GS}), Braeburn (λ_{BR}) and Royal Gala (λ_{GA}), following Eq (3).

$$R = C_{GD}\lambda_{GD} + C_{GS}\lambda_{GS} + C_{BR}\lambda_{BR} + C_{GA}\lambda_{GA} \quad (3)$$

2.6 Spectral multivariate regression

Spectral pre-processing and multivariate regression were performed with several packages ('prospectr' (Stevens & Ramirez-Lopez, 2013), 'pls' (Mevik, Wehrens, & Liland, 2019), 'Cubist' (Kuhn, Weston, Keefer, Coulter, & Quinlan, 2014) and 'caret' (Kuhn, 2015)) of the R software (version 2.6.2) (R Core Team, 2019). As demonstrated in previous works (Bureau et al., 2013; Ncama, Opara, Tesfay, Fawole, & Magwaza, 2017), the wavelengths from 400 to 2500 nm of Vis-NIR and the wavenumbers from 900 to 1800 cm^{-1} in MIR were selected (**Figure 3**). For all spectral datasets, standard normal variate (SNV), resampling (intervals= 5, 10, 15), and derivative transform calculation (Savitzky–Golay method, gap size = 11, 21, 31, 41) of first or second order were compared before multivariate regression. SNV pre-processing applied on the Vis-NIR and MIR data had the best performances to predict puree quality and was then systematically used.

The partial least square (PLS), Cubist and RF regression models were developed to i) detect the proportions of each apple varieties in puree samples (**Table 2**) and predict the quality characteristics of formulated purees based on ii) the acquired Vis-NIR, MIR and their combined infrared spectra (CB) (**Table 3**) or iii) the reconstructed Vis-NIR, MIR and CB spectra (**Table 4**). All aforementioned spectral matrices (Vis-NIRS, MIRS and CB) corresponded to the same reference dataset. The set of all modelling spectra (324 spectra) was randomly split, with two-thirds of the dataset (216 spectra) used for calibration and a third (108 spectra) for internal validation. Then, calibrated models were further validated with the external prediction set (162

spectra). The procedure was repeated 10 times in order to obtain the dispersion of values giving an idea of the model stability and robustness. The developed models performance was then described by the 10-times averaged values of the determination coefficients of internal validation (R_v^2) and external prediction (R_p^2), root mean square error of prediction (RMSEP), RPD (Residual Predictive Deviation) value as described by Nicolai (Nicolai et al., 2007). During model training, the variable importance (VIP) for each puree characteristics were computed using the 'varImp' function by 'caret' package in R software (Kuhn, 2015), which could be applied both on PLS and machine learning regressions (Parmley, Higgins, Ganapathysubramanian, Sarkar, & Singh, 2019).

3. Results and discussion

3.1 Characteristics of single-cultivar purees and formulated purees

After puree processing, the four different cultivars provided a large variability of appearance, in particular color and texture **(Figure S-1)**. According to PCA results taking into account their rheological and biochemical characteristics **(Figure S-2)**, 'Royal Gala' (GA) purees were clearly discriminated from the other purees along the first principal component (PC1), with significantly (p <0.001) lower TA, pH, glucose, malic acid and viscosity (η_{50} and η_{100}) (**Table S-1**). Particularly, the values of viscosity at a shear rate of 50 s^{-1} (η_{50}), which is commonly used to describe the in-mouth texture perception of fluid foods (Chen & Engelen, 2012), were much more lower in GA purees (547 ± 13 $Pa.s^{-1}$) than in 'Golden Delicious' (GD) (839 ± 53 $Pa.s^{-1}$) and 'Granny Smith' (GS) (904 ± 31 $Pa.s^{-1}$) purees (**Table S-1**). As expected, the

viscosity and global quality (SSC and TA) of the formulated purees were affected when prepared with GA purees (**Figure S-3**). For example, the formulated GA-GD (group C) or GA-GS purees (group E) provided a high range of viscosity **(Figure S-3c and d)** and composition (**Figure S-3e and f**), but with a limited variation of color (a* and b* values) **(Figure S-3a and b)**.

Remarkable changes (p <0.001) of color parameters (L*, a* and b*) allowed the separation of 'Braeburn' (BR) purees and the others along the second principal component (**Figure S-2 and Table S-2**). Particularly the redness (a* values) of formulated puree groups (**Figure S-3a**), the admixture of BR (groups B, D and F) introduced more intensive variations (from -4.33 to 2.35) than the others (groups A, C and E, from -4.77 to -1.52). The limited variations of yellowness (b* values) in formulated GD-GA purees resulted in differences below the visual detection threshold **(Figure S-3b)**. Consequently, different strategies of puree formulation, especially the mixtures with 'Royal Gala' or with 'Braeburn' purees, could provide variability in taste, texture and color.

3.2 Characteristics of formulated purees: determination of composed single-cultivar puree proportions

In this part, the ability of SNV pre-processed Vis-NIR, MIR and CB coupled with PLS, Cubist and RF regressions was compared to estimate the proportions of single-cultivar in all formulated purees (**Table 2**).

Both, Vis-NIR and MIR techniques were potentially able to estimate the proportions of single-variety puree in the formulated purees with good models presenting robust

determination coefficients for both internal validation (R_v^2) and external validation (R_v^2), acceptable RMSEP (<10%) and RPD values at least higher than 2.5 (Nicolai et al., 2007). For Vis-NIR technique, two regression methods, PLS and RF, showed an acceptable ability to estimate proportions of GS (R_P^2> 0.92, RPD> 3.4, RMSEP< 9.2%) and of BR (R_P^2> 0.95, RPD> 4.2, RMSEP< 7.9%) varieties in all formulated purees, based on the VIP wavelengths at 412 nm, 524 nm and 672 nm (**Figure S-4b and c**). The predictive errors obtained here for the mixture of two cultivars of the same species, apple, were lower than those obtained earlier for the mixture of two species, namely apple/raspberry (11.3%) (Contal, León, & Downey, 2002). The poor Vis-NIRS prediction results for GD (RMSEP> 17.4%, RPD< 1.7) and GA (RMSEP> 16.2%, RPD< 2.1) were probably due to their similar color (**Figure S-1**). As the VIP wavelengths of Vis-NIR models were mainly dominated in the visible spectral region (412-672 nm), the color variations were not enough to be used for prediction of proportions in formulated purees, especially in the group C (GD-GA) (**Figure S-3a & b**).

MIR provided a better prediction of the proportions of single-cultivar purees in the formulated purees than Vis-NIR. Moreover, the regression method affected the prediction results of MIR. PLS gave better prediction results (RMSEP<8.1%, RPD> 3.6) than Cubist (RMSEP<15.1%, RPD> 2.3) and RF (RMSEP<10.6%, RPD> 2.7). Particularly, MIRS combined with PLS reached the lowest determination error (RMSEP=2.7%, RPD=11.4) for GS compared with other cultivars (GD, GA, BR). The highest VIP values (**Figure S-5c**) at 1723 $cm^{-}1$, 1065 cm^{-1} and 1034 cm^{-1}

attributed respectively to malic acid, fructose and glucose (Bureau, Cozzolino, & Clark, 2019; Clark, 2016), were consistent with the existence of marked differences in chemical composition (SD and significance) between purees containing GS (**Table S-2**). The excellent PLS predictions obtained for BR (RMSEP=4.3%, RPD=7.7) were based on the VIP wavenumbers at 998 cm^{-1} and 1084 cm^{-1} related to sucrose and fructose (Bureau, Cozzolino, & Clark, 2019) (**Figure S-5b**). Besides the aforementioned spectral signal, the satisfactory assessments of GD and GA proportions (RMSEP<8.1%, RPD>3.6) were linked to the MIRS region between 1750 and 1650 cm^{-1} related to organic acids, pectins, proteins, phenolics and absorbed water (**Figure S-5a & d**) (Abidi, Cabrales, & Haigler, 2014; Canteri, Renard, Le Bourvellec, & Bureau, 2019; Kačuráková et al., 1999).

The CB spectra, including Vis-NIR and MIR regions, coupled with PLS (RPD>2.8, RMSEP<11.5%) and RF (RPD>3.0, RMSEP<9.5%) provided a satisfactory assessment of the proportions of single-cultivar purees (**Table 2**). However, the results on CB were not as good as for MIR only.

Consequently, to predict proportions of single-cultivar purees, Vis-NIR was suitable for the formulated samples presenting large diversity in the color range, with the use of Braeburn and Granny Smith apples for example, and under vacuum processing conditions providing a good puree color preservation. MIRS coupled with PLS was evidenced as a powerful tool to provide excellent estimations of puree proportions, mainly based on differing concentrations of individual sugars and acid. Combining Vis-NIR and MIR did not improve prediction.

3.3 Characteristics of formulated purees: prediction of quality traits

As previously, the different spectral areas, Vis-NIR, MIR or CB, of all formulated purees coupled with the different regression methods, PLS, Cubist and RF, were compared for their ability to predict color, rheological and biochemical characteristics of formulated purees (**Table 3**). MIR spectra coupled with PLS obtained the best predictions in comparison with Vis-NIR and CB, except for color. Indeed concerning the color parameters, a good prediction of a* values was obtained for all spectral areas with a RPD decreasing order Vis-NIR (RPD>4.0), CB (RPD>3.6) and MIR (RPD>3.3) for both PLS and machine learning regressions (Cubist and Random forest). Particularly, the best prediction of a* values was obtained on CB with PLS models (R_P^2=0.96, RPD=5.0), slightly better than in Vis-NIR (R_P^2=0.95, RPD=4.7).

MIR spectra coupled with PLS gave the best prediction (R_P^2>0.90, RPD>4.1) of the rheological parameters (η_{50} and η_{100}) (**Table 3**). The identified VIP wavenumbers were 1026, 1065, 1113 and 1720 cm^{-1} (**Figure S-6**). These dominant carbohydrate bands centered at 1000-1200 cm^{-1}, associated with C-OH and C-O-C vibration of glucose and fructose (Bureau, Cozzolino, & Clark, 2019), have also been identified to predict viscosity of tomato purees (Ayvaz et al., 2016). And an acceptable estimation of DMC was observed for all developed MIR models (RMSEP< 0.003, RPD>2.7).

For biochemical parameters, MIR coupled with PLS allowed a very good prediction of SSC (RMSEP=0.1, RPD=5.1) in accordance with previous results of apple and tomato purees (Ayvaz et al., 2016; Lan, Renard, Jaillais, Leca, & Bureau, 2020). In apples, SSC is strongly correlated to the presence of sugars, namely fructose, sucrose

388 and glucose. The two main sugars, fructose and sucrose, were satisfactorily predicted
389 with PLS (RPD>3.0) and the non-linear regressions, Cubist and RF (RPD>2.9).
390 However, MIR could not predict the glucose content (RPD<2.4) (**Table 3**).
391 Considering the different expressions of acidity such as pH, TA and malic acid content,
392 MIR coupled with PLS provided their excellent prediction with R_P^2>0.92 and
393 RPD>4.0. It can be noticed that Vis-NIRS gave also acceptable prediction of TA and
394 malic acid (R_P^2>0.87, RPD>2.9), better than our previous results in NIRS on apple
395 purees (Lan, Jaillais, Leca, Renard, & Bureau, 2020).
396 In comparison with Vis-NIRS and MIRS, the slight improvements of using the
397 combined spectra (CB) concerned only the prediction of a* values (**Table 3**).
398 Combining Vis-NIRS and MIRS spectra offered little improvement or even degraded
399 the results in comparison with MIRS alone for analyzing puree viscosity and chemical
400 variations, (**Table 3**). These conclusions were in accordance with previous works on
401 forage (Reeves, 1997) and beers (Iñón, Garrigues, & Guardia, 2006). They can be
402 explained by i) the limited ability to balance the important variables after combination
403 of two spectral domains with different resolutions (**Figure S-6**); and ii) the
404 involvement of non-relevant or unimportant spectral regions which disturbed the
405 calibration modelling by producing more noise.
406 In summary, MIRS coupled with PLS had promising ability to well estimate viscosity,
407 a* color parameter, DMC, SSC, pH, TA, malic acid, sucrose and glucose of
408 formulated purees, but not for fructose. Acceptable assessments of a*, TA, malic acid

409 and glucose were obtained with the Vis-NIR region, in which sensors could be easily
410 adapted for fruit processing.

411 **3.4 Characteristics of formulated purees: prediction of quality traits based on the**
412 **reconstructed spectra**

413 In order to compute the concentration profiles of relevant single-cultivar puree
414 compositions, MCR-ALS was applied on the Vis-NIR and MIR spectra of all
415 formulated purees and of the four single-cultivar purees, using two approaches: the 54
416 averaged formulated puree spectra and the 4 averaged single-cultivar puree spectra of
417 a) each week or b) over the three weeks. These two methods (a and b) obtained
418 similar concentrations, indicating their robustness over different processing weeks.
419 Results are only shown for method b taking into account different processing periods
420 (**Table S-3**)**.** Based on that, in total 486 spectra of formulated purees were
421 reconstructed based on their corresponding 36 single-cultivar spectra (4 varieties x 3
422 replications x 3 weeks).

423 Accurate predictions of the concentrations were obtained with MIRS. These
424 predictios were highly related to the proportions of the single-cultivar purees (**Table**
425 **S-3**). However, the results were not acceptable with Vis-NIRS (**Table S-3**). The
426 limited ability of Vis-NIRS was due to the high similarity in color between GA and
427 GD and so a poor prediction of the proportions GA/GD in formulated purees (**Table**
428 **2**). The concentration profiles of MIRS in each group (A-E) appeared to follow a
429 non-linear relationship along the variation of puree proportions.

430 Prediction models were then developed using these reconstructed MIR spectra and the

reference data characterized on the formulated purees (**Table 4**). Overall, reconstructed MIR spectra with PLS regression better predicted the puree characteristics than Cubist and RF regressions. What stands out in these results was the highly accurate PLS predictions (R_p^2>0.85, RPD>4.0) of rheological parameters (η_{50} and η_{100}) from reconstructed spectra (**Table 4**), which were close to those obtained from the spectra of formulated purees (R_p^2>0.90, RPD>4.1) (**Table 3**). Particularly, similar MIRS fingerprint wavenumbers were obtained in reconstructed spectra and directly on formulated purees described above, mainly 1720, 1113, 1065 and 1026 cm^{-1} related to acid and sugars (Bureau, Cozzolino, & Clark, 2019). The prediction of DMC was acceptable (RPD>2.5) as mentioned above with real spectra in **Table 3**. For color, a good prediction of a* value was obtained with both, PLS (R_p^2=0.92, PRD=3.5) and machine learning methods (R_p^2>0.89, PRD>3.2) but not for L* and b*. For SSC, although the slight lower R_p^2 and RPD values than the best results obtained directly on MIR spectra (RMSEP=0.13, RPD=5.1) (**Table 3**), the PLS and Cubist models had an acceptable ability to estimate it for all formulated purees (RMESP<0.20, RPD>4.1) (**Table 4**). Considering the global acidity parameters, acceptable PLS predictions (R_p^2>0.88, PRD>3.2) were obtained for pH and TA, with a lower performance than directly on real spectra (R_p^2>0.92, PRD>4.0 in **Table 3**). For individual sugars and acids, PLS models showed an excellent prediction of glucose and malic acids (R_p^2>0.94, RPD>4.3), and an acceptable prediction of sucrose (R_p^2=0.86, RPD=2.8) but not for fructose (RPD<2.5). The specific wavenumbers at 1034 cm^{-1} for glucose, 1723 cm^{-1} for malic acid and 998 cm^{-1} for sucrose, mainly

453 contributed to the PLS models both from reconstructed spectra and directly on puree
454 spectra. The decrease of prediction accuracy was possibly owing to the non-negativity
455 of the concentration profiles which could constrain the spectral reconstruction (Le
456 Dréau, Dupuy, Artaud, Ollivier, & Kister, 2009). Briefly, MIR spectra coupled with
457 the concentration profiles of MCR-ALS showed a potential way to directly estimate
458 the viscosity, a* color parameter, SSC, TA, malic acid, pH, fructose and glucose for
459 formulated purees depending only on the spectral information of the single-cultivar
460 purees.
461 Compared to the previous prediction models obtained on the real spectra of
462 formulated purees (**Table 3**), highly consistent specific fingerprints and acceptable
463 prediction results (**Table 4**) provided a justifiable explanation to use the spectra
464 reconstruction of formulated purees from spectra of single-cultivar purees. MCR-ALS
465 has been used in other ways to identify precisely the chemical species or track their
466 evolutions (Garrido, Rius, & Larrechi, 2008; de Juan & Tauler, 2006). Here, it was
467 firstly used with the concentration profiles to reconstruct spectra of processed
468 products based on the spectra of raw materials.

469 **4. Conclusion**

470 This was the first detailed work to show the ability of infrared spectroscopy coupled
471 to suitable chemometric methods as a powerful tool to trace different composed
472 cultivars and estimate their corresponding compositions in apple purees. Moreover, an
473 innovative chemometric method based on MCR-ALS was developed to reach
474 simultaneous targets in terms of composition (in % of different cultivars) and

475 physico-chemical properties (rheology, SSC, TA, DMC) of final puree products. As
476 far as we know, this was the first report concerning the control of the final fruit
477 product quality variations depending on the spectral information of the initial purees
478 using a spectral reconstruction approach.
479 Vis-NIR on formulated purees could detect the composed single cultivars purees with
480 large color differences, such as 'Granny Smith' (R_P^2> 0.92, RPD> 3.4, RMSEP< 9.2%)
481 and 'Braeburn' (R_P^2> 0.95, RPD> 4.2, RMSEP< 7.9%), but not for 'Golden
482 Delicious' and 'Royal Gala'. MIR had the potential to trace the composed apple
483 varieties with the excellent evaluations of 'Granny Smith' and 'Braeburn proportions'
484 (RMSEP<4.3%, RPD> 7.7) and the satisfactory assessments of 'Golden Delicious'
485 and 'Royal Gala' proportions (RMSEP<8.1%, RPD>3.6). And MIR could also predict
486 the internal quality (SSC, TA, DMC, viscosity, pH, fructose, malic acid) of formulated
487 purees coupled with PLS and machine learning regressions.
488 Innovatively, MIR technique opens the possibility to control and guidance the final
489 puree characteristics by simply scanning the single-cultivar apple purees, in order to
490 maintain the product quality or to drive the development of new products in apple
491 industry. For instance, after acquiring MIR spectra of the four single-cultivar purees,
492 our developed PLS models might be used in industry: i) to formulate purees with
493 defined SSC and viscosity (e.g. 15.0 ± 0.3 °Brix and 1500 ± 100 $Pa.s^{-1}$, which might
494 be reached with the formulate solutions as 75% GD-25% GS, 80% GD-20% BR and
495 90% GD-10% GA purees); or ii) to compare *in silico* the results of different puree
496 formulation strategies, such as 33.3% GD and 66.6% GS purees (low redness, high

497 acidity and viscosity) versus another strategy of 80% BR and 20% GA purees (more
498 redness, low acidity, low viscosity), depending on the market.
499 Further challenging works will be to investigate the possibility to reconstruct spectra
500 of final processed purees based on spectra acquired directly on raw apples to provide
501 non-destructive information guidance.
502

Acknowledgements

The authors thank Patrice Reling, Xuwei Liu, Barbara Gouble, Marielle Boge, Caroline Garcia and Gisèle Riqueau (INRAE, SQPOV unit) for their technical help. The 'Interfaces' project is an Agropolis Fondation Flashship 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). Weijie Lan was supported by a doctoral grant from Chinese Scholarship Council.

References

Abidi, N., Cabrales, L., & Haigler, C. H. (2014). Changes in the cell wall and cellulose content of developing cotton fibers investigated by FTIR spectroscopy. *Carbohydrate Polymers, 100*, 9-16. https://doi.org/10.1016/j.carbpol.2013.01.074.

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.

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://pubs.acs.org/doi/abs/ 10.1021/ac402428s.

Canteri, M. H. G., Renard, C. M. G. C., Le Bourvellec, C., & Bureau, S. (2019). ATR-FTIR spectroscopy to determine cell wall composition: Application on a large diversity of fruits and vegetables. *Carbohydrate Polymers, 212*, 186-196. https://doi.org/10.1016/j.carbpol.2019.02.021.

Chen, J., & Engelen, L. (2012). *Food oral processing: fundamentals of eating and sensory perception*: John Wiley & Sons.

Clark, C. J. (2016). Fast determination by Fourier-transform infrared spectroscopy of sugar–acid composition of citrus juices for determination of industry maturity standards. *New Zealand journal of crop and horticultural science, 44*(1), 69-82. https://doi.org/10.1080/01140671.2015.1131725.

Contal, L., León, V., & Downey, G. (2002). Detection and Quantification of Apple Adulteration in Strawberry and Raspberry Purées Using Visible and near Infrared Spectroscopy. *Journal of Near Infrared Spectroscopy, 10*(4), 289-299. https://doi.org/10.1255/jnirs.345.

de Santana, F. B., Borges Neto, W., & Poppi, R. J. (2019). Random forest as one-class classifier and infrared spectroscopy for food adulteration detection. *Food Chemistry, 293*, 323-332. https://doi.org/10.1016/j.foodchem.2019.04.073.

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, L., To, N., Symoneaux, R., Renard, C. M. G. C., Biau, N., & Cuvelier, G. (2011). Effect of processing on rheological, structural and sensory properties of apple puree. *Procedia Food Science, 1*, 513-520. https://doi.org/10.1016/j.profoo.2011.09.078.

Garrido, M., Rius, F. X., & Larrechi, M. S. (2008). Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and Bioanalytical Chemistry, 390*(8), 2059-2066.

https://doi.org/10.1007/s00216-008-1955-6.

Goh, K. M., Maulidiani, M., Rudiyanto, R., Wong, Y. H., Ang, M. Y., Yew, W. M., Abas, F., Lai, O. M., Wang, Y., & Tan, C. P. (2019). Rapid assessment of total MCPD esters in palm-based cooking oil using ATR-FTIR application and chemometric analysis. *Talanta, 198*, 215-223. https://doi.org/10.1016/j.talanta.2019.01.111.

Iñón, F. A., Garrigues, S., & de la Guardia, M. (2006). Combination of mid- and near-infrared spectroscopy for the determination of the quality properties of beers. *Analytica Chimica Acta, 571*(2), 167-174. https://doi.org/10.1016/j.aca.2006.04.070.

de Juan, A., & Tauler, R. (2006). Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. *Critical Reviews in Analytical Chemistry*, 36(3-4), 163-176. https://doi.org/ 10.1080/10408340600970005.

Kačuráková, M., Wellner, N., Ebringerová, A., Hromádková, Z., Wilson, R. H., & Belton, P. S. (1999). Characterisation of xylan-type polysaccharides and associated cell wall components by FT-IR and FT-Raman spectroscopies. *Food Hydrocolloids, 13*(1), 35-41. https://doi.org/10.1016/S0268-005X(98)00067-8.

Kemsley, E. K., Holland, J. K., Defernez, M., & Wilson, R. H. (1996). Detection of adulteration of raspberry purees using infrared spectroscopy and chemometrics. *Journal of Agricultural and Food Chemistry, 44*(12), 3864-3870. https://doi.org/10.1021/jf960089l.

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

Kuhn, M., Weston, S., Keefer, C., Coulter, N., & Quinlan, R. (2014). Cubist: Rule- And Instance-Based Regression Modeling. R package version 0.2.3. https://CRAN.R-project.org/package=Cubist.

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., 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.

Le Bourvellec, C., Bouzerzour, K., Ginies, C., Regis, S., Plé, Y., & Renard, C. M. G. C. (2011). Phenolic and polysaccharidic composition of applesauce is close to that of apple flesh. *Journal of Food Composition and Analysis, 24*(4-5), 537-547. https://doi.org/10.1016/j.jfca.2010.12.012.

Le Dréau, Y., Dupuy, N., Artaud, J., Ollivier, D., & Kister, J. (2009). Infrared study of aging of edible oils by oxidative spectroscopic index and MCR-ALS chemometric method. *Talanta, 77*(5), 1748-1756. https://doi.org/10.1016/j.talanta.2008.10.012.

Loncaric, A., Dugalic, K., Mihaljevic, I., Jakobek, L., & Pilizota, V. (2014). Effects of sugar addition on total polyphenol content and antioxidant activity of frozen and freeze-dried apple puree. *Journal of Agricultural and Food Chemistry, 62*(7), 1674-1682. https://doi.org/10.1021/jf405003u.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2019). pls: Partial least squares and principal component regression. R package version 2.7-2. https://CRAN.R-project.org/package=pls.

Ncama, K., Opara, U. L., Tesfay, S. Z., Fawole, O. A., & Magwaza, L. S. (2017). Application of

Vis/NIR spectroscopy for predicting sweetness and flavour parameters of 'Valencia' orange (Citrus sinensis) and 'Star Ruby' grapefruit (Citrus x paradisi Macfad). *Journal of Food Engineering, 193*, 86-94. https://doi.org/10.1016/j.jfoodeng.2016.08.015.

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.

O'Sullivan, M. (2016). A handbook for sensory and consumer-driven new product development: Innovative technologies for the food and beverage industry (pp 156-159): Woodhead Publishing.

Opatová, H., Voldřich, M., Dobiáš, J., & Čurda, D. (1992). Quality changes during the storage of apple puree. *Food /Nahrung, 36*(2), 129-134. https://doi.org/10.1002/food.19920360205.

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., & Singh, A. K. (2019). Machine Learning Approach for Prescriptive Plant Breeding. *Scientific Reports, 9*(1), 17132. https://doi.org/10.1038/s41598-019-53451-4.

Picouet, P. A., Landl, A., Abadias, M., Castellari, M., & Viñas, I. (2009). Minimal processing of a Granny Smith apple purée by microwave heating. *Innovative Food Science & Emerging Technologies, 10*(4), 545-550. https://doi.org/10.1016/j.ifset.2009.05.007.

R Core Team, R. C. (2019). R: A Language and Environment for Statistical Computing. *1*(1358), 34.

Reeves, J. B. (1997). Concatenation of Near- and Mid-Infrared Spectra To Improve Calibrations for Determining Forage Composition. *Journal of Agricultural and Food Chemistry, 45*(5), 1711-1714. https://doi.org/10.1021/jf9604246.

Rembiałkowska, E., Hallmann, E., & Rusaczonek, A. (2007). Influence of processing on bioactive substances content and antioxidant properties of apple purée from organic and conventional production in Poland. https://orgprints.org/10417/.

Silva, A. C., Lourenço, A. S., & de Araujo, M. C. U. (2018). Simultaneous voltammetric determination of four organic acids in fruit juices using multiway calibration. *Food Chemistry, 266*, 232-239. https://doi.org/10.1016/j.foodchem.2018.06.005.

Stevens, A., & Ramirez-Lopez, L. (2013). An introduction to the prospectr packageR package Vignette R package version 0.1. 3. https://CRAN. R-project. org/package= prospectr.

Figure captions:

Figure 1. Experimental scheme of purees reformation, quality characterizations and spectral acquisition.

Figure 2. Process of VIS-NIRS and MIRS data by multivariate curve resolution-alternative least square (MCR-ALS) and spectral reconstruction of reformulates puree samples.

Figure 3. Overview of the applied methodology of VIS-NIR and MIR spectra pre-processing and multivariate regression.

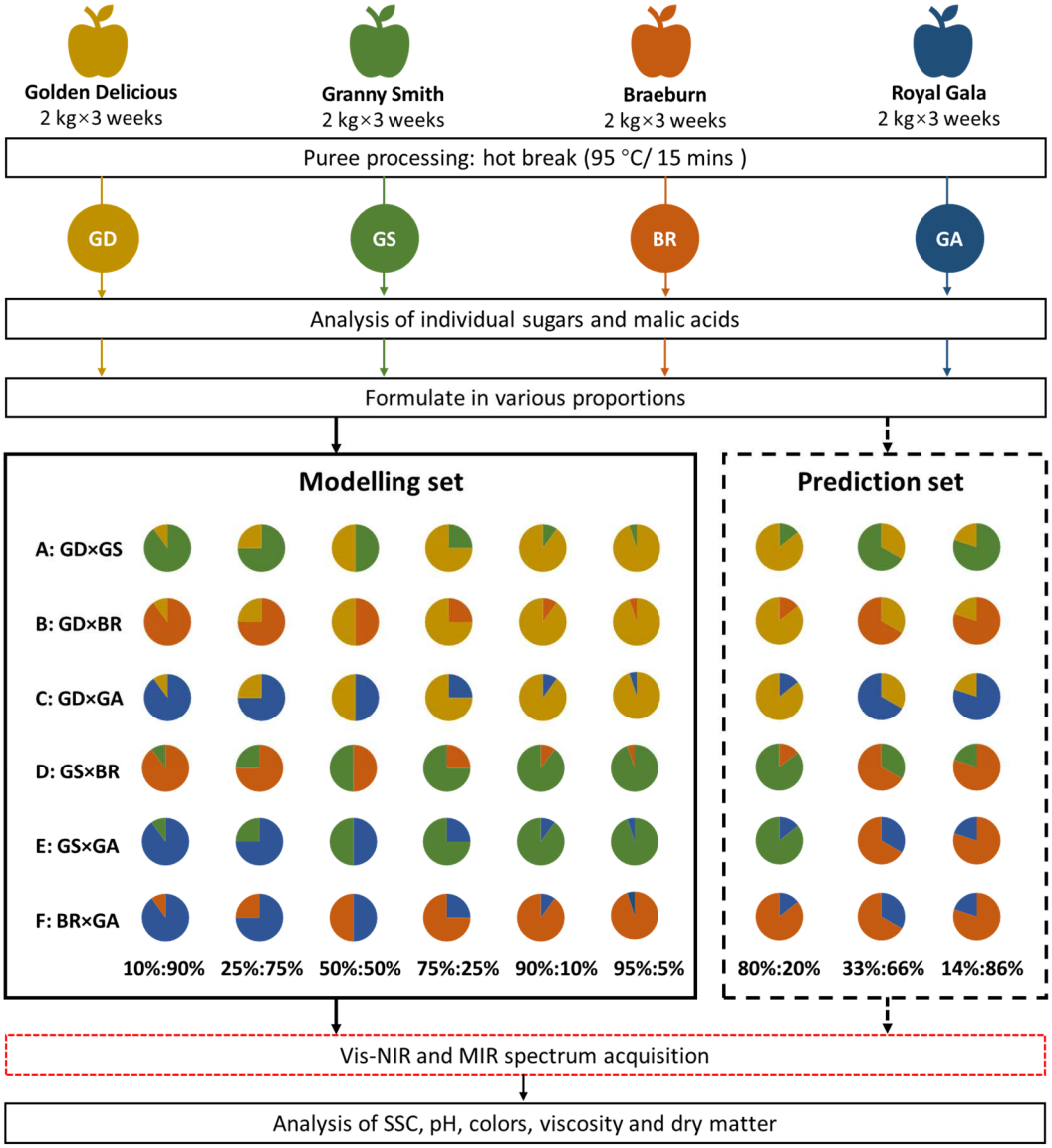
Golden Delicious
2 kg×3 weeks
Granny Smith
2 kg×3 weeks
Braeburn
2 kg×3 weeks
Royal Gala
2 kg×3 weeks
Puree processing: hot break (95 °C/ 15 mins)
GD
GS
BR
GA
Analysis of individual sugars and malic acids
Formulate in various proportions
Modelling set
Prediction set
A: GD×GS
B: GD×BR
C: GD×GA
D: GS×BR
E: GS×GA
F: BR×GA
10%:90%
25%:75%
50%:50%
75%:25%
90%:10%
95%:5%
80%:20%
33%:66%
14%:86%
Vis-NIR and MIR spectrum acquisition
Analysis of SSC, pH, colors, viscosity and dry matter

Figure 1

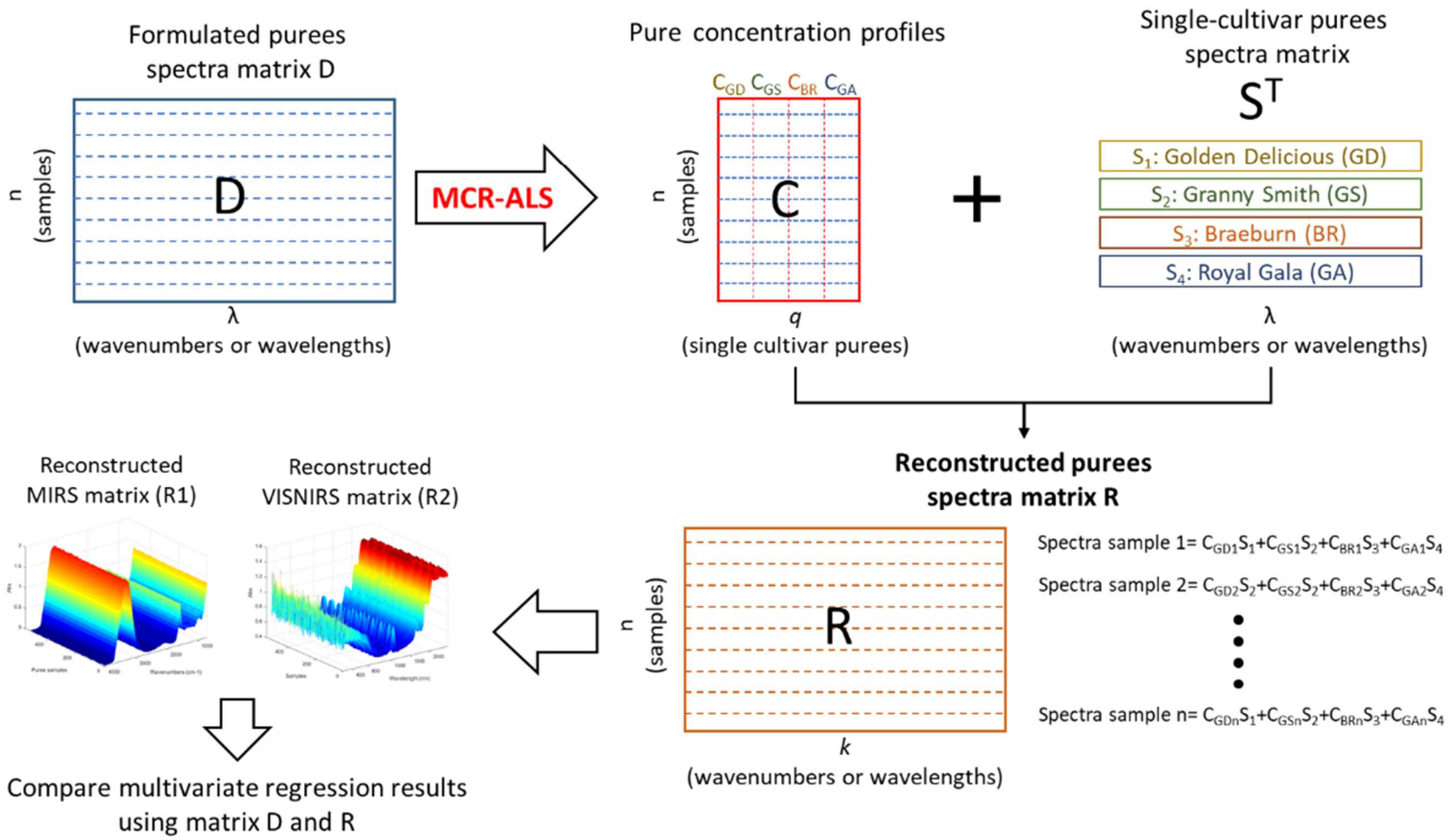
Formulated purees
spectra matrix D
n
(samples)
D
λ
(wavenumbers or wavelengths)
MCR-ALS
Pure concentration profiles
C_{GD} C_{GS} C_{BR} C_{GA}
n
(samples)
C
q
(single cultivar purees)
+
Single-cultivar purees
spectra matrix
S^T
S_1: Golden Delicious (GD)
S_2: Granny Smith (GS)
S_3: Braeburn (BR)
S_4: Royal Gala (GA)
λ
(wavenumbers or wavelengths)
Reconstructed purees
spectra matrix R
n
(samples)
R
k
(wavenumbers or wavelengths)
Spectra sample 1= $C_{GD1}S_1+C_{GS1}S_2+C_{BR1}S_3+C_{GA1}S_4$
Spectra sample 2= $C_{GD2}S_2+C_{GS2}S_2+C_{BR2}S_3+C_{GA2}S_4$
Spectra sample n= $C_{GDn}S_1+C_{GSn}S_2+C_{BRn}S_3+C_{GAn}S_4$
Reconstructed
MIRS matrix (R1)
Reconstructed
VISNIRS matrix (R2)
Compare multivariate regression results
using matrix D and R

Figure 2

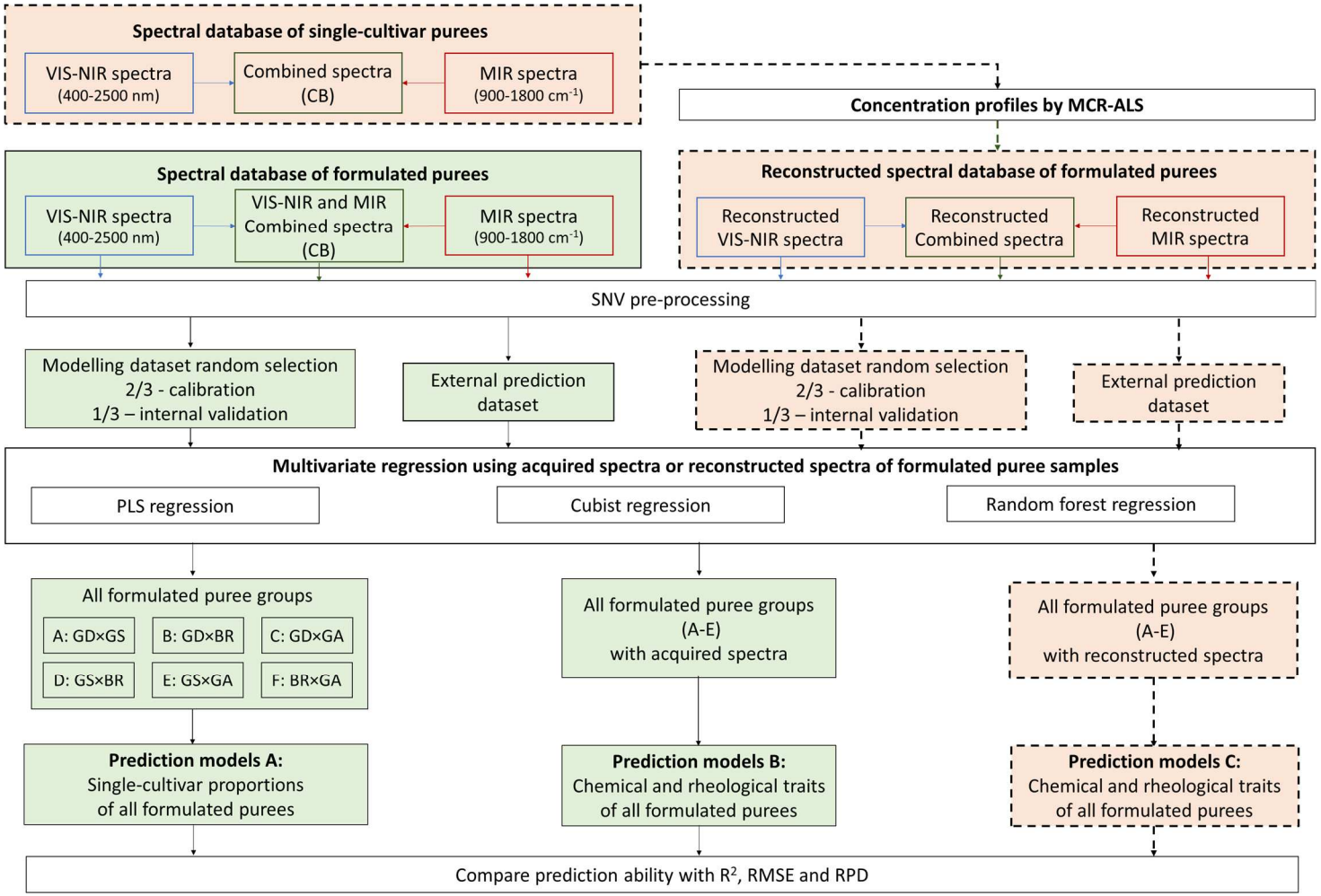
Spectral database of single-cultivar purees
VIS-NIR spectra (400-2500 nm)
Combined spectra (CB)
MIR spectra (900-1800 cm-1)
Concentration profiles by MCR-ALS
Spectral database of formulated purees
VIS-NIR spectra (400-2500 nm)
VIS-NIR and MIR Combined spectra (CB)
MIR spectra (900-1800 cm-1)
Reconstructed spectral database of formulated purees
Reconstructed VIS-NIR spectra
Reconstructed Combined spectra
Reconstructed MIR spectra
SNV pre-processing
Modelling dataset random selection 2/3 - calibration 1/3 – internal validation
External prediction dataset
Modelling dataset random selection 2/3 - calibration 1/3 – internal validation
External prediction dataset
Multivariate regression using acquired spectra or reconstructed spectra of formulated puree samples
PLS regression
Cubist regression
Random forest regression
All formulated puree groups
A: GD×GS
B: GD×BR
C: GD×GA
D: GS×BR
E: GS×GA
F: BR×GA
All formulated puree groups (A-E) with acquired spectra
All formulated puree groups (A-E) with reconstructed spectra
Prediction models A: Single-cultivar proportions of all formulated purees
Prediction models B: Chemical and rheological traits of all formulated purees
Prediction models C: Chemical and rheological traits of all formulated purees
Compare prediction ability with R2, RMSE and RPD

Figure 3

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

Common names	Abbreviations
‘Golden Delicious’ purees	GD
‘Granny Smith’ purees	GS
‘Braeburn’ purees	BR
‘Royal Gala’ purees	GA
partial least square	PLS
random forest regression	RF
combination of Vis-NIR and MIR	CB
multivariate curve resolution-alternative least square	MCR-ALS
dry matter contents	DMC
soluble solid contents	SSC
titratable acidity	TA
purees viscosity at a control share rate of 50 s^{-1}	η_{50}
purees viscosity at a control share rate of 100 s^{-1}	η_{100}
standard deviation values	SD
Principal component analysis	PCA
fresh weight	FW
standard normal variate	SNV
determination coefficients of internal validation	R_v^2
determination coefficients of external prediction	R_p^2
root mean square error of prediction	RMSEP
variable importance	VIP
Residual Predictive Deviation	RPD

Table 2. Prediction of the proportions (%) of single-cultivar purees in all formulated purees based on VIS-NIR (400- 2500 nm), MIR (900- 1800 cm^{-1}) and their combined spectra (CB; VIS-NIR-MIR). Comparison of three regression models (PLS, Cubist and Random forest)

Single-cultivar	Spectra	PLSR				Cubist				Random forest			
		R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD
GD	Vis-NIR	0.66	0.60	19.0	1.5	0.82	0.58	19.4	1.5	0.88	0.64	17.4	1.7
	MIR	0.94	0.92	8.1	3.6	0.95	0.86	11.3	2.6	0.94	0.87	10.6	2.7
	CB	0.91	0.88	10.3	2.8	0.93	0.82	12.7	2.3	0.96	0.90	9.5	3.0
BR	Vis-NIR	0.97	0.95	7.5	4.4	0.98	0.93	8.7	3.8	0.97	0.95	7.9	4.2
	MIR	0.99	0.98	4.3	7.7	0.98	0.97	5.2	6.3	0.98	0.95	7.6	4.3
	CB	0.99	0.98	5.0	6.6	1.00	0.97	5.6	5.9	0.98	0.97	6.1	5.4
GS	Vis-NIR	0.93	0.92	9.2	3.4	0.97	0.89	10.5	3.0	0.97	0.94	8.2	3.8
	MIR	0.99	0.99	2.7	11.4	0.99	0.93	8.1	3.8	0.98	0.97	5.3	5.8
	CB	0.99	0.98	4.3	7.3	0.99	0.98	4.9	6.4	0.98	0.97	5.8	5.4
GA	Vis-NIR	0.79	0.65	16.2	2.1	0.67	0.68	20.2	1.7	0.75	0.73	18.5	1.9
	MIR	0.96	0.94	7.4	4.7	0.90	0.82	15.1	2.3	0.91	0.90	10.3	3.4
	CB	0.89	0.83	11.5	3.0	0.88	0.79	16.2	2.2	0.94	0.92	9.4	3.7

Notes: single-cultivar purees of Golden Delicious named 'GD', Braeburn named 'BR', Granny Smith named 'GS', Royal Gala named 'GA'. All results corresponded to the averaged values of 10 replicates. R_v^2: determination coefficient of the validation test (internal); R_p^2: determination coefficient of the prediction test (external); RMSEP: root mean square error of prediction test (external) expressed as the puree proportions (%); RPD: the residual predictive deviation of prediction test. (external).

Table 3. Prediction of chemical and rheological parameters of all formulated purees using Vis-NIR (400-2500 nm), MIR (900-1800 cm^{-1}) or their combined spectra (CB) and regression methods, PLS, Cubist or Random forest.

Parameter	Spectra	Range	SD	PLSR				Cubist				Random forest			
				R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD
L*	Vis-NIR	41.6-48.6	1.5	0.81	0.70	0.8	1.6	0.87	0.63	0.9	1.4	0.88	0.75	0.6	1.9
	MIR			0.88	0.80	0.6	2.0	0.96	0.83	0.6	2.2	0.94	0.80	0.6	2.2
	CB			0.89	0.79	0.6	2.1	0.95	0.79	0.6	1.9	0.94	0.83	0.5	2.4
a*	Vis-NIR	(-4.8)-2.4	2	0.97	0.96	0.4	4.7	0.98	0.94	0.5	4.0	0.96	0.94	0.5	4.1
	MIR			0.96	0.94	0.5	4.0	0.98	0.92	0.5	3.6	0.97	0.91	0.6	3.3
	CB			0.98	0.96	0.4	5.0	0.99	0.93	0.5	3.6	0.98	0.94	0.5	4.1
b*	Vis-NIR	9.6-18.4	1.7	0.62	0.55	1.2	1.5	0.76	0.46	1.5	1.3	0.72	0.53	1.3	1.4
	MIR			0.67	0.56	1.2	1.5	0.86	0.48	1.4	1.3	0.84	0.62	1.1	1.6
	CB			0.67	0.53	1.3	1.5	0.88	0.46	1.4	1.3	0.81	0.57	1.2	1.5
Viscosity η_{50}	Vis-NIR	526-1029	119	0.79	0.81	54.6	2.2	0.85	0.85	49.8	2.4	0.82	0.78	57.8	2.1
	MIR			0.94	0.90	29.8	4.1	0.95	0.89	39.4	3.1	0.9	0.87	43.6	2.8
	CB			0.91	0.87	43.5	2.8	0.93	0.88	43.2	2.8	0.91	0.89	42.8	2.8
Viscosity η_{100}	Vis-NIR	834-1721	210	0.73	0.74	108.0	2.0	0.87	0.79	98.9	2.2	0.82	0.75	109.3	1.9
	MIR			0.94	0.91	52.0	4.1	0.96	0.86	81.2	2.6	0.90	0.88	74.4	2.9
	CB			0.88	0.87	79.6	2.7	0.91	0.87	76.5	2.8	0.91	0.88	77.3	2.8
DMC (g/g FW)	Vis-NIR	0.14-0.17	0.009	0.85	0.79	0.004	2.1	0.81	0.75	0.004	1.9	0.79	0.77	0.004	2.0
	MIR			0.93	0.89	0.003	3.1	0.91	0.88	0.003	2.7	0.93	0.90	0.003	3.0
	CB			0.85	0.83	0.003	2.5	0.96	0.83	0.003	2.5	0.93	0.87	0.003	2.8
SSC (°Brix)	Vis-NIR	12.1-15.3	0.7	0.61	0.53	0.5	1.5	0.79	0.56	0.5	1.3	0.78	0.62	0.5	1.5
	MIR			0.96	0.95	0.1	5.1	0.96	0.93	0.2	3.9	0.94	0.94	0.2	4.1
	CB			0.89	0.94	0.2	4.0	0.95	0.92	0.2	3.4	0.95	0.96	0.1	4.4

	Vis-NIR			0.37	0.38	7.3	1.2	0.52	0.25	8.4	1.1	0.70	0.50	6.3	1.4
fructose (g/kg FW)	MIR	40.2-80.3	9.1	0.82	0.78	3.7	2.4	0.93	0.81	4.0	2.2	0.92	0.70	4.8	1.8
	CB			0.67	0.56	5.8	1.5	0.83	0.74	4.4	2.0	0.91	0.76	4.4	2.0
	Vis-NIR			0.54	0.49	3.9	1.4	0.69	0.52	4.3	1.3	0.76	0.46	4.0	1.4
sucrose (g/kg FW)	MIR	33.2-57.3	5.5	0.89	0.89	1.8	3.0	0.88	0.89	1.8	2.9	0.92	0.92	1.6	3.3
	CB			0.60	0.67	3.1	1.7	0.92	0.87	2.0	2.6	0.87	0.78	2.5	2.1
	Vis-NIR			0.92	0.93	1.0	3.6	0.96	0.87	1.3	2.6	0.91	0.89	1.2	2.9
glucose (g/kg FW)	MIR	13.2-28.3	3.7	0.98	0.98	0.5	6.7	0.99	0.97	0.6	5.6	0.97	0.94	0.9	4.1
	CB			0.95	0.93	1.0	3.7	0.98	0.96	0.7	4.9	0.98	0.95	0.8	4.4
	Vis-NIR			0.84	0.83	0.1	2.4	0.94	0.76	0.1	2.0	0.82	0.76	0.1	2.0
pH	MIR	3.39-4.47	0.23	0.94	0.92	0.1	4.0	0.89	0.85	0.1	2.5	0.94	0.92	0.1	3.4
	CB			0.83	0.86	0.1	2.7	0.96	0.67	0.1	1.7	0.9	0.85	0.1	2.4
	Vis-NIR			0.93	0.87	5.0	2.9	0.95	0.90	5.1	3.1	0.96	0.89	5.2	3.0
TA (meq/kg FW)	MIR	28.0-94.8	16.2	0.99	0.96	3.5	4.3	0.99	0.94	3.9	3.9	0.96	0.91	4.7	3.3
	CB			0.95	0.91	4.9	3.1	0.98	0.95	3.8	4.0	0.96	0.9	4.8	3.2
	Vis-NIR			0.90	0.88	0.5	2.9	0.91	0.85	0.5	2.6	0.94	0.87	0.5	2.8
malic acid (g/kg FW)	MIR	3.0-8.8	1.3	0.97	0.97	0.2	5.9	0.95	0.92	0.4	3.7	0.94	0.94	0.3	4.2
	CB			0.92	0.92	0.4	3.4	0.91	0.84	0.5	2.5	0.96	0.93	0.4	3.7

Notes: all results corresponded to the averaged values of 10 replicates. R_v^2: determination coefficient of the validation test (internal); R_p^2: determination coefficient of the prediction test (external); RMSEP: root mean square error of prediction test (external); RPD: the residual predictive deviation of prediction test. (external).

Table 4. Prediction results of chemical and rheological parameters of all formulated purees from the reconstructed MIR spectra computed by the concentrations of MCR-ALS and the spectra of single-cultivar purees.

Parameter	Range	SD	PLSR				Cubist				Random forest			
			R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD	R_v^2	R_p^2	RMSEP	RPD
L*	41.6-48.6	1.5	0.91	0.86	0.5	2.4	0.9	0.83	0.6	1.9	0.86	0.78	0.6	2.1
a*	(-4.8)-2.4	2	0.92	0.92	0.5	3.5	0.94	0.89	0.6	3.2	0.93	0.91	0.6	3.4
b*	9.6-18.4	1.7	0.62	0.59	1.2	1.6	0.56	0.48	1.2	1.5	0.58	0.54	1.2	1.5
Viscosity ŋ50	526-1029	119	0.93	0.86	32.3	4.0	0.86	0.82	45.6	3.1	0.86	0.79	47.4	2.8
Viscosity ŋ100	834-1721	210	0.94	0.85	55.5	4.0	0.86	0.83	81	2.8	0.85	0.78	85.3	2.7
DMC (g/g FW)	0.14-0.17	0.009	0.87	0.85	0.003	2.7	0.85	0.84	0.003	2.6	0.89	0.82	0.004	2.5
SSC (°Brix)	12.1-15.3	0.7	0.95	0.9	0.2	4.1	0.9	0.85	0.2	4.1	0.79	0.73	0.3	2.3
fructose (g/kg FW)	40.2-80.3	9.1	0.84	0.79	4.0	2.1	0.88	0.82	3.7	2.5	0.83	0.8	3.7	2.3
sucrose (g/kg FW)	33.2-57.3	5.5	0.88	0.86	2.0	2.8	0.87	0.85	2.1	2.7	0.88	0.83	2.1	2.7
glucose (g/kg FW)	13.2-28.3	3.7	0.94	0.94	0.9	4.3	0.97	0.9	1.1	3.2	0.93	0.94	0.9	3.7
pH	3.39-4.47	0.23	0.89	0.88	0.1	3.2	0.89	0.83	0.1	2.8	0.86	0.79	0.1	2.7
TA (meq/kg FW)	28.0-94.8	16.2	0.92	0.91	4.4	3.4	0.91	0.88	5.9	2.7	0.92	0.92	4.4	3.4
malic (g/kg FW)	3.0-8.8	1.3	0.95	0.93	0.3	4.7	0.94	0.87	0.4	3.9	0.95	0.95	0.3	4.3

Notes: all results corresponded to the averaged values of 10 replicates. R_v^2: determination coefficient of the validation test (internal); R_p^2: determination coefficient of the prediction test (external); RMSEP: root mean square error of prediction test (external); RPD: the residual predictive deviation of prediction test (external).

Supplementary Tables:

Table S-1. Mean values with the characteristics of single-cultivar purees differed significantly using Tukey's test.

Cultivar	Viscosity η50	Viscosity η100	L*	a*	b*	SSC (°Brix)	DMC (g/g FW)	pH	TA (meq/kg FW)	malic acid (g/kg FW)	fructose (g/kg FW)	sucrose (g/kg FW)	glucose (g/kg FW)
GD	838.6± 69.6 a	1388.5± 138.2 ab	47.0± 0.3 a	-4.1± 0.3 c	15.8± 0.5 a	14.5± 0.4 a	0.167± 0.003 a	3.9± 0.1 b	53.9± 0.8 c	6.0± 0.1 b	72.8± 6.2 a	46.7± 6.6 a	18.7± 0.8 b
GS	904.2± 18.8 a	1501.2± 18.9 a	45.1± 0.8 b	-4.3± 0.9 c	14.4± 1.4 a	13.2± 0.3 b	0.152± 0.003 b	3.6± 0.2 c	89.1± 1.3 a	8.1± 0.5 a	58.6± 14.9 ab	37.4± 4.3 b	26.8± 1.3 a
BR	736.8± 61.1 b	1229.1± 106.2 b	42.2± 0.4 c	1.8± 0.3 a	10.2± 1.2 b	13.1± 0.5 b	0.151± 0.005 b	3.7± 0.2 bc	62.7± 1.5 b	5.9± 0.1 b	59.4± 6.6 ab	50.0± 3.7 a	17.2± 1.5 b
GA	547.1± 38.0 c	860.9± 59.5 c	45.6± 0.5 b	-2.5± 0.2 b	15.0± 0.8 a	12.4± 0.6 b	0.143± 0.008 b	4.3± 0.2 a	29.3± 1.3 d	3.5± 0.7 c	49.3± 4.0 b	36.0± 1.3 b	14.6± 1.3 c

Note: Data are expressed as puree fresh weight (FW) ± standard deviation. Puree cultivars: Golden Delicious ('GD'); Granny Smith ('GS'), Braeburn ('BR') and Royal Gala ('GA').

Table S-2. Mean values with the characteristics of formulated puree groups differed significantly using Tukey's test.

Groups	Viscosity η50	Viscosity η100	L*	a*	b*	SSC (°Brix)	DMC (g/g FW)	pH	TA (meq/kg FW)	malic acid (g/kg FW)	fructose (g/kg FW)	sucrose (g/kg FW)	glucose (g/kg FW)
A: GD×GS	854.1± 70.3 a	1412.8± 124.5 a	46.3± 1.0 a	-4.5± 0.2 d	15.0± 1.2 a	13.9± 0.7 a	0.160± 0.007 a	3.7± 0.1 e	70.0± 10.5 b	7.0± 0.8 a	66.1± 6.4 a	42.3± 3.5 b	22.6± 2.8 a
B: GD×BR	767.2 ± 67.4 b	1278.8± 128.7 b	45.2± 1.6 b	-1.3± 1.9 b	13.9± 1.9 b	13.9± 0.5 a	0.160± 0.006 a	3.7± 0.1 d	58.7± 4.3 c	5.9± 0.1 b	66.4± 5.5 a	48.2± 3.1 a	18.0± 0.9 c
C: GD×GA	684.7± 90.5 c	1127.0± 172.3 cd	46.1± 0.9 a	-3.2± 0.7 c	14.7± 0.7 a	13.4± 0.7 b	0.156± 0.008 b	4.1± 0.2 a	42.7± 7.9 e	4.8± 0.8 c	61.6± 11.7 b	41.6± 4.7 b	16.8± 1.6 cd
D: GS×BR	853.8± 99.2 a	1408.5± 173.0 a	43.6± 1.0 c	-1.4± 2.0 b	12.4± 1.4 c	13.2± 0.5 b	0.154± 0.007 d	3.6± 0.1 e	76.6± 7.5 a	7.2± 0.8 a	59.0± 2.9 b	43.4± 5.1 b	22.3± 3.1 ab
E: GS×GA	743.2± 115.2 b	1209.1± 197.0 bc	45.1± 0.5 b	-3.5± 0.8 c	14.4± 1.1 ab	12.9± 0.5 c	0.150± 0.007 c	3.8± 0.2 c	60.5± 17.8 c	5.9± 1.5 b	54.2± 8.7 c	36.7± 2.0 c	21.0± 4.0 b
F: BR×GA	651.7± 86.2 c	1061.4± 161.1 d	44.1± 1.2 c	-0.3± 1.5 a	12.8± 1.7 c	12.8± 0.3 c	0.148± 0.003 c	4.0± 0.2 b	48.7± 10.2 d	4.7± 0.8 c	54.6± 7.2 c	43.3± 5.6 b	16.0± 1.1 d

Note: Data are expressed as puree fresh weight (FW) ± standard deviation. Puree cultivars: Golden Delicious ('GD'); Granny Smith ('GS'), Braeburn ('BR') and Royal Gala ('GA').

Table S-3. The VIS-NIR (400-2500 nm) and MIR (900-1800 cm^{-1}) spectral concentration profiles of each apple cultivar in formulated puree obtained from MCR-ALS.

Groups	Proportions	MIRS (900-1800 cm^{-1})				VIS-NIRS (400-2500 nm)			
		BR	GA	GD	GS	BR	GA	GD	GS
A: GD×GS	95%:5%	0.012	0.000	0.925	0.057	0.000	0.162	0.635	0.193
	90%:10%	0.010	0.000	0.911	0.073	0.000	0.275	0.381	0.333
	75%:25%	0.015	0.049	0.790	0.143	0.000	0.000	0.833	0.169
	50%:50%	0.015	0.000	0.566	0.415	0.000	0.000	0.595	0.406
	25%:75%	0.018	0.000	0.328	0.649	0.000	0.000	0.151	0.857
	10%:90%	0.075	0.000	0.100	0.809	0.000	0.000	0.254	0.731
	80%:20%	0.008	0.000	0.772	0.211	0.012	0.231	0.358	0.398
	33%:66%	0.079	0.003	0.337	0.571	0.033	0.061	0.043	0.868
	14%:86%	0.090	0.120	0.179	0.603	0.005	0.010	0.000	0.979
B: GD:BR	95%:5%	0.050	0.004	0.942	0.004	0.034	0.243	0.564	0.148
	90%:10%	0.081	0.010	0.897	0.000	0.000	0.250	0.621	0.133
	75%:25%	0.235	0.017	0.738	0.000	0.146	0.017	0.812	0.034
	50%:50%	0.495	0.004	0.492	0.000	0.252	0.000	0.764	0.000
	25%:75%	0.711	0.003	0.279	0.000	0.657	0.159	0.073	0.117
	10%:90%	0.847	0.013	0.135	0.000	0.866	0.079	0.000	0.058
	80%:20%	0.141	0.024	0.826	0.000	0.144	0.000	0.865	0.000
	33%:66%	0.515	0.000	0.425	0.058	0.599	0.164	0.103	0.133
	14%:86%	0.827	0.000	0.171	0.002	0.673	0.000	0.328	0.000
C: GD:GA	95%:5%	0.000	0.062	0.933	0.000	0.000	0.078	0.815	0.101
	90%:10%	0.048	0.047	0.897	0.005	0.000	0.000	0.995	0.000
	75%:25%	0.028	0.239	0.728	0.000	0.000	0.407	0.533	0.061
	50%:50%	0.006	0.474	0.500	0.016	0.019	0.584	0.272	0.113
	25%:75%	0.000	0.732	0.263	0.000	0.000	0.601	0.416	0.000
	10%:90%	0.000	0.907	0.087	0.000	0.000	0.521	0.483	0.000
	80%:20%	0.019	0.165	0.807	0.005	0.000	0.528	0.281	0.183
	33%:66%	0.003	0.642	0.352	0.001	0.000	0.367	0.632	0.000
	14%:86%	0.000	0.836	0.142	0.021	0.000	0.549	0.461	0.000
D: GS:BR	95%:5%	0.220	0.090	0.002	0.677	0.000	0.000	0.068	0.927
	90%:10%	0.283	0.259	0.000	0.444	0.001	0.000	0.000	0.994
	75%:25%	0.403	0.163	0.013	0.410	0.205	0.000	0.011	0.768
	50%:50%	0.636	0.138	0.000	0.217	0.417	0.000	0.092	0.492
	25%:75%	0.782	0.067	0.037	0.106	0.631	0.000	0.064	0.312
	10%:90%	0.951	0.005	0.000	0.040	0.853	0.000	0.000	0.144
	80%:20%	0.341	0.162	0.000	0.488	0.143	0.000	0.031	0.822
	33%:66%	0.738	0.082	0.000	0.171	0.543	0.000	0.201	0.249
	14%:86%	0.846	0.000	0.000	0.144	0.668	0.000	0.352	0.000
E: GS:GA	95%:5%	0.000	0.265	0.000	0.725	0.000	0.000	0.142	0.849
	90%:10%	0.006	0.219	0.000	0.787	0.000	0.000	0.178	0.822
	75%:25%	0.030	0.419	0.000	0.537	0.091	0.071	0.058	0.773

	50%:50%	0.009	0.655	0.000	0.334	0.093	0.000	0.560	0.340
	25%:75%	0.003	0.820	0.000	0.169	0.024	0.436	0.348	0.195
	10%:90%	0.047	0.702	0.238	0.000	0.000	0.068	0.943	0.000
	80%:20%	0.050	0.365	0.000	0.574	0.000	0.000	0.286	0.721
	33%:66%	0.000	0.799	0.000	0.193	0.000	0.453	0.114	0.431
	14%:86%	0.000	0.991	0.000	0.000	0.000	0.428	0.530	0.046
	95%:5%	0.785	0.209	0.000	0.000	0.706	0.000	0.314	0.000
	90%:10%	0.849	0.149	0.000	0.000	0.737	0.131	0.076	0.058
	75%:25%	0.599	0.398	0.000	0.000	0.513	0.000	0.496	0.000
	50%:50%	0.406	0.593	0.000	0.000	0.292	0.620	0.000	0.097
F: BR:GA	25%:75%	0.138	0.857	0.000	0.000	0.000	0.541	0.471	0.000
	10%:90%	0.051	0.946	0.000	0.000	0.000	0.533	0.475	0.000
	80%:20%	0.732	0.263	0.000	0.000	0.649	0.194	0.000	0.162
	33%:66%	0.254	0.745	0.000	0.000	0.176	0.644	0.033	0.136
	14%:86%	0.067	0.932	0.000	0.000	0.804	0.000	0.183	0.014

Puree cultivars: Golden Delicious ('GD'); Granny Smith ('GS'), Braeburn ('BR') and Royal Gala ('GA').