



SHORT COMMUNICATION

Accurate varietal classification and quantification of key quality compounds of grape extracts using the absorbance-transmittance fluorescence excitation emission matrix (A-TEEM) method and machine learning

Adam M. Gilmore^{1*}, Qiang Sui², Bryant Blair² and Bruce S. Pan²

¹ HORIBA Instruments Incorporated, Piscataway NJ 08854 USA

² E & J Gallo Winery, Modesto, CA 95354 USA



*correspondence:
adam.gilmore@horiba.com

Associate editor:
Victor de Freitas



Received:
26 May 2022

Accepted:
6 October 2022

Published:
9 November 2022



This article is published under the **Creative Commons licence (CC BY 4.0)**.

Use of all or part of the content of this article must mention the authors, the year of publication, the title, the name of the journal, the volume, the pages and the DOI in compliance with the information given above.

ABSTRACT

Rapid and accurate quantification of grape berry phenolics, anthocyanins and tannins and identification of grape varieties are both important for effective quality control of harvesting and initial processing for winemaking. Current reference technologies, including High-Performance Liquid Chromatography (HPLC), can be rate-limiting and too complex and expensive for effective field operations. In this paper, we analyse robotically prepared grape extracts from several key varieties (n = Calibration/n = Prediction samples), including Cabernet-Sauvignon (64/10), Grenache (16/4), Malbec (14/4), Merlot (56/10), Petite Sirah (52/10), Pinot noir (54/8), Syrah (20/2), Teroldego (14/2) and Zinfandel (62/12). Key phenolic and anthocyanin parameters measured by HPLC included Catechin, Epicatechin, Quercetin Glycosides, Malvidin 3-glucoside, Total Anthocyanins and Polymeric Tannins. Split samples diluted 50-fold in 50 % EtOH pH 2 were analysed in parallel using the A-TEEM method following Multi-block Data Fusion of the absorbance and unfolded EEM data. A-TEEM chemical data were calibrated (n = 390) using Extreme Gradient Boosting (XGB) Regression and evaluated based on the Root Mean Square Error of the Prediction (RMSEP), the Relative Error of Prediction (REP) and Coefficient of Variation (R2P) of the Prediction data (n = 62). The regression results yielded an average Relative Error of Prediction (REP) of 5.89 ± 2.47 % and an R2P of 0.941 ± 0.025 . While we consider the REP values to be in the acceptable range at significantly < 10 %, we acknowledge that both the grape extraction method repeatability and HPLC reference method sample repeatability (5-8 % RSD) likely constituted the major sources of variation compared to the A-TEEM instrumental sample repeatability (< 2 % RSD). The varietal classification was analysed using Agglomerative Hierarchical Cluster Analysis (HCA) and XGB discrimination analysis of the multi-block data. The classification results yielded 100 % True Positive and True Negative responses for the Calibration and Prediction Data for all tested varieties. We conclude that the A-TEEM method requires a minimum of sample preparation and rapid acquisition times (< 1 min) and can serve as an accurate secondary method for both grape varietal identification and phenolic quantification. Importantly, the software application of the regression and classification models can be effectively automated for operators.

KEYWORDS: multi-block modelling, multivariate analysis, discrimination, regression, winemaking

INTRODUCTION

Facilitation of grape quality evaluation is a significant concern, especially for large-scale winemakers dealing with multiple vineyards, varieties and targeted products and quality levels. Planning harvest and sorting grapes for varying target products call for a rapid and accurate methodology. Whereas High-Performance Liquid Chromatography (HPLC) is recognised as the primary reference method for wine phenolic analysis, it may be too expensive, time-consuming and complex for effective implementation scale-up. Hence secondary methods, including Fourier Transform Infrared (FTIR), Near and Mid-Infrared (NIR and MIR), Nuclear Magnetic Resonance (NMR) and UV-VIS spectroscopies, have been examined with respect to their ability to be calibrated against HPLC and other reference chemistry methods (Aleixandre-Tudo *et al.*, 2017; Harbertson *et al.*, 2003; Harbertson and Spayd, 2006; Miramont *et al.*, 2020; Rouxinol *et al.*, 2022; Sen *et al.*, 2016). The consensus is that while these secondary spectroscopic techniques provide generalised compound class information, they do not exhibit highly correlated molecular specificity for the key phenolic compounds generally regarded as quality parameters. However, in several reports, as reviewed by Ranaweera *et al.* (2021a), UV-VIS, FTIR and NIR, as well as NMR spectroscopies, have provided significant classification results for grape and wine varieties. Interestingly reports of successful data fusion for UV-VIS and FTIR for effective wine classification have also been reported (Geană *et al.*, 2019; Sen *et al.*, 2016). HPLC has been used for arguably the most highly effective varietal classification, albeit as mentioned above, the method is relatively slow and laborious compared to spectroscopic methods (Peng *et al.*, 2002; de Villiers *et al.*, 2005; Rouxinol *et al.*, 2022; Urcan *et al.*, 2016).

Recent developments in the areas of phenolic quantification and varietal and regional classification, focusing only on fermented wine products, have been documented using a new patented technique known as Absorbance-Transmittance fluorescence Excitation-Emission Matrix (A-TEEM) spectroscopy (Gilmore and Tong, 2019). A-TEEM is typically coupled with machine learning methods for both effective wine classification (Niimi *et al.*, 2020; Ranaweera *et al.*, 2021a; Ranaweera *et al.*, 2021b; Ranaweera *et al.*, 2022) and specific phenolic quantification with repeatability statistics on par with HPLC and other basic chemistry methods (Ranaweera *et al.*, 2021b; Ranaweera *et al.*, 2022; Schober *et al.*, 2022). The A-TEEM method allows rapid (s-min) acquisition rates and requires a very small amount of test material. The A-TEEM's potential to identify and quantify specific chemicals is reinforced by several key features, including primarily the automated correction of fluorescence inner-filter effects (IFEs), which is critical for linear quantification (Gilmore, 2014; Gilmore and Tong, 2019). However, there is also the fact that the A-TEEM, when applied under Beer-Lambert linear concentration conditions, simultaneously evaluates five optical molecular parameters, including 1) the absorbance extinction coefficients, 2) the fluorescence

quantum efficiencies, 3) absorbance spectral shape and the fluorescence, 4) excitation- and 5) emission-spectral shapes. Importantly, simultaneous analysis of the absorbance and fluorescence variables using a 'multi-block' model or data-fusion approach facilitates an effective means to incorporate and leverage all five parameters in machine learning applications which can be significant regarding the ability to identify and quantify specific quality marker compounds in complex grape extract matrices (Ranaweera *et al.*, 2021b; Schober *et al.*, 2022).

In this study, we select a set of the major compounds that are important to measure industrial grape quality (Cleary *et al.*, 2015). These include 1) Polymeric tannins which in wine are mostly grape-derived, are extracted from skins and seeds during fermentation and are essential because of their astringent properties; 2) Quercetin glycosides, including quercetin-3-O-glucoside and quercetin-3-O-glucuronide, which are flavonols extracted from skins, and contribute to wine's velvety astringency (Hufnagel and Hofmann, 2008a; Hufnagel and Hofmann, 2008b); 3) (+)-Catechin and (-)-epicatechin, which are the monomeric fractions of procyanidins, and are associated with wine astringency and bitterness (Kallithraka and Clifford, 2007; Hufnagel and Hofmann, 2008a) and finally, 4) anthocyanins, including malvidin-3-O-glucoside, which are important for wine colour, and also for forming polymeric pigments and more stable pigmented tannins such as pyranoanthocyanins as wine ages (He, 2012).

Examples of A-TEEM finished wine analysis using the machine learning method based on a decision-tree algorithm known as Extreme Gradient Boosting (XGB) (Chen and Guestrin, 2016), including Discrimination Analysis (XGBDA) and regression, have shown the potential to classify wine grape varieties and regions accurately as well as accurately quantify phenolic compounds, respectively (Ranaweera *et al.*, 2021a; Ranaweera *et al.*, 2021b; Ranaweera *et al.*, 2022; Schober *et al.*, 2022). In this study, we investigated the capacity of A-TEEM using multi-block data organisation using both Agglomerative Hierarchical Cluster Analysis (HCA) and XGBDA for grape varietal classification and XGB regression for estimating key chemical compound quality markers using unfermented grape extracts. The efficiency of grape berry phenolic extraction used in this paper is evaluated and compared to earlier finished wine extraction studies to evaluate whether berry extractions lead to more variance and/or lower spike-recovery rates for key phenolic quality-marker compounds. The main sources of variance in the compound model predictions in this paper are evaluated and discussed with respect to their potential impact on field and laboratory implementation for grape inspection and quality assessment.

MATERIALS AND METHODS

1. Chemicals

All solvents, including ethanol, methanol, acetonitrile (chromatography quality), hydrochloric acid and further

chemicals and consumables, including for UV-vis analysis, were purchased from VWR International, USA. Deionised water was obtained by the system Purelab Flex (Elga Labwater, Woodridge, USA). HPLC reference standards, gallic acid mono hydrate (CAS 149-91-7, $\geq 97.5\%$), caffeic acid (CAS: 331-39-5, $\geq 95\%$), (+)-catechin (CAS: 225937-10-0, $\geq 98\%$), (-)-epicatechin (490-46-0, $\geq 99\%$), quercetin hydrate (CAS: 522-12-3, $\geq 90\%$), quercetin dihydrate (6151-25-3, $\geq 98\%$), were purchased from Sigma Aldrich (St. Louis, USA) and malvidin-3-O-glucoside chloride (7228-78-6, $\geq 95\%$) was purchased from Indofine Chemical Company (Hillsborough Township, USA).

2. Grape berry extraction

Fresh grapes collected from different regions of northern California were destemmed to obtain whole berries. Grapes berries were extracted with the modified Iland method (Iland *et al.*, 2004). Briefly, grapes were destemmed on a destemmer, then homogenised with a Vitamix T&G® 2 commercial blender to obtain grape homogenate. One gram of each grape homogenate was acutely weighed in a 15 ml test tube, to which 10 ml 50 % ethanolic solution (pH 2.0 adjusted with 37 % HCl) was added. The test tube was shaken on a shaker (Dual Action Shaker, Labline, Melrose Park, USA) for 1 hour, with motor speed set at 8, to extract colour and phenolics. After extraction, the extract test tube was centrifuged on a centrifuge (Allegra 6, Beckman Coulter, Pasadena, USA) at 4000 RPM for 10 min. One aliquot of the extract supernatant was transferred to an Eppendorf vial by a liquid handling robot (Star, Hamilton Company, Reno, USA) for further centrifugation at 14,000 RMP for 5 min on a microcentrifuge (Microfuge 18, Beckman Coulter, Pasadena, USA). One aliquot of the extract supernatant was transferred to an HPLC vial (Thermo Fisher Scientific, Waltham, USA) for HPLC analysis, and the other aliquot to a 10ml glass vial (Wheaton, Millville, NJ, USA) for Aqualog acquisition. The Aqualog samples were diluted 1:50 in ethanolic aqueous solution (50 % v/v, pH 2.0 adjusted with 37 % HCl) and mixed using the liquid handling robot.

3. Analysis of phenolic compounds by HPLC

Grape extract samples were filtered through Captiva Filter Vials (0.2 μm , Agilent Technologies, Santa Clara, USA) prior to measurement. Analysis was performed by a High-Performance Liquid Chromatography (HPLC) coupled to a Diode Array Detector (DAD). The method was based on a phenolic profile adapted from Peng's method (Peng, 2002). The system was an Agilent 1200 Infinity Series (degasser G1379B, autosampler G1367C, column compartment 1316B, DAD G1315C with standard flow cell, Agilent Technologies, Santa Clara, USA). A precolumn (Zorbax Eclipse XDB C18 4.6 \times 12.5 mm, 5 μm) was coupled to an analytical column Zorbax Eclipse (XDB-C18, 4.6 \times 50 mm, 1.8 μm , both Agilent Technologies Santa Clara, USA), with two mobile phases of water/phosphoric acid (99.5:0.5 v/v) as Phase A and acetonitrile/phosphoric acid (99.5:0.5 v/v) as Phase B. The flow was set to 1.0 ml/min and the column temperature at 50 °C. The binary gradient of the mobile phases was set at the following proportions (percentage of

Phase A): at 0 min 95 %, at 10 min 81 %, 10.25-12.5 min 67 %, 13-15 min 5 % and 15.5-19.5 min 95 %. The injection volume was 4 μl for all samples. The calibration curve was based on 6 points covering the range of 1.0 mg/L to 200 mg/L for all compounds except catechin, with a range of 1.0 mg/L to 1000 mg/L. All compounds were quantitated by their corresponding standards except polymeric tannins, which were expressed in catechin equivalence, quercetin glycosides that were expressed in quercetrin equivalency and total anthocyanins that were expressed in malvidin 3-O-glucoside equivalency. After every twentieth sample, an additional continuous calibration verification (CCV) sample was spiked with a known concentration of (+)-catechin and quercetrin. Recovery rates were calculated from the difference between unspiked and spiked samples divided by spiked concentrations. Recovery rates of 97 %-103 % were achieved for CCV samples. Inter-day (3-day) and intra-day RSD were approximately 8 % and 5 %, respectively, for routine grape extract samples and compounds in this study.

4. A-TEEM analysis

The 1:50 dilution factor was determined to establish a linear concentration relationship according to the Beer–Lambert law for the absorbance and fluorescence signal and IFE correction. Analysis was performed with an Aqualog spectrophotometer (Aqualog-UV-800-C, Horiba Instruments Inc, NJ USA), which simultaneously collected absorbance-transmission (A-T) and fluorescence excitation–emission (EEM) data. The excitation wavelength was set to a range of 240-750 nm with an increment of 5 nm, and the emission wavelength was set to 250-800 nm with an increment of 4.66 nm; the emission axis was interpolated to 5 nm spacing. The measurement was under medium gain and with an integration time of 0.55 sec per single excitation point. Samples contained in 10 ml vials were maintained at 20 °C within and analysed by a Fast-01 HPLC autosampler (HORIBA Instruments Inc., NJ USA). Aqualog data acquisition was carried out with the instrument software, which is based on software Origin (Aqualog V4.3, Horiba Instruments Inc., NJ, USA) and featured automatic spectral pre-processing, including correction of inner filter effects (IFE) and Rayleigh masking (RM) before further data pre-processing and calibration modelling. Solvent blanks were recorded for background signal subtraction, and the EEMs were normalised daily by water Raman scattering units for the specified conditions obtained by the measurement of a standard sealed water cuvette (Starna RM H20, Starna Cells, Atascadero, USA). A-TEEM measurement repeatability was monitored by the relative standard deviation (RSD%) of the absorbance intensity at 520 nm; it was systematically < 2 % and not considered as an impact factor.

5. Grape sampling and experimental plan

In total, 226 grape samples, each sourced from unique vineyards and/or at different maturation stages from a given vineyard for each variety, were analysed in this work with respect to phenolic chemistry regression; the calibration/validation data set included 195/31 samples. For grape variety classification, only varieties with > 10 unique examples were evaluated, which totalled 207 samples; see Table 1 for

TABLE 1. Confusion matrix for Extreme Gradient Boosting Discrimination Analysis of key grape variety extracts. The number of calibration and validation sample files are represented by Cal (n) and Val (n), respectively. Each sample was repeated in two files and all validation sample files were excluded from the calibration data.

Variety	Cal (n)	Val (n)	TPR	TNR	FPR	FNR	Err	P	F1
Cabernet-Sauvignon	64	10	1	1	0	0	0	1	1
Grenache	16	4	1	1	0	0	0	1	1
Malbec	14	4	1	1	0	0	0	1	1
Merlot	56	10	1	1	0	0	0	1	1
Petite Sirah	52	10	1	1	0	0	0	1	1
Pinot noir	54	8	1	1	0	0	0	1	1
Syrah	20	2	1	1	0	0	0	1	1
Teroldego	14	2	1	1	0	0	0	1	1
Zinfandel	62	12	1	1	0	0	0	1	1

TPR: proportion of positive cases that were correctly identified (Sensitivity), = $TP/(TP+FN)$

FPR: proportion of negatives cases that were incorrectly classified as positive, = $FP/(FP+TN)$

TNR: proportion of negatives cases that were classified correctly (Specificity), = $TN/(TN+FP)$

FNR: proportion of positive cases that were incorrectly classified as negative, = $FN/(FN+TP)$

Err: Misclassification error = proportion of samples which were incorrectly classified, = 1-accuracy, = $(FP+FN)/(TP+TN+FP+FN)$

P: Precision, = $TP/(TP+FP)$

F1: F1 Score, = $2*TP/(2*TP+FP+FN)$

calibration/validation splits for each variety. All regression and classification samples were measured in duplicate, and all duplicates remained exclusive to either the calibration or validation data sets to prevent overfitting.

6. Statistical and chemometric analyses

CIE 1931 analysis was performed on the absorbance spectral data, including adjustment by the dilution factor in Aqualog v4.3 software (HORIBA Instruments Inc., USA). The water Raman scattering unit normalised and IFE-corrected three-dimensional EEM data (matrix of 111×103 variables) were unfolded into a two-dimensional coordinate-based format for emission and excitation to intensity (11,433 variables), respectively. Absorbance data (103 variables) was baseline corrected and scaled by the 1:50 dilution factor. The unfolded EEM and absorbance data variables were concatenated as the final multi-block data (11,536 variables). The multi-block spectral data (X block) were regressed against the reference chemistry concentration tables (Y block) to obtain each individual compound prediction model. Regression used the Extreme Gradient Boosting (XGB) algorithm, and all X- and Y-block data pre-processing and XGB model settings are listed in Supplementary Table 1 (Table S1). Table S1 also lists the XGBDA and Agglomerative Hierarchical Cluster Analysis (HCA) pre-processing and model configurations. Both XGBDA and regression models implemented X-block compression using a Partial Least Squares (PLS) model with an optimal number of Latent Variables (LVs) to improve sensitivity and diminish the risk of overfitting. All pre-processing, clustering, regression and classification modelling were carried out in Eigenvector Solo v. 8.9.2 or v. 9.0 (Eigenvector Inc., USA).

RESULTS

Figure 1 shows typical example absorbance profiles (A) and CIE 1931 x-y colour coordinates (B) from A-TEEM samples for Cabernet-Sauvignon, Merlot, Pinot noir and Zinfandel grapes. The absorbance profiles all exhibit the peak signal below 250 nm, a major peak around 280 nm attributed to phenolic and flavan-3-ols compounds, a minor shoulder in the 300–400 nm range associated mostly with flavonols and a second major peak around 520 nm associated with anthocyanins. There were slight differences noted among the different varieties in Panel A that were associated with the different colour mapping coordinates shown in Panel B. Importantly, although these typical examples were clearly unique, the varietal differences among all samples in this study were not clearly discerned solely by the absorbance profiles or CIE coordinates when plotted together (not shown); hence additional evaluation of the EEM data is pursued below.

Figure 2 shows the EEM contour plots for the same typical samples as shown in Figure 1, and within the peak-normalised scaling shown, each variety exhibited a unique fingerprint. Cabernet-Sauvignon exhibited the strongest peak intensity (8.1), followed by Pinot noir (6.7), Zinfandel (6.3) and Merlot (5.24). All four varieties exhibited two contour peaks with major excitation/emission coordinates at $< 240/315$ nm and $280/315$ nm. Interestingly, both the Merlot and Pinot noir also exhibited distinct minor contour peaks with excitation/emission coordinates at $260/375$ nm that were much less pronounced in the Cabernet-Sauvignon and Zinfandel varieties. Supplementary Figure S1 shows the same EEM data as Figure 2 scaled to a peak value of 1 for all four

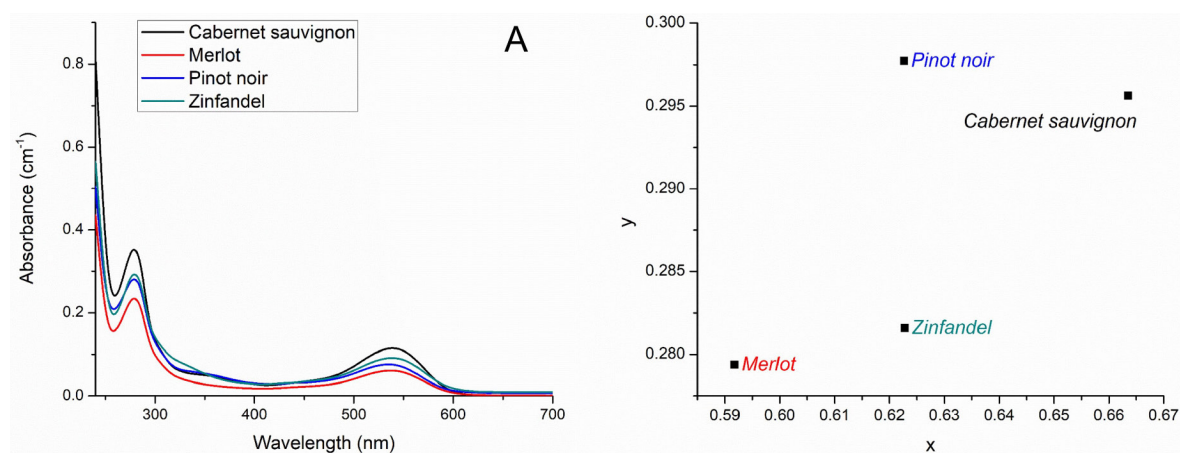


FIGURE 1. Typical Absorbance spectra (A) and CIE 1931 x,y coordinate indices (B) for Cabernet-Sauvignon, Merlot, Pinot noir and Zinfandel grape extract samples. The data in Panel A represent the extracts diluted 50 fold in 50 % EtOH pH 2 solvent whereas the CIE indices in Panel B were adjusted for the dilution factor.

varieties to illustrate that multiple minor contours, including the anthocyanin compounds in the red, can also be observed beyond the limited scaling shown in Figure 2 to distinguish the four varieties further visually.

As with the absorbance and colour profiles in Figure 1, the EEM contours in Figure 2 and Figure S1 can only elicit a limited amount of varietal information by simple visual inspection of one sample at a time. Therefore, we analysed the complete sample set first by Agglomerative Hierarchical Cluster Analysis (HCA) in Figure 3; the results from Figure 3 were evaluated from the concatenated multi-block absorbance and unfolded EEM variables for the combined calibration

and validation data. Here the dendrogram represents the variance-weighted distribution between clusters for the nine varieties identified in the legend. Each variety was exclusively clustered together to support that the A-TEEM multi-block data set contained enough unique chemical fingerprint information to distinguish the different varieties clearly. The apparent varietal linkages clearly show that Pinot noir is the most unique, with Cabernet-Sauvignon initiating the second tier and exhibiting relatively closer linkages to all the other varieties. Supplemental Figure S2 shows that the mean sum (A) and normalised (B) distributions of the key quality marker compound concentrations were uniquely

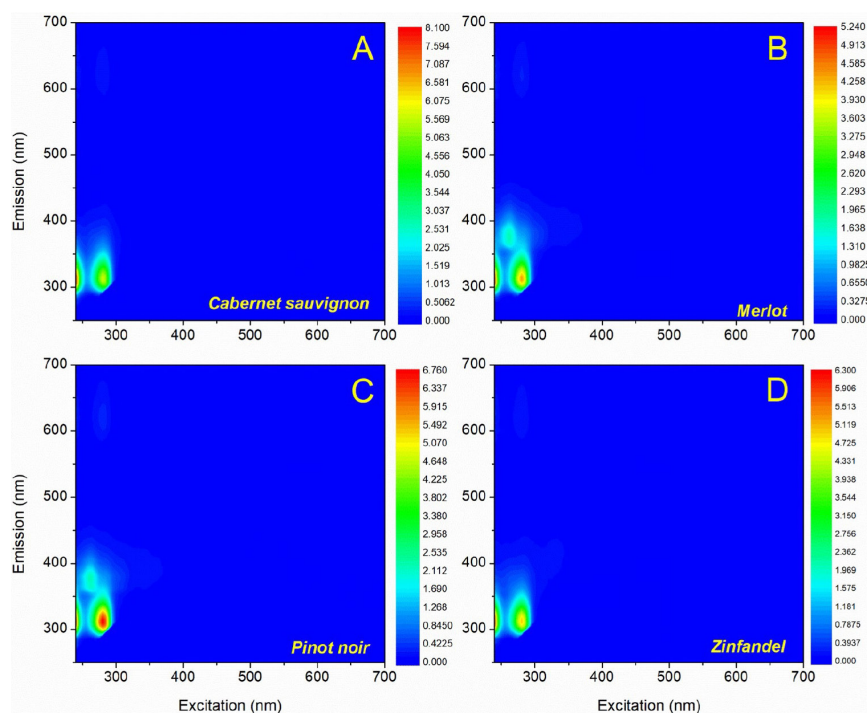


FIGURE 2. Typical fluorescence Excitation-Emission matrix contour plots for the same Cabernet-Sauvignon (A), Merlot (B), Pinot noir (C) and Zinfandel (D) grape extract samples shown in Figure 1 each scaled to the peak EEM contour values.

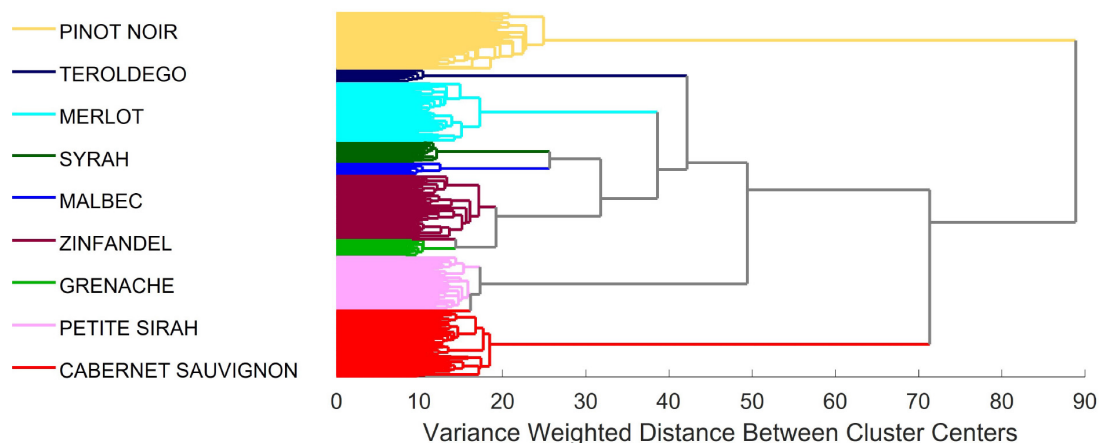


FIGURE 3. Agglomerative Hierarchical Cluster Analysis (HCA) dendrogram for the multi-block A-TEEM data representing the varieties shown in the legend. The data included both the calibration ($n = 390$) and validation ($n = 62$) file sets.

distributed among the tested varieties. The data in Figures 3 and S2 are consistent with other published observations that Pinot noir exhibits a unique phenolic composition, being notably enriched in flavan-3-ols, including catechin and epicatechin, and with generally reduced total anthocyanins (Cliff *et al.*, 2007; Rouxinol *et al.*, 2022; Urcan *et al.*, 2016).

To further investigate the A-TEEM's capacity to classify the varieties in a quantitative manner, we subjected the same data shown in Figure 3 to XGBDA with the results shown in Table 1. Using the methods described above, Table 1 reports no errors in the identification of the calibration set (not shown) or the validation data set for each of the nine varieties analysed as indicated by True Positive Proportion (TPR) and True Negative Proportion (TNR) values of 1 and False Positive Proportion (FPR) and False Negative Proportion (FNR) values of 0 which correspond to Error (Err) values of 0, and Probability (P) and F1 values of 1, respectively. Hence, the HCA (Figure 3) and XGBDA (Table 1) data both support that the A-TEEM accurately qualified the nine grape varieties in the tested samples.

The quantification of key quality chemical markers for grapes was another key topic of investigation for the A-TEEM multi-block model approach. In Figure 4, we show the XGB regression prediction data set values for (A) Polymeric Tannins, (B) Malvidin-3-Glucoside and Total Anthocyanins, (C) Flavan-3-ols, Catechin and Epicatechin and (D) Quercetin Glycosides. Polymeric Tannins ranged from 150 to 350 mg/L, Total anthocyanins ranged above 185 mg/L, Catechin ranged up to around 55 mg/L and Quercetin Glycosides were observed up to around 14 mg/L. For all six target compounds, the predicted linear and the quantitative performance prediction statistics are described in Table 2. Table 2 data are sorted from top to bottom by the maximum range for each compound. The R^2P was > 0.917 for all compounds with an average of 0.9411 ± 0.0247 which equates to an average REP of $< 5.9\%$. For all models in Table 1, the R^2 for calibration was greater than or equal to the R^2P . It is notable that the target for an ideal secondary method for this primary HPLC method should be around 5

to 8%, given the previously described typical spike recovery and Inter- and Intra-day repeatability statistics (RSD) of this HPLC method with grape berry extracts.

DISCUSSION

Both the A-TEEM varietal classification and phenolic regression results with unfermented grape extracts presented here are consistent with previously published work with finished wines (Ranaweera, 2021a; Ranaweera, 2021b; Schober *et al.*, 2022). In contrast, however, this work is a special industrially focused application exclusively with vineyard and maturation-stage specific commercial grape extracts with significance for rapidly and accurately establishing both grape quality and authenticating varietal characteristics. Both the varietal classification and phenolic quantification capabilities are important for advising vineyard management and berry harvesting and sorting for targeted product quality control and formulation.

The combined use of the multi-block data organisation and use of the XGB algorithm in this paper yielded satisfactory regression results, which were consistent with the established performance of the primary HPLC method with respect to observed inter- (8%) and intra-day (5%) RSD associated with recovery of the measured compounds from raw berry extracts. Importantly, the A-TEEM repeatability for any given sample was measured to be $< 2\%$ Relative Standard Deviation (RSD) as evidence that the mean 5.89% REP for the phenolics in Table 2 was mostly associated with error propagation associated with the overall berry extraction and HPLC methodology. It is also notable that this reference HPLC method required approximately 20 min, while the A-TEEM analysis was limited to less than one. Importantly, the REP ($\approx 6\%$) for the overall phenolic analysis of finished wines in Schober *et al.* (2022) compared quite closely to the phenolic REP ($\approx 5.89\%$) in this grape extract study.

Here the significant resolution of each of the nine grape varieties with HCA and XGBDA supports the potential for the A-TEEM method to be employed for authentication of

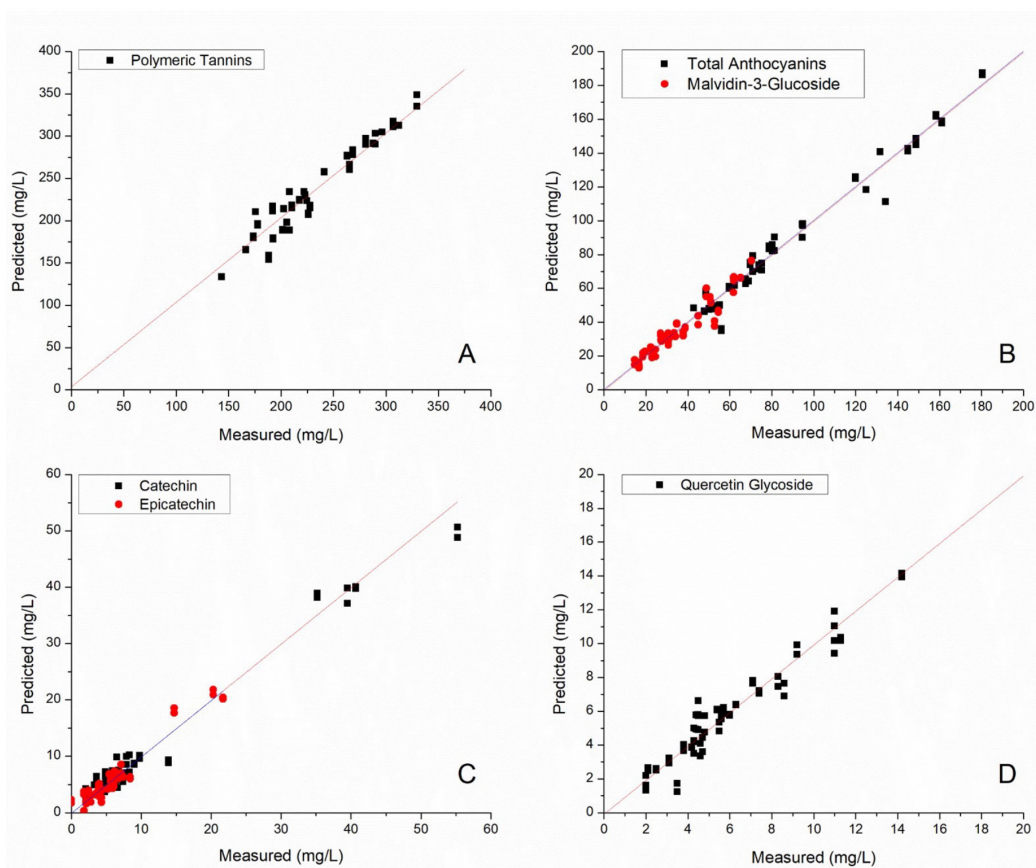


FIGURE 4. Extreme Gradient Boosting regression plots for the prediction data set for key phenolic and anthocyanin compounds. All lines were constrained to a slope of unity and intercept of 0 mg/L. The compound identities are listed in the legends. The regression statistics are contained in Table 2.

grape lots, especially for labelling of pure varietal products as well as for blending as had been shown earlier for finished wine products (Ranaweera *et al.*, 2022). It is evident from the unique A-TEEM fingerprint characteristics that the successful discrimination is primarily based on the unique phenolic compositions of each variety. It is also important to consider the samples evaluated here for each variety, and samples were sourced from different vineyards and harvest

dates (all from the same year) which lends credence to the concept that the varietal discrimination is primarily based on the genetic characteristics as opposed to being more strongly associated with vineyard or growth (maturation). This is especially evident with the previously referenced uniqueness of the Pinot noir variety that consistently exhibits much higher concentrations of the flavan-3-ols Catechin and Epicatechin as well as other significant differences in the

TABLE 2. Extreme Gradient Boosting Regression analysis statistics for key quality-associated phenolic and anthocyanin parameters including the R², the Relative Error of Prediction (REP), the Root Mean Square Error or Standard Deviation (RMSE or SD) and the maximum concentration range (Max Range) for the prediction set. The calibration/validation sets included 390/62 sample files representing 195/31 samples, respectively. The table is sorted by the Max Range parameter.

Compound/Parameter	R ² P	REP%	RMSE SD (mg/L)	Max Range (mg/L)
Polymeric Tannins	0.9244	7.56	14.38	348.51
Total Anthocyanins	0.9655	3.45	7.58	187.35
Malvidin-3-Glucoside	0.9173	8.27	4.74	76.46
Catechin	0.9783	2.18	1.09	50.61
Epicatechin	0.9316	6.84	1.33	21.81
Quercetin Glycosides	0.9293	7.07	0.79	14.16
Mean	0.9411	5.89		
SD	0.0247	2.47		

polyphenolic profiles (Yang *et al.*, 2009), including lack of acylated anthocyanins (Boss *et al.*, 1996; Mazza *et al.*, 1999), compared to the other varieties in this study.

In conclusion, this study helps to establish further the A-TEEM coupled with multi-block data fusion and machine learning, especially XGB, as an effective tool to monitor incoming grape samples rapidly. This can facilitate analysis faster than HPLC with similar degrees of precision for phenolic concentrations and, more importantly, with the added capacity for rapid, accurate grape variety identification. Future work may extend to the rapid measurement of phenolic compounds in grapes and authentication by the A-TEEM method on a wider geographical scale. Importantly, the individual regression and discrimination models, once calibrated, can both be applied with commercially available software in an automated fashion that is amenable to field or remote-laboratory operations for near real-time monitoring (Schober *et al.*, 2022).

ACKNOWLEDGEMENTS

We thank our colleagues and interns from Winegrowing Research at E&J Gallo Winery who collected and helped to process the grapes for this study. We would also like to show our gratitude to Dr. Nick Dokoozlian, Vice President, E&J Gallo Winery for his support to advance the research in grapes and wine with A-TEEM technology.

REFERENCES

- Alexandre-Tudo, J.L., Nieuwoudt H., Aleixandre, J. L., & du Toit, W. (2017) Chemometric compositional analysis of phenolic compounds in fermenting samples and wines using different infrared spectroscopy techniques. *Talanta*, 176, 526–536. <http://dx.doi.org/10.1016/j.talanta.2017.08.065>
- Boss, P.K., Davies, C., & Robinson, S.P. (1996). Anthocyanin composition and anthocyanin pathway gene expression in grapevine sports differing in berry skin colour. *Australian Journal of Grape and Wine Research*, 2, 163–170. <https://doi.org/10.1111/j.1755-0238.1996.tb00104.x>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, USA. <https://doi.org/10.1145/2939672.2939785>
- Cleary, M., Chong, H., Ebisuda, N., Dokoozlian, N., Loscos, N., Pan, B., Santino, D., Sui, Q., & Yonker, C. (2015). Objective Chemical Measures of Grape Quality. *ACS Symposium Series*, 120, 365–378. <https://doi.org/10.1021/bk-2015-1203.ch023>
- Cliff, M.A., King, M.C., & Schlosser, J. (2007) Anthocyanin, phenolic composition, colour measurement and sensory analysis of BC commercial red wines. *Food Research International*, 40, 92–100. <https://doi.org/10.1016/j.foodres.2006.08.002>
- de Villiers, A., Majek, P., Lynen, F., Crouch, A., Lauer, H., & Sandra, P. (2005) Classification of South African red and white wines according to grape variety based on the non-coloured phenolic content. *Eur Food Res Technol*, 221, 520–528. <https://doi.org/10.1007/s00217-005-1169-5>
- Geană, E-I., Ciucure, C.T., Apetrei, C., & Artem, V. (2019). Application of Spectroscopic UV-Vis and FT-IR Screening Techniques Coupled with Multivariate Statistical Analysis for Red Wine Authentication: Varietal and Vintage Year Discrimination. *Molecules*, 24(22), 4166. <https://doi.org/10.3390/molecules24224166>
- Gilmore, A. (2014). How to Collect National Institute of Standards and Technology (NIST) Traceable Fluorescence Excitation and Emission Spectra. In Y. Engelborghs & A. J. W. G. Visser (Eds.), *Fluorescence Spectroscopy and Microscopy: Methods and Protocols* (pp. 371–417). <https://doi.org/10.1007/978-1-62703-649-8>
- Gilmore, A., & Tong, X. (2019) System and Method for Fluorescence and Absorbance Analysis. United States Patent No. US8901513B2
- Harbertson, J. F., & Spayd, S. (2006). Measuring Phenolics in the Winery. *American Journal of Enology and Viticulture*, 3(57), 280–288.
- Harbertson, J. F., Picciotto, E. A., & Adams, D. O. (2003). Measurement of Polymeric Pigments in Grape Berry Extracts and Wines Using a Protein Precipitation Assay Combined with Bisulfite Bleaching. *American Journal of Enology and Viticulture*, 54(4), 301–306.
- He, F. (2012) Anthocyanins and their variation in red wines I. Monomeric anthocyanins and their color expression. *Molecules*, 17(2), 1571–1601. <https://doi.org/10.3390/molecules17021571>
- Hufnagel, J.C., & Hofmann, T. (2008a). Orosensory-Directed Identification of Astringent Mouthfeel and Bitter-Tasting Compounds in Red Wine. *Journal of Agricultural and Food Chemistry*, 56(3), 1376–1386. <https://doi.org/10.1021/jf073031n>
- Hufnagel, J. C., & Hofmann, T. (2008b). Quantitative Reconstruction Of the Nonvolatile Sensometabolome of a Red Wine. *Journal of Agricultural and Food Chemistry*, 56, 9190–9199. <https://doi.org/10.1021/jf801742w>
- Iland, P., Bruer, N., Edwards, G., Weeks, S., & Wilkes, E. (2004) *Chemical Analysis of Grapes and Wine: Techniques and Concepts*; Patrick Iland Wine Promotions PTY Ltd.: Campbelltown, Australia.
- Kallithraka, S., & Clifford, M.N. (2007). Evaluation of bitterness and astringency of (+)-catechin and (-)-epicatechin in red wine and in model solution. *Journal of Sensory Studies*, 12(1), 25–27. <https://doi.org/10.1111/j.1745-459X.1997.tb00051.x>
- Mazza, G., Fukumoto, L., Delaquis, P., Girard, B., & Ewert, B. (1999) Anthocyanins, phenolics, and color of Cabernet Franc, Merlot, and Pinot noir wines from British Columbia. *J. Agric. Food Chem.* 47, 4009–4017. <https://doi.org/10.1021/jf990449f>
- Miramont, C., Jourdes, M., & Teissedre, P-L. (2020) Development of UV-vis and FTIR Partial Least Squares models: comparison and combination of two spectroscopy techniques with chemometrics for polyphenols quantification in red wine. *Oeno One*, 4, 779–792. <https://doi.org/10.20870/oeno-one.2020.54.4.3731>
- Niimi, J., Tomic, O., Næs, T., Bastian, S. E. P., Jeffery, D. W., Nicholson, E. L., Maffei, S. M., & Boss, P. K. (2020). Objective measures of grape quality: From Cabernet-Sauvignon grape composition to wine sensory characteristics. *Food Science and Technology*, 123, 109105. <https://doi.org/10.1016/j.lwt.2020.109105>
- Peng, Z, Iland, P. G., Oberholster, A., Sefton, M. A., & Waters, E. J. (2002). Analysis of pigmented polymers in red wine by reverse phase HPLC. *Australian Journal of Grape and Wine Research*, 8, 70–75. <https://doi.org/10.1111/j.1755-0238.2002.tb00213.x>
- Ranaweera, K. R. R., Capone, D.L., Bastian, S.E.P., Cozzolino, D., & Jeffery, D.W. (2021a). A Review of Wine Authentication Using Spectroscopic Approaches in Combination with Chemometrics. *Molecules*, 26, 4334. <https://doi.org/10.3390/molecules26144334>
- Ranaweera, K. R. R., Gilmore, A. M., Capone, D. L., Bastian, S. E. P., & Jeffery, D. W. (2021b). Spectrofluorometric analysis combined

with machine learning for geographical and varietal authentication, and prediction of phenolic compound concentrations in red wine. *Food Chemistry*, 361, 130149. <https://doi.org/10.1016/j.foodchem.2021.130149>

Ranaweera, K.R.R., Gilmore, A.M., Bastian, S.E.P., Capone, D.L., & Jeffery, D.W. (2022). Spectrofluorometric analysis to trace the molecular fingerprint of wine during the winemaking process and recognise the blending percentage of different varietal wines. *OENO One*, 56(1), 189-196. <https://doi.org/10.20870/oeno-one.2022.56.1.4904>

Rouxinol, M.I., Martins, M.R., Murta, G.C., Barroso, J.M., & Rato, A.E. (2022). Quality Assessment of Red Wine Grapes through NIR Spectroscopy. *Agronomy*, 12(3), 637. <https://doi.org/10.3390/agronomy12030637>

Schober, D., Gilmore, A., Chen, L., Zincker, J., & Gonzalez, A. (2022). Determination of Cabernet-Sauvignon Wine Quality Parameters in

Chile by Absorbance-Transmission and Fluorescence Excitation Emission MATRIX (A-Teem) Spectroscopy. *Food Chemistry*, 133101. <https://doi.org/10.1016/j.foodchem.2022.133101>

Sen, I., Ozturk, B., Tokatli, F., & Ozen, B. (2016). Combination of visible and mid-infrared spectra for the prediction of chemical parameters of wines. *Talanta*, 161, 130–137. <http://dx.doi.org/10.1016/j.talanta.2016.08.057>

Urcan, D.E., Lung, M-L., Giacosa, S., Torchio, F., Ferrandino, A., Vincenzi, S., Segade, S-R., Pop, N., & Rolle, L. (2016). Phenolic Substances, Flavor Compounds, and Textural Properties of Three Native Romanian Wine Grape Varieties. *International Journal of Food Properties*, 19, 76–98. <https://doi.org/10.1080/10942912.2015.1019626>

Yang, J., Martinson, T.E., & Liu, R.H. (2009). Phytochemical profiles and antioxidant activities of wine grapes. *Food Chemistry*, 116, 332–339. <https://doi.org/10.1016/j.foodchem.2009.02.021>