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

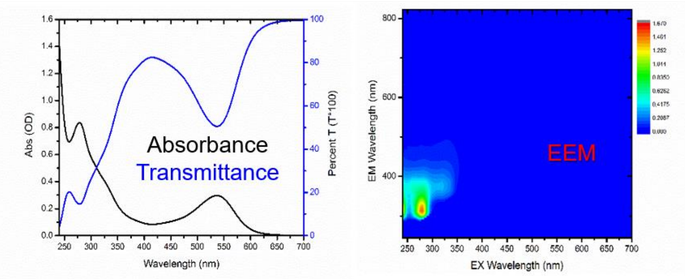
Adam Gilmore[1*], Qiang Sui[2], Bryant Blair[2], Bruce Pan[2]
[1]HORIBA Instruments Incorporated, Piscataway NJ 08854 USA,[2]E & J Gallo Winery, Modesto CA 95354 USA
*corresponding author: adam.gilmore@horiba.com

ABSTRACT

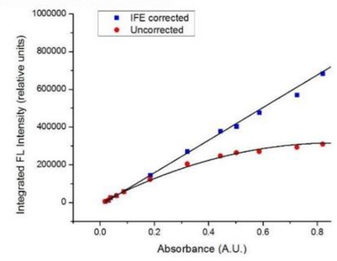
Rapid, accurate quantification of grape berry phenolics, anthocyanins and tannins, and discrimination of grape varieties are both important for effective quality control of harvesting and initial processing for winemaking. Current reference technologies including High Performance Liquid Chromatography (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), Petit syrah (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 regressions were calibrated (n = 390) using Extreme Gradient Boost (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 (R^2P) of the Prediction data (n = 62). The regression results yielded an average Relative Error of Prediction (REP) of 5.89 ± 2.47 % and R^2P 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). 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.

A-TEEM: The Synergy of Absorbance and Fluorescence

***Patented* Instrument Method:**
Absorbance: measures all colored species independent of fluorescence
Transmittance: measures color and chromaticity
EEM: measures excitation and emission of all fluorescent components



Inner-Filter Effect Correction:
Expands linear concentration range for fluorescence
High sensitivity
Can resolve compounds in ppt to ppb range from high background (ppm) matrices.
Compound libraries and model databases independent of concentration



HORIBA Scientific © 2022 HORIBA, Ltd. All rights reserved.

Unique Hierarchical Clustering for Grape Varieties

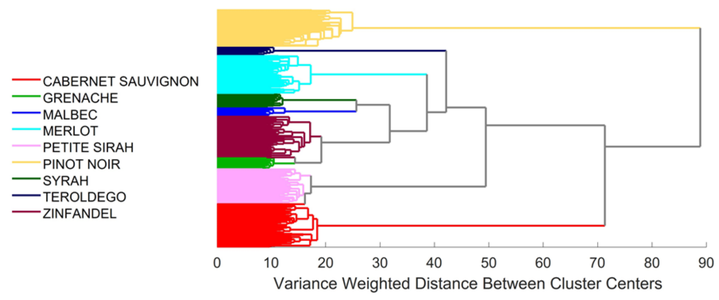


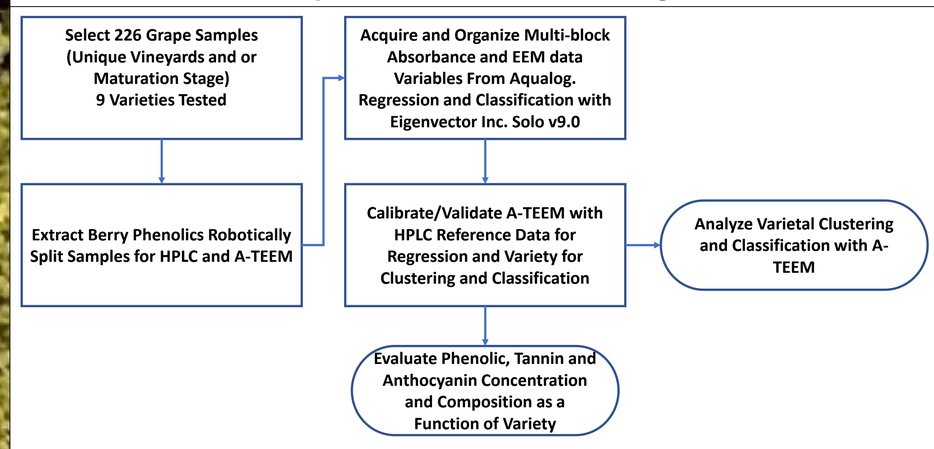
Figure 3. Agglomerative Hierarchical Cluster Analysis (HCA) dendrogram for the multi-block A-TEEM data representing the varieties shown in the legend. The data included both the calibration (n=390) and validation (n=62) file sets.

Accurate Regression Statistics for Phenolic Analysis

Table 2. Extreme Gradient Boost Regression analysis statistics for key quality-associated phenolic and anthocyanin parameters including the R^2, the Relative Error of Prediction (REP), The Root Mean Square Error or Standard Deviation (RMSE 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^2P	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		

Experimental Methods and Design



Accurate Classification of Grape Varieties

Table 1. Confusion matrix for Extreme Gradient Boost 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. *See Footnote 1 for definitions of column parameters, TPR, TNR, FPR, FNR, Err, P and F1.

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
Petit 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

*Footnote of Table 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)

Unique Predicted Phenolic Composition for Grape Varieties

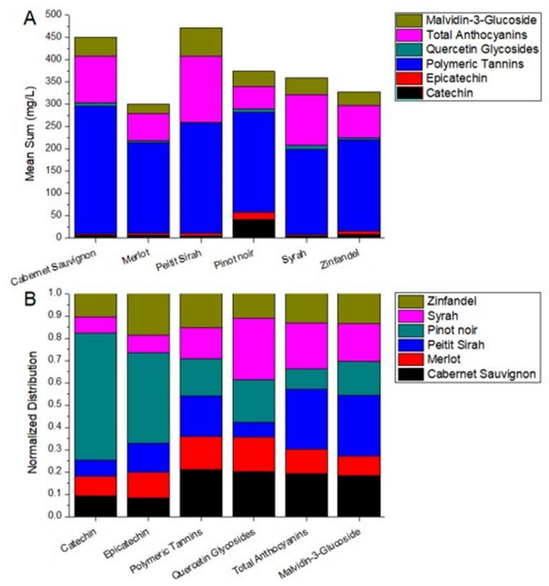


Figure 5. Panel A shows the Mean Sum distribution comprising the six quality compound marker concentrations listed in the legend plotted as the averages from the validation data set for each of the six varieties tested on the x-axis. Panel B shows the same data as Panel A organized as the normalized distribution of each of the six quality marker compounds on the x-axis for the six varieties listed in the legend.

Unique Absorbance Spectra and Chromaticity of Grape Varieties

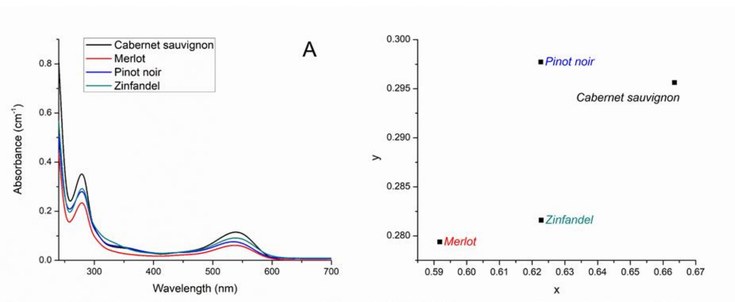


Figure 1. Typical Absorbance spectra (A) and CIE 1931 x,y coordinate indices (B) for *Cabernet sauvignon, Merlot, Pinot noir and Zinfandel* grape extract samples. The data in Panel A represent the extracts diluted 50 fold in 50% EtOH pH 2 solvent whereas the CIE indices in Panel B were adjusted for the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

Unique EEM Fingerprints for Grape Varieties

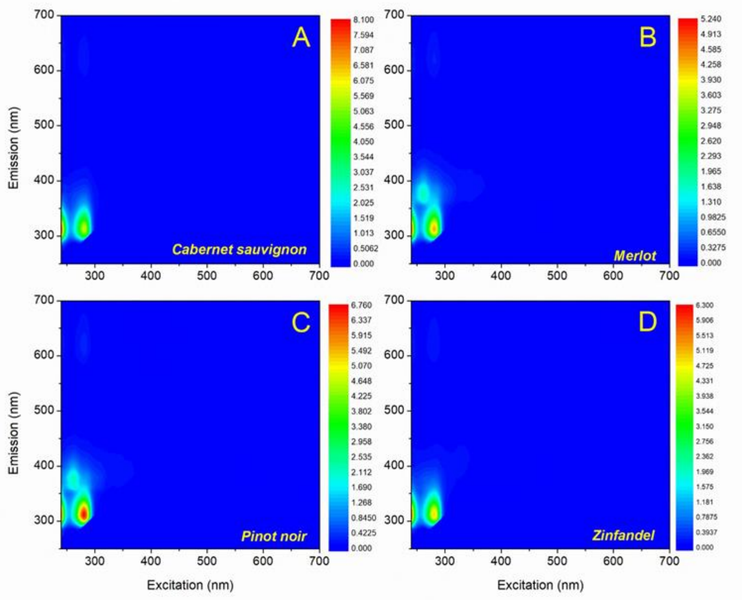


Figure 2. Typical fluorescence Excitation-Emission matrix contour plots for the same *Cabernet sauvignon (A), Merlot (B), Pinot noir (C) and Zinfandel (D)* grape extract samples shown in Figure 1 each scaled to the peak EEM contour values.

Linear Prediction of Flavan-3-ols, Flavonols, Tannins and Anthocyanins

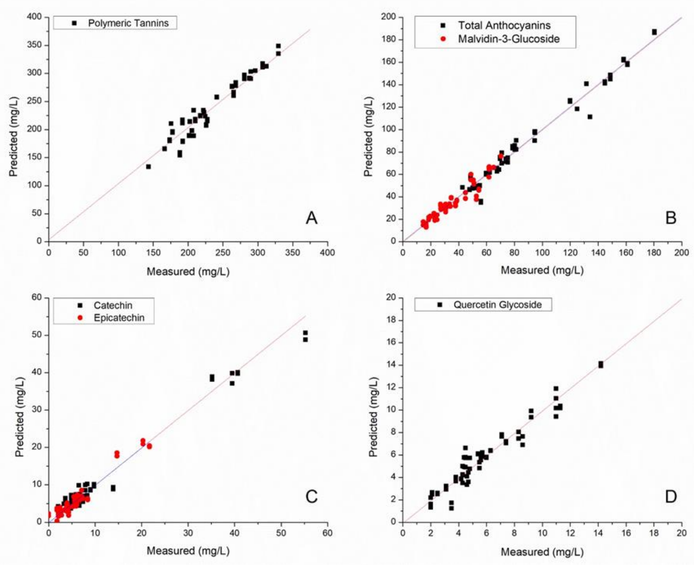


Figure 4. Extreme Gradient Boost regression plots for the prediction data set for key phenolic and anthocyanin compounds. All lines were constrained to a slope of unity and intercept of 0 mg/L. The compound identities are listed in the legends. The regression statistics are contained in Table 2.

Conclusions

- The A-TEEM accurately classified 9 varieties of grape juice extracts.
- The A-TEEM accurately predicted flavan-3-ol, flavonol, tannin and anthocyanin concentrations with a relative prediction error (5.89%) that was consistent with the reference HPLC repeatability (intraday 5% and inter-3-day 8%).
- The A-TEEM acquisition scan time was less than 1 min.
- The A-TEEM qualifies as a suitable method for grape quality evaluation and varietal authentication.

E&J Gallo Winery

HORIBA Explore the future