The quantification of red wine phenolics using fluorescence spectroscopy with chemometrics

by

Isabel Anne dos Santos

Thesis presented in partial fulfilment of the requirements for the degree of
Master of Agricultural Sciences

at
Stellenbosch University
Department of Viticulture and Oenology, Faculty of AgriSciences

Supervisor: Prof. Wessel du Toit
Co-supervisor: Dr. José Luis Aleixandre-Tudo and Dr. Gurthwin Bosman

March 2021

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2021

Summary

The organoleptic and perceived quality characteristics of red wine are largely influenced by important phenolic compounds extracted throughout fermentation from the grape berry to the final wine matrix. These complex secondary metabolites have resulted in numerous equally complex analysis methods, the implementation of which are yet to form part of routine phenolic analysis during winemaking. In this study, front-face fluorescence spectroscopy was investigated for its suitability in quantifying phenolic parameters of unaltered samples and the subsequent implications for non-invasive analysis throughout fermentation.

A front-face accessory and fluorescence spectrophotometer were successfully optimised in order to analyse samples directly, eliminating the need for sample dilution as with conventional fluorescence spectroscopy. A diverse dataset comprising 289 fermenting musts and wine were analysed using the optimised fluorescence protocol and the most commonly used UV-Vis spectrophotometric methods for the following phenolic parameters; total phenolics, total condensed tannins, total anthocyanins, colour density and polymeric pigments. Different statistical analysis methods were explored for their suitability in model development, specifically Parallel Factor Analysis (PARAFAC) and a gradient boosting machine learning algorithm (XGBoost). Subsequent to the investigation of the most optimal chemometric method, a machine learning pipeline was generated to develop accurate regression models per phenolic parameter. Successful models were obtained for total phenolics, total condensed tannins and total anthocyanins while polymeric pigments and colour density require further investigation and refinement. Following model development and optimisation, an external validation experiment monitoring a Cabernet Sauvignon fermentation was used to examine prediction accuracy under fermentation conditions, specifically investigating the effect of carbon dioxide and must turbidity. No effect of sample preparation treatment was found and the potential for analysing unaltered samples directly during fermentation was possible.

Fluorescent properties of fermenting musts and wines were explored and the responsible spectral regions of interest tentatively identified. Differences in fluorescence between musts and wines were found and upon closer inspection, unique changes were monitored and identified throughout fermentation using the Cabernet Sauvignon experiment. The unique fluorescent profiles of wines is widely accepted, and the classification of South African red wine cultivars was successfully conducted using Neighbourhood Component Analysis (NCA). These results may have beneficial implications for authentication and quality control by industry bodies.

Overall, front-face fluorescence spectroscopy holds several advantages including it being non-invasive, user-friendly, relatively economical, rapid and accurate, and thus presents itself as a

promising alternative to the current phenolic analysis methods with the added benefit of direct phenolic analysis throughout red wine fermentation. The potential for implementation within on-line automated systems or portable optical devices may be of interest to producers and allow for monitoring of phenolic content and extraction directly from the fermentation vessel throughout red wine production.

Biographical sketch

Isabel dos Santos was born on the 7th of December 1996 in Boksburg, where she grew up and matriculated in 2014 from St. Dominic's School for Girls. Following her heritage from the island of Madeira, she went on to study BScAgric Viticulture and Oenology at Stellenbosch University where she graduated top of her class in 2018. Her postgraduate studies commenced the following year and she was able to complete her Master's in Oenology.

Acknowledgements

I wish to express my sincere gratitude and appreciation to the following persons and institutions:

- My supervisor, **Prof. Wessel du Toit**, for all your help, guidance and mentorship throughout this journey.

- My co-supervisor, **Dr. José Luis Aleixandre-Tudo**, for all your wisdom, critical thinking, attentiveness and constant availability to help.

- My co-supervisor, **Dr. Gurthwin Bosman**, for your guidance in the domain of physics, constant open-door policy and your overall positive demeanour that kept me motivated throughout.

- **Dr. Federico Marini** from the University of Rome, for your time, effort and help with PARAFAC all the way from Italy.

- My loving mom, **Rita dos Santos**, for your constant motivation, sacrifices and strength that have made me who I am today, my big brothers, **Paul** and **Edward**, for being the best role models I could ever have, **Benjamin Nicol** for your unwavering love and support and all my friends who have been with me throughout this journey.

- My friend, **Ayesha Shaikh**, for joining me along this Masters journey and constantly keeping me on my toes with your inspiring work ethic and positivity.

- My brother, **Paul dos Santos**, for your incredible brain and data science skills, helping me to better navigate the world of machine learning.

- My guardian angels, **Manuel dos Santos** and **Adelina de Abreu**. Thank you for giving me strength, I know you are beaming down at this achievement.

- A special thank you to the **PA and Alize Malan Trust** for your generous financial support that allowed me to pursue this Masters as well as **Winetech** for the research funding that made these studies possible.

Preface

This thesis is presented as a compilation of 5 chapters. Each chapter is introduced separately and is written according to the style of Harvard citation.

Chapter 1 **General Introduction and project aims**

Chapter 2 **Literature review**
Red wine phenolics and their quantification methods

Chapter 3 **Research results**
The direct quantification of red wine phenolics using fluorescence spectroscopy with chemometrics

Chapter 4 **Research results**
Non-invasive fluorescence spectroscopy to quantify phenolic content under real-time fermentation conditions

Chapter 5 **General discussion and conclusions**

Table of Contents

Chapter 1

Introduction and project aims

1.1 INTRODUCTION

Grape berry growth and development is characterised by a three stage, double sigmoidal curve, during which the evolution of numerous solutes and metabolites occurs (Kennedy, 2002; Garrido and Borges, 2013). While primary metabolites are essential for plant growth and survival, the roles secondary metabolites play in wine aroma and taste attributes are of great importance due to their influence on perceived quality by the consumer. Phenolic compounds are a diverse and complex group of secondary metabolites found in grapes and wine and can be categorised according to two groups; flavonoids and non-flavonoids (Garrido and Borges, 2013). These compounds influence colour and mouthfeel properties, such as astringency and bitterness, as well as the ageing potential of wines (Aleixandre-Tudo *et al*., 2017) and an understanding of appropriate winemaking practices to implement at various stages or phenolic levels may hugely impact the final wine phenolic profile (Sacchi *et al*., 2005).

Currently, phenolic analysis methods do not form part of routine wine analysis as a result of several drawbacks, including the need for trained personnel, expensive reagents and equipment, and lengthy sample preparation and analysis time (Harbertson and Spayd, 2006). Due to the complexity of phenolic compounds, numerous analysis methods have been developed in order to extract the most relevant phenolic information by reducing complex phenolic chemistry to the measurement of a number of parameters, the focus of which having been on total phenolics, tannins, anthocyanins, polymeric pigments and colour density. The simple spectrophotometric methods most often used are UV-Vis based and rely on the spectral properties of the aromatic ring present in all phenolic compounds, allowing for differentiation between phenolic groups according to characteristic wavelength peaks (Harbertson and Spayd, 2006; Aleixandre-Tudo *et al*., 2017). Alternatives such as High-performance liquid chromatography (HPLC) are extremely sensitive and accurate but are rarely implemented outside of research applications (Aleixandre-Tudo *et al*., 2018). As the need for rapid, accurate, user-friendly and cost effective methods increases, the applicability of spectroscopy coupled with multivariate statistical analysis (chemometrics) presents itself as a suitable option. UV-Visible and infrared spectroscopies with chemometrics have previously been studied and deemed suitable in the analysis of phenolic compounds (Romera-fernández *et al*., 2012; Dambergs *et al*., 2012; Daniel, 2015).

Fluorescence spectroscopy is non-destructive, user-friendly, cost effective and highly sensitive when compared to other spectrophotometric methods and its benefits have deemed it a useful method of analysis in the authentication and quality control of many food science

disciplines (Strasburg and Ludescher, 1995; Airado-Rodríguez *et al.*, 2011; Karoui and Blecker, 2011). Fluorescent properties of the wine matrix have been studied and both the qualitative and quantitative analysis of the responsible components is of increasing interest, with two general spectral regions having been identified within the wine excitation-emission matrix (Airado-Rodŕiguez *et al.*, 2009; Airado-Rodríguez *et al.*, 2011). Understanding the principles and limitations of the instrumentation for fluorescence spectroscopy is an important aspect in obtaining the highest quality data possible. Sample geometry is one of the most important considerations of fluorescence analysis, with the conventional right angle technique able to measure only clear or diluted samples and therefore unable to analyse samples in their truest form (Airado-Rodríguez *et al.*, 2011). Front-face analysis developed by Parker (1968) overcomes this problem by changing the angle of incidence. As a result, the complex wine matrix is kept intact and the subsequent analysis of unaltered samples may be of significant benefit in non-invasive analysis during and throughout fermentation.

Chemometrics allows for the interpretation and decomposition of complex datasets while analysing multiple variables simultaneously in a considerably reduced analysis time (Aleixandre-Tudo *et al.*, 2017). Combining spectroscopy with chemometrics has the potential to create accurate and robust regression models capable of quantifying red wine phenolics for real-time winemaking decisions. Despite the research into wine fluorescence, gaps remain in the quantification of general parameters such as total condensed tannins, which may be of more use to producers compared to those of individual chemical compounds, such as the quantification of catechin or epicatechin by Cabrera-Bañegil *et al*. (2019). The majority of wine fluorescence research has focused on qualitative applications such as classification and discrimination according to cultivar, wine style or appellation (Letort et al., 2006; Airado-Rodríguez *et al*., 2011), while non-invasive quantification of phenolic compounds requires more investigation.

1.2 AIMS AND OBJECTIVES

Considering the lack of phenolic analysis conducted within the South African wine industry, the main aim of this research was to determine the suitability of fluorescence spectroscopy for the quantification of phenolic content of unaltered red wine samples and the subsequent implications for non-invasive analysis during fermentation. The five phenolic parameters of interest included total phenolics, total condensed tannins, total anthocyanins, colour density and polymeric pigments. To successfully achieve this goal, several objectives had to be met, including:

I. The modification of conventional fluorescence spectroscopy to front-face fluorescence spectroscopy and the calibration and optimisation of this protocol.

II. The analysis of red wine cultivars during fermentation and of finished wines using fluorescence spectroscopy and conventional UV-Vis spectrophotometric methods.

III. The exploration of chemometrics and development of accurate prediction models.

IV. Investigating the influence of sample composition on fluorescence analysis, specifically with regards to potential interference from turbidity and carbon dioxide produced during fermentation, and the subsequent implications for real-time and on-site applications.

1.3 REFERENCES

Airado-Rodŕiguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2009. Usefulness of fluorescence excitation-emission matrices in combination with parafac, as fingerprints of red wines. Journal of Agricultural and Food Chemistry, 57(5), 1711–1720.

Airado-Rodríguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2011. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. Journal of Food Composition and Analysis, 24(2), 257–264.

Aleixandre-Tudo, J. L., Buica, A., Nieuwoudt, H., Aleixandre, J. L. and du Toit, W., 2017. Spectrophotometric analysis of phenolic compounds in grapes and wines. Journal of Agricultural and Food Chemistry, 65(20), 4009–4026.

Aleixandre-Tudo, J. L., Nieuwoudt, H., Olivieri, A., Aleixandre, J. L. and du Toit, W., 2018. Phenolic profiling of grapes, fermenting samples and wines using UV-Visible spectroscopy with chemometrics. Food Control, 85, 11–22.

Cabrera-Bañegil, M., Valdés-Sánchez, E., Moreno, D., Airado-Rodríguez, D. and Durán-Merás, I., 2019. Front-face fluorescence excitation-emission matrices in combination with three-way chemometrics for the discrimination and prediction of phenolic response to vineyard agronomic practices. Food Chemistry, 270, 162–172.

Dambergs, R. G., Mercurio, M. D., Kassara, S., Cozzolino, D. and Smith, P. A., 2012. Rapid measurement of methyl cellulose precipitable tannins using ultraviolet spectroscopy with chemometrics: Application to red wine and inter-laboratory calibration transfer. Applied Spectroscopy, 66(6), 656-664.

Daniel, C., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules, 20(1), 726–737.

Garrido, J. and Borges, F., 2013. Wine and grape polyphenols - A chemical perspective. Food Research International, 54(2), 1844–1858.

Harbertson, J. F. and Spayd, S., 2006. Measuring phenolics in the winery. American Journal of Enology and Viticulture, 57(3), 280-288.

Karoui, R. and Blecker, C., 2011. Fluorescence spectroscopy measurement for quality assessment of food systems — a review. . Food and Bioprocess Technology, 4(3), 364–386.

Kennedy, B., 2002. Understanding grape berry development. Practical Winery and Vineyard, (4), 1-5.

Letort, A., Laguet, A., Lebecque, A. and Serra, J. N., 2006 Investigation of variety, typicality and vintage of French and German wines using front-face fluorescence spectroscopy. Analytica Chimica Acta, 563, 292–299.

Parker, C.A., 1968. Apparatus and experimental methods. In: Parker, C.A. (Ed.), Photoluminescence of Solutions with Applications to Photochemistry and Analytical Chemistry, 128–302.

Romera-fernández, M. Berrueta, L. A., Garmón-lobato, S., Gallo, B., Vicente, F. and Moreda, J. M., 2012. Feasibility study of FT-MIR spectroscopy and PLS-R for the fast determination of anthocyanins in wine. Talanta, 88, 303–310.

Sacchi, K. L., Bisson, L. F. and Adams, D. O., 2005. . A review of the effect of winemaking techniques. American Journal of Enology and Viticulture, 56(3), 197–206.

Strasburg, G. M. and Ludescher, R. D., 1995. Theory and applications of fluorescence spectroscopy in food research. Trends in Food Science and Technology, 6(3), 69-75.

Chapter 2

Literature Review

Red wine phenolics and their quantification methods

2.1 INTRODUCTION

The grapevine berry is a compartmentalised organ housing numerous phenolic compounds integral for red winemaking. The final wine phenolic composition is dependent on numerous factors including grape cultivar and chemical composition at harvest, viticultural factors influencing berry development and ripening (climate, soil, irrigation), and winemaking practices including fermentation and ageing (Garrido and Borges, 2013). The French Paradox, conceptualised in the late 1980s, stimulated interest in the health benefits and thus research of red wine phenolics (Guilford and Pezzuto, 2011). Today, the antioxidant properties of phenolic compounds are widely accepted. Additionally, phenolic compounds are known to play a crucial role in red wine organoleptic properties with regards to colour, both the intensity and stability thereof, as well as mouthfeel properties such as bitterness and astringency. Phenolic analysis is therefore of great importance during winemaking with regards to decision making and implementation of practices throughout processing in order to elevate the perceived quality of red wine.

This literature review aims to discuss the phenolics present in grapes and wine, highlighting the current spectrophotometric analysis methods versus more advanced alternative methods, with a focus on fluorescence spectroscopy and chemometrics as well as their growing potential within the wine industry.

2.2 PHENOLICS IN GRAPES AND WINE

Phenolic compounds are a diverse group of secondary metabolites found in grapes and wine that can be classified into two families: flavonoids (flavonols, flavan-3-ols and anthocyanins) and non-flavonoids (phenolic acids and stilbenes) (Garrido and Borges, 2013; Aleixandre-Tudo *et al*., 2018). Flavonoids are found in higher concentrations than non-flavonoids and therefore make a greater contribution to the final wine quality. Flavonoids, as depicted in **Figure 1**, are characterised by a C6-C3-C6 skeleton and two benzene rings connected to a heterocyclic pyran ring (Cheynier *et al*., 2006). The diversity of flavonoid subclasses arises from the rearrangement and oxidation state of this pyran ring (Garrido and Borges, 2013). The grape skin accumulates anthocyanins, tannins and hydroxycinnamates (Adams, 2006) while the phenolic compounds found within the seed include flavan-3-ols, catechin, epicatechin and epicatechin-gallate. These may be found in both monomeric and polymeric forms (Downey *et al*., 2003) while differing in their structure and complexity.

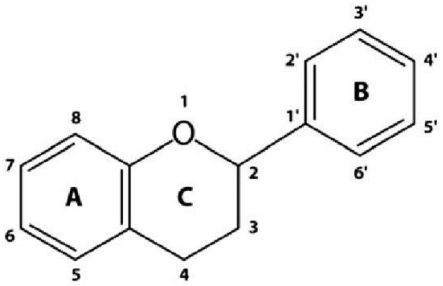


Figure 2.1. Basic flavonoid structure including carbon numbering.

2.2.1 FLAVONOIDS

Flavonols

These phenolic compounds play an important role in the colour of red wine, enhancing extraction of anthocyanins and contributing to co-pigmentation reactions despite being the less abundant class of flavonoids (Schwarz *et al*., 2005). Kaempferol, quercetin and myricetin are the most common flavonols found in wine and are found in grapes as their corresponding glucoside, galactoside and glucuronide derivatives (Adams, 2006).

Flavan-3-ols and condensed tannins

Flavan-3-ols found in the grape skin and seeds include (+)-catechin, (−)-epicatechin, (+)-gallocatechin, (−)-epigallocatechin, and (−)-epicatechin-3-O-gallate (Downey *et al*., 2003). Monomeric flavan-3-ols combine via 4-6 and 4-8 interflavan bonds to form high molecular weight polymers, commonly known as tannins (Adams, 2006). Tannins are the most abundant group of soluble polyphenols in grape berries and have a large influence on the final wine quality (Adams, 2006). Grape seed tannins have terminal subunits of catechin, epicatechin and epicatechin-gallate while skin tannin terminal subunits are primarily catechin *(*Downey *et al*., 2003). Catechin, epicatechin, epicatechin-gallate and epigallocatechin are all detected in grapes as extension subunits, however, epigallocatechin has only been detected in grape skin (Downey *et al*., 2003). Grape seeds have a significantly higher tannin content but possess smaller, less polymerised tannins while skin tannins are larger and have a greater mean degree of polymerisation (mDP) (Kennedy *et al*., 2001). Due to the negative correlation of mDP with bitterness and the positive correlation with astringency, seed tannins are perceived as more bitter while skin tannins are perceived as more astringent (Peleg *et al*., 1999; Pascual, *et al*., 2016).

Tannins can be classified into two groups: condensed and hydrolysable tannins. Condensed tannins or “proanthocyanidins” are grape-derived, naturally occurring oligomers and polymers of flavan-3-ols as described above (Garrido and Borges, 2013). Hydrolysable tannins are oak

derived and comprise of basic gallic and ellagic acid units usually esterified with D-glucose, namely gallotannins and ellagitannins (Garrido and Borges, 2013). Barrel ageing promotes the extraction of these hydrolysable tannins which when hydrolysed, release their gallic and ellagic acid units into the wine to readily react with condensed tannins and anthocyanins, improving colour stability (Ribéreau-Gayon *et al*., 2006).

Anthocyanins

Anthocyanins are the pigments responsible for the red colour of grapes and wine (Garrido and Borges, 2013). Véraison initiates anthocyanin accumulation within only the thick-walled hypodermal cells in red grape varieties, while teinturier varieties store anthocyanins both in the skin and pulp (Adams, 2006). The five most common free anthocyanins found in red wines are cyanidin, peonidin, delphinidin, petunidin and malvidin (Garrido and Borges, 2013) with malvidin-3-glucoside being the most abundant form. This monomeric anthocyanin exists in a dynamic equilibrium influenced by factors such as temperature, sulphur dioxide, the presence of oxygen and most importantly, pH (Aleixandre-Tudo *et al*., 2017). As a result, four chemical states of the anthocyanin may be found including the flavylium cation (red colour), quinoidal base (blue-purple), carbinol pseudobase (colourless) and chalcone (pale yellow) (Ribéreau-Gayon *et al*., 2006).

Anthocyanins are highly reactive and participate in numerous chemical reactions and associations (Aleixandre-Tudo *et al*., 2017). During ageing, anthocyanins polymerise and form more complex, stable compounds such as pyranoanthocyanins and polymeric pigments (Ribéreau-Gayon *et al*., 2006). These compounds are more resistant to colour changes induced by pH shifts or bleaching by sulphur dioxide and therefore retain and enhance the wine colour (Garrido and Borges, 2013). Over time, monomeric anthocyanins possessing a red-purple hue evolve into tawny, brick-red pyranoanthocyanins and polymeric pigments (Quaglieri *et al*., 2017). These anthocyanin derived pigments are formed when tannins and anthocyanins bind via either direct condensation or acetaldehyde mediated reactions, such as the Baeyer reaction (Monagas *et al*., 2005).

2.2.2 NON-FLAVONOIDS

Phenolic acids

Phenolic acids form the most abundant class of non-flavonoids and can be classified into two groups: hydroxycinnamic and hydroxybenzoic acids (Garrido and Borges, 2013). Hydroxycinnamic acids and their tartaric esters are the main class of non-flavonoids in red wine and the main acids include caftaric, p-coutaric and fertaric acids as both *trans* and *cis*

isomers (Garrido and Borges, 2013). Hydroxycinnamate accumulation occurs until véraison and declines during ripening which may be due to catabolism or its utilization in other phenolic compound biosynthesis (Adams, 2006). These compounds are involved in co-pigmentation reactions, contribute partial astringency and bitterness and act as precursors to volatile phenols (Kennedy, 2002; Garrido and Borges, 2013).

The most abundant hydroxybenzoic acids include para-hydroxybenzoic, protocatechuic, vanillic, gallic, and syringic acids (Garrido and Borges, 2013). Gallic acid is of the greatest phenolic importance due to its presence and involvement in hydrolysable tannins (Garrido and Borges, 2013).

Stilbenes

Stilbenes consist of two aromatic rings linked via an ethene bridge and several stilbene-like compounds have been identified in grapes and wine, including resveratrol, astringin and viniferins (Garrido and Borges, 2013). Resveratrol is the most widely studied stilbene in grapes and wine due to its antioxidant health benefits, including positive effects on cardiovascular and neurological diseases as well as possessing anticancer properties (Garrido and Borges, 2013; Fabjanowicz *et al*., 2018). The trans- isomer of resveratrol occurs naturally in grapes and may transform into the cis- isomer during winemaking and ageing (Fabjanowicz *et al*., 2018). These compounds are often classified as phytoalexins due to their accumulation in response to fungal infection, however, accumulation may also be induced by mechanical vine stresses and excessive UV radiation exposure (Adams, 2006; Garrido and Borges, 2013; Fabjanowicz *et al*., 2018).

2.2.3 IMPORTANCE TO WINEMAKING

The evolution of phenolics, from those present in the grapes to those found within the final wine matrix, is influenced by several factors and winemaking techniques. Different processing practices can alter the concentration and composition of phenolic compounds throughout the vinification process (Garrido and Borges, 2013) and thus, an understanding of phenolic composition within the berry may aid in decision-making for improved or reduced extraction. Monitoring this extraction of phenolic compounds during the winemaking process may therefore aid in the timely implementation of appropriate techniques, briefly discussed below, for their desired effect on the final wine phenolic profile.

Winemaking Techniques

Phenolic extraction relies on adequate berry rupturing and occurs as a result of the subsequent contact between solid grape parts and must (Aleixandre-Tudo *et al.*, 2018). The concentration of phenolic compounds increases throughout fermentation due to their greater solubility in ethanol (Sacchi *et al*., 2005). Anthocyanin extraction reaches a maximum peak earlier on in fermentation, followed by a noticeable decline thereafter (Sacchi *et al*., 2005; Aleixandre-Tudo *et al.*, 2018) while tannin extraction continues with extended skin and seed contact (Sacchi *et al*., 2005). Skin tannins follow a sigmoidal extraction curve until reaching a saturation plateau while seed tannins follow a linear extraction, highlighting the need for seed coat hydration for extraction to occur (Cadot *et al*., 2006).

Several different winemaking techniques have been studied for their influence on phenolic extraction. Pre-fermentative treatments include must or grape freezing for increased skin bursting and contact between solid grape parts and must, cold maceration for improved aqueous extraction, the addition of pectolytic enzymes for increased juice yield and cell wall breakdown as well as the removal of juice prior to fermentation, termed *saignée*, for an increased skin to juice ratio (Sacchi *et al*., 2005).

Fermentative treatments include various pump-over and punch-down frequencies, as well as *delestage* (rack and return), for the dispersal of trapped heat within the fermentation cap and increased mixing between the juice and skins. Other practices include different yeast selections which may have implications for phenolic absorption, carbonic maceration for partial intracellular fermentation as well as thermovinification for increased cell membrane damage and an enhanced release of phenolic compounds (Sacchi *et al*., 2005). Fermentation temperature has also been studied as a factor in phenolic extraction, in which greater temperatures resulted in greater phenolic extraction (Sacchi *et al*., 2005).

The post-fermentative technique of extended maceration is applied for increased phenolic extraction due to increased skin and seed to wine contact (Sacchi *et al*., 2005). Of these applied methods, cold maceration, carbonic maceration, yeast selection and skin and juice mixing practices have shown variable results with regards to the phenolic profiles of finished wine and may be particularly variable according to cultivar (Sacchi *et al*., 2005).

Impact on Wine Attributes

Phenolic compounds take part in numerous reactions throughout fermentation and ageing, including condensation, oxidation, adsorption and precipitation reactions (Pérez-Magariño and González-San José, 2004; Aleixandre-Tudo *et al.*, 2018). These evolutionary changes of phenolics have implications for final wine traits, specifically with regards to colour and mouthfeel. By the end of alcoholic fermentation, 25% of monomeric anthocyanins have

polymerised and this increases to around 40% the following year (Monagas *et al*., 2005). These polymerised anthocyanins, together with other anthocyanin derivatives or polymeric pigments, have increased stability and resistance and therefore, play important roles in the visual aspect of wine colour.

Astringency and bitterness are important sensorial attributes of red wine. The concentration, mDP and galloylation of proanthocyanidins affects the intensity of these attributes (Cliff *et al*., 2007). Bitterness is a taste mediated by mouth receptors while astringency is a mouthfeel sensation commonly described as dry and rough due to the interaction between proanthocyanidins and salivary glycoproteins (Vidal *et al*., 2003). Astringency tends to decrease during wine ageing, leading to "smoother, softer" wines. Currently, three potential mechanisms for this reduced sensation are proposed including cleavage reactions reducing tannin size, molecular rearrangement resulting in bulkier tannins unable to react with salivary proteins either due to steric hindrance or tannin-anthocyanin reactions and lastly, the formation of new polymeric pigments (Aleixandre-Tudo *et al*., 2017). Due to the sensorial importance of phenolic compounds, a greater understanding of their attributes and the techniques to modify them may result in improved red wine decision making. Monitoring phenolic compounds and their evolution throughout the winemaking process may therefore be of great benefit to the producer.

2.3 PHENOLIC ANALYSIS METHODS

A top priority of phenolic analysis research has been to establish reliable, robust and sensitive methods. The complexity and diverse range of phenolic compounds has resulted in the development of numerous established analytical methods. However, due to the need for trained personnel and often expensive equipment and reagents, the routine analysis of phenolic measurements is not a widespread practice in winemaking. Of the limited wineries conducting phenolic analysis, few perform on-site measurements while most utilize external analytical laboratories (Harbertson and Spayd, 2006).

Standard phenolic analysis currently conducted can therefore, in this case, be described as the simple spectrophotometric methods most often used in the analysis of phenolic compounds. UV-Vis spectroscopy, involving the absorption of light in the ultraviolet and adjacent visible spectra, has been reported to successfully analyse phenolic compounds by several authors due to the spectral properties of phenolics (Aleixandre-Tudo *et al*., 2018; Beaver and Harbertson, 2016; Dambergs *et al*., 2012, Harbertson and Spayd, 2006). The aromatic ring shared by all phenolic compounds allows for the absorption of light within the

ultraviolet region and the transitions occurring within the OH-groups allow for the differentiation of phenolic classes via characteristic peaks at various wavelengths (Harbertson and Spayd, 2006; Aleixandre-Tudo *et al*., 2017). The benefits of UV-Vis spectroscopy (simple, reliable, rapid and cost effective) have allowed it to become a widespread application (Harbertson and Spayd, 2006; Aleixandre-Tudo *et al*., 2017).

Due to the increasing need for rapid and accurate methods of analysis that are more cost-effective and user-friendly, other alternatives such as spectroscopy combined with chemometrics are becoming increasingly investigated in both academic and industry domains. The advantages of being able to measure phenolic compounds throughout fermentation more frequently as well as on site include process control, monitoring and optimisation. On-line systems implemented for process control in wineries are becoming increasingly popular and therefore requires suitable technology with desirable features such as speed, accuracy and non-destructive analysis methods (Daniel, 2015; Aleixandre-Tudo *et al*., 2019). The spectrophotometric methods described below illustrate the current phenolic analysis techniques employed together with their benefits and drawbacks.

2.3.1 SPECTROPHOTOMETRIC ANALYSIS

Protocols for Tannin Analysis

The most widely used tannin analysis protocols include the acid hydrolysis, BSA (bovine serum albumin) and MCP (methyl cellulose precipitable) tannin assays. Acid hydrolysis is based on the transformation of proanthocyanidins to carbocations, followed by a conversion to anthocyanins when heated in an acid medium (Ribéreau-Gayon and Stonestreet, 1965). The BSA tannin assay relies on tannin-protein interactions and their subsequent precipitation (Harbertson *et al*., 2002) while the MCP tannin assay is a precipitation-based method relying on polymer-tannin interactions, specifically between tannins and methyl cellulose (Sarneckis *et al*., 2006).

Acid hydrolysis is the most simple and user-friendly method while the BSA and MCP assays have the benefit of estimating astringency due to positive correlations between the two (Kennedy *et al*., 2006; Mercurio and Smith, 2008; Harbertson *et al*., 2015). The MCP assay requires fewer and simpler reagents than its BSA counterpart and is therefore often preferred in the application of routine analysis. Additional drawbacks of BSA include frequent over- or under-estimation of tannin levels (Harbertson *et al*., 2015; Jensen *et al*., 2008; Mercurio and Smith, 2008) while precipitation based methods have the overall disadvantage of being unable to be implemented within inline systems.

Protocols for Total Phenolic Analysis

Total phenolics may be quantified using either the Folin-Ciocalteu assay (Singleton and Rossi, 1965) which relies on the donation of electrons during redox reactions or the Total Phenolics Index (TPI) which relies on the characteristic absorption of UV light by phenolic compounds at 280 nm (Somers and Evans, 1974). Both methods share the risk of over-estimation due to other non-selective compounds interfering with redox reactions or possessing light absorbing abilities at 280 nm (De Beer *et al*., 2004; Somers and Evans, 1974). Of the two, TPI is considered a simpler and more user–friendly option.

Protocols for Anthocyanin Analysis

Anthocyanins are easily quantifiable using spectrophotometric methods due to their characteristic absorption peak between 490 nm and 550 nm (Giusti and Wrolstad, 2001), resulting in a sufficient standard measurement at 520 nm. Three commonly employed methods include bisulfite bleaching (Ribéreau-Gayon and Stonestreet, 1965), the hydrochloric acid method (Iland *et al*., 2000; Cliff *et al*., 2007) and the Modified Somers Assay (Mercurio et al, 2007). These methods rely on the equilibria of anthocyanins and the mechanisms activated and influenced by pH shifts as well as the bleaching effect of bisulfite. These methods have all shown reliability due to the sensitivity of anthocyanins to pH changes and free sulphur dioxide (SO_2). The modification of the original Somers and Evans procedure by Mercurio *et al.* (2007), simplifies the protocol by incorporating a buffer solution (12% v/v ethanol, 0.5 g/L w/v tartaric acid at pH 3.4) thereby increasing ease of use. The Modified Somers Assay presents itself as an extensive analysis method that holistically evaluates the equilibria of anthocyanins via additions of bisulfite, acetaldehyde and hydrochloric acid in order to quantify anthocyanins in their original form, as sulphur resistant pigments, as liberated anthocyanins previously bound by sulphur and finally, in their coloured red flavylium form.

Protocols for Colour Analysis

Red wine colour experiences numerous transitions over time as a result of the interactions between anthocyanins and other phenolic compounds (Harbertson and Spayd, 2006). Colour density (Glories, 2016) and the CIELab Colour Space (CIE, 1978; Harbertson and Spayd, 2006) are commonly used methods for colour analysis. Both methods have the advantage of analysing samples without requiring dilution. Colour density is the estimation of total colour using the sum of absorbances at three wavelengths, namely 420 nm (yellow colouration), 520 nm (red colouration) and 620 nm (blue colouration). Wine hue (A_{420nm}/A_{520nm}) can be used as an indication of polymerisation during ageing due to the shift of monomeric anthocyanins in the red flavylium form to more stable, yellow/orange polymeric pigments. CIELab is a widely accepted tri-stimulus colour specification method developed by the Commission Internationale

de l'Eclairage which converts spectral results into tri-stimulus values of lightness and chromatic components. The sample's colour can then be identified and matched to what is visually perceived.

HPLC Analysis

High-performance liquid chromatography (HPLC) is a selective and accurate analytical technique involving the separation of compounds within a column and the characterisation of these separated compounds over the UV-Vis region. The analysis of phenolic compounds is typically conducted by a reversed phase (RP-HPLC) C_{18} column with a binary elution system (Ibern-Gómez *et al*., 2002; Burin *et al*., 2011). The benefits of this method include highly selective and accurate qualitative and quantitative analyses, however, the requirement of sample preparation, expensive equipment and reagents, highly skilled personnel and lengthy analysis time create major drawbacks (Aleixandre-Tudo *et al*., 2018) and therefore, limits the use of HPLC for routine analysis outside of research applications.

2.3.2 ABSORPTION SPECTROSCOPY ANALYSIS

Absorption spectroscopy analysis has been widely studied for its applicability to grape and wine analysis. The combination of spectroscopy with chemometrics allows for the deconstruction and improved interpretation of complex data sets. Spectroscopy coupled with multivariate calibration holds several advantages over the numerous spectrophotometric methods currently used, including more rapid analysis, the simultaneous measurement of several analytes at once, non-destructive technique and requiring minimal sample preparation (Aleixandre-Tudo *et al*., 2018; Gishen *et al*., 2005). These advantages have presented the potential for automation and on-line systems as well as optical portable devices (Giovenzana *et al*., 2013).

Infrared spectroscopy, involving the absorption of light in the infrared spectral region, is particular sensitive to the fundamental molecular vibrations of specific functional groups (Daniel, 2015). Both mid infrared (MIR) and near infrared (NIR) spectroscopies have been studied for their applicability to phenolic analysis while more recently, the combination of Fourier transform mid infrared spectroscopy (FT-MIR) with chemometrics has been deemed useful (Romera-fernández *et al*., 2012; Daniel, 2015).

2.4 FLUORESCENCE SPECTROSCOPY

Fluorescence spectroscopy, a type of emission spectroscopy, has been used extensively in chemistry and biochemistry disciplines due to its success in analysing the structures, functions and reactivity's of several compounds, ranging from small biological molecules to polymers and proteins (Strasburg and Ludescher, 1995). This success allowed fluorescence spectroscopy to branch out into food science disciplines to meet the growing demands for improved quality and food safety throughout increasingly industrialised food supply chains, while being used for the chemical characterisation of compounds as well as authenticity and quality control (Airado-Rodŕiguez *et al.*, 2009; Karoui and Blecker, 2011). Several food science applications have been successful, including analysing the stability of cheese and yoghurt during storage (Christensen *et al*., 2003; Christensen *et al*., 2005) as well as discriminating between virgin and pure olive oils (Guimet *et al.*, 2004). The advantages of fluorescence spectroscopy include it being non-destructive, easy to use, relatively inexpensive, rapid and highly sensitive when compared to other spectrophotometric methods, all of which highlight the potential for use within online systems and devices (Karoui and Blecker, 2011; Strasburg and Ludescher, 1995; Airado-Rodríguez *et al*., 2011).

2.4.1 PRINCIPLES OF FLUORESCENCE

Fluorescence spectroscopy typically involves the absorption of light (excitation) within the UV-Visible spectrum which excites the fluorophore (fluorescent chemical compound or compound of interest) followed by energy redistribution and decay with accompanied emission of light (emission). Typically the entire fluorescence spectrum is measured as a function of the excitation wavelengths. The detected result is thus an excitation-emission matrix (EEM) or fluorescent landscape of the sample. The Jablonski diagram below (**Figure 2.2**) illustrates the electronic transitions taking place within a fluorescent molecule. The first stage including energy absorption and electron excitation, allows electron transitioning from the ground state (S_0) to an excited state (S_n) (Albani, 2008). As the excited electrons return to a lower energy level (S_1), energy is dissipated into the surrounding environment in a process known as internal conversion. The return of these electrons to ground state follows various processes including the emission of a photon (fluorescence), the dissipation of non-radiative heat into the medium, energy transfer to surrounding molecules (quenching) and the transition to an excited triplet state (T_1). This triplet state forms part of an alternative de-excitation method known as phosphorescence.

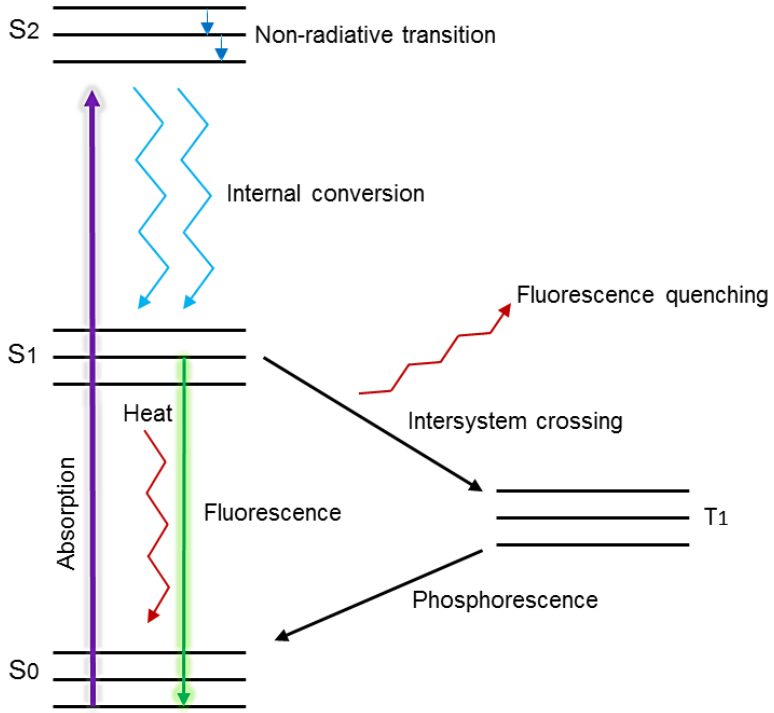


Figure 2.2. Jablonski diagram adapted from literature (Albani, 2008).

2.4.2 FACTORS AFFECTING FLUORESCENCE INTENSITY

The intensity of fluorescence emission is directly proportional to the quantum yield, simply defined as the probability of photon emission by the excited fluorophore (Strasburg and Ludescher, 1995). Intensity is also related to the fluorescence lifetime of a molecule, defined as the time spent in the excited state of the fluorophore's electrons before returning to ground state and can range from nano- to picoseconds (Albani, 2008). The three following factors have implications for fluorescence intensity, namely quenching, the local environment and scatter phenomena.

Any process resulting in the deactivation of the excited molecule via intra- or intermolecular interactions is defined as quenching (Karoui and Blecker, 2011). Static quenching involves the formation of non-fluorescent complexes between electrons and quencher molecules in the ground state, thereby inhibiting excitation, while dynamic quenching involves deactivation of the electron post-excitation via collisions or inter-molecular interactions (Albani, 2008; Karoui and Blecker, 2011).

Environmental factors such as temperature, pH and colour, impact fluorescence intensity due to the high sensitivity of fluorophores to their surrounding environments (Strasburg and Ludescher, 1995; Karoui and Blecker, 2011). Higher sample temperatures during analysis

may increase quenching due to greater velocities of collisions during fluorescence (Karoui and Blecker, 2011). Different fluorophores may fluoresce greater at certain pH levels, such as hydroxyl aromatic compounds, while sample colour influences absorption of the excitation beam, thereby influencing both the shape and the intensity of the fluorescence emission (Karoui and Blecker, 2011). These factors have numerous implications in the interference of obtaining true excitation-emission spectra and should be considered during sample preparation.

Rayleigh and Raman scattering are light scatter phenomena most problematic to fluorescence spectroscopy. In Rayleigh scattering, light is scattered by particles much smaller than the wavelength of light and does not involve energy loss. Due to this elasticity, first-order Rayleigh scatter ($\lambda_{ex} = \lambda_{em}$) is observed as the diagonal line running through the emission landscape (**Figure 2.3**) as excitation wavelength closely equals the emission wavelength (Karoui and Blecker, 2011). Second-order Rayleigh scattering occurs at twice the excitation wavelength ($2\lambda_{ex} = \lambda_{em}$) and is generally an instrumental detector issue, which is an artefact from the second order diffraction of the grating used in the detector. Raman scattering is inelastic light scattering as a result of light interaction with specific vibrational states of a molecule. The influence of Raman scattering is often negligible due to its weaker contribution to the fluorescent landscape (Karoui and Blecker, 2011) . Pre-processing of spectral data involving the removal of scattered interference is essential for the success of fluorescence analysis and is often one of the first stages in the pre-treatment of such complex data sets (Airado-Rodríguez *et al.*, 2011; Cabrera-Bañegil *et al.*, 2017).

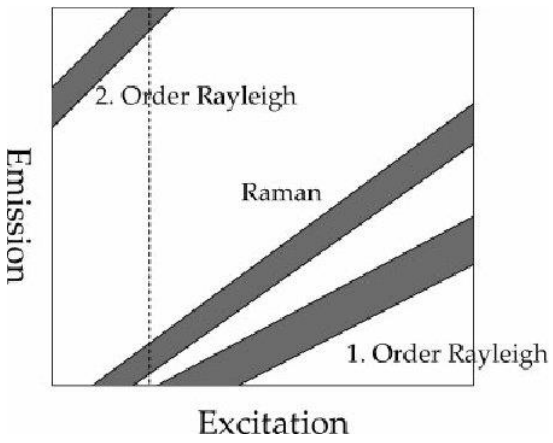


Figure 2.3. The scatter phenomena locations within a fluorescence EEM (Buhram *et al.*, 2006).

2.4.3 INSTRUMENTATION

Several considerations are highly important when looking at the instrumentation of spectrofluorometers due to their direct influence on the success of fluorescence analysis. Additionally, it is important to note that every spectrofluorometer is unique as a result of the non-uniform spectral output of light sources and the wavelength-dependency of monochromators and detectors (Lakowicz, 2013). An understanding of the numerous components and their effects on the instrument's spectral output allows for improved control by the user. The basic setup of a spectrofluorometer includes a light source, motorised monochromators, sample chamber, a detector and appropriate quantification devices (**Figure 2.4**). In recent times, there has been an increasing focus towards the development of smaller, more compact devices that allow all components to be encased within a single enclosure (Lakowicz, 2013; Bridgeman *et al.*, 2015).

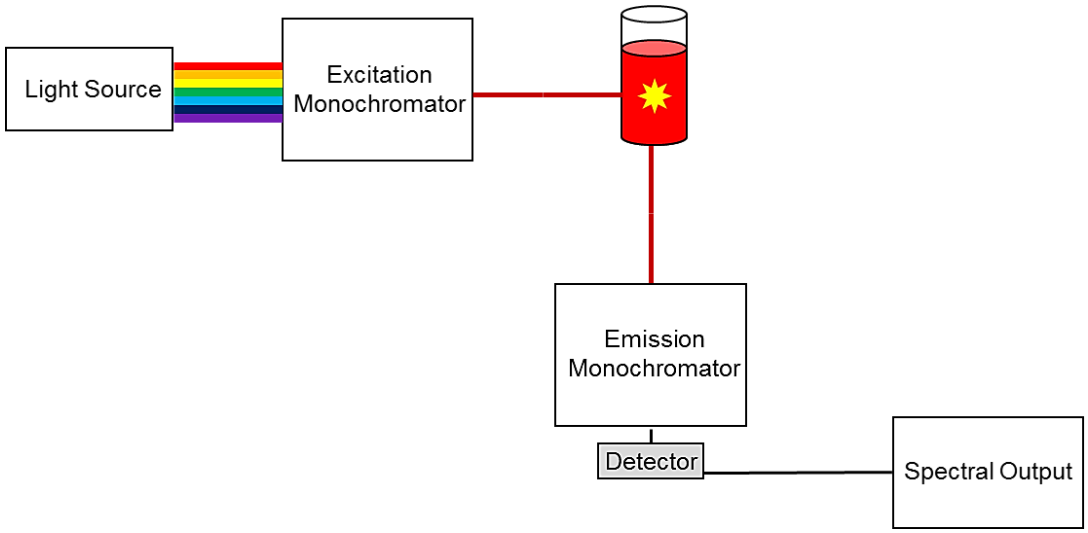


Figure 2.4. Basic instrumentation of a spectrofluorometer adapted from literature.

Currently, the most ideal light sources are high-pressure xenon arc lamps (Lakowicz, 2013). These lamps are able to emit a relatively continuous spectrum of light between 250 nm and 700 nm. As the need for more compact and user-friendly devices increases, a shift to pulsed xenon lamps has become common. These flash lamps hold several advantages other than their compact size, such as having a greater peak intensity, generating less heat while using less power and the potential for reducing photo-damage to samples due to their pulsed rather than continuous nature (Lakowicz, 2013). Additionally, the use of LED light sources is becoming increasingly popular due to their energy efficiency, lower cost and longer lifetime (Bridgeman *et al.*, 2015).

Monochromators are used to select for a specific wavelength from a multi-coloured or white light source. Slit widths of the excitation and emission monochromators can be fine-tuned to

allow for more or less light passage. Larger slit widths allow for increased light intensity but risk an increased signal to noise ratio while smaller slit widths may improve spectral resolution but risk losing light intensity (Lakowicz, 2013). The dispersion of light in a monochromator is achieved using either planar or concave diffraction gratings and important considerations include their effects on efficiency, dispersion and stray light levels. A higher efficiency allows for the detection of lower light levels while dispersion is greatly affected by the type of grating used. Stray light can be defined as any additional light passing through the monochromators other than the selected wavelength and is a key consideration for the success of fluorescence analysis due to its effect on fluorescent interference, specifically scatter phenomena and their effect on emission intensities (Lakowicz, 2013). Filters may be used to reduce light scattering as well as the analysis of control samples which is an important method for ruling out spectral noise from interfering compounds.

Once the sample has been excited, the emitted light is captured by a highly sensitive light detector, typically photomultiplier tubes, consisting of a photocathode and a series of dynodes (Lakowicz, 2013). Photons from the fluorescence emission are detected by the sensitive photocathode and the subsequently generated electrons are amplified via each successive dynode. The detected signal is thereafter quantified, displayed and stored within the appropriate electronic device.

2.4.4 FLUORESCENCE SPECTROSCOPY IN WINE

The complex wine matrix consists of several naturally occurring fluorescent compounds. Polyphenols form the largest concentration while vitamins (specifically B-complex) and amino acids also possess fluorescent properties (Airado-Rodríguez *et al.*, 2011). Previous research has been conducted to determine the fluorescent capabilities of the wine matrix, including identifying spectral regions of interest, and to subsequently measure the responsible fluorescent compounds. The majority of wine fluorescence research has been qualitative, focusing on classification and discrimination tasks while more recently, quantitative applications have been explored. Airado-Rodríguez *et al*. (2011) revealed four fluorophores responsible for the main fluorescence of Spanish wines and determined the potential for discriminating according to appellation. Cabrera-Bañegil *et al*. (2019) were able to distinguish between water-stressed and irrigated vines while quantifying catechin, epicatechin and resveratrol. Another study determined a good correlation for the quantification of vanillic acid, caffeic acid, epicatechin and resveratrol (Cabrera-Bañegil *et al*., 2017). Letort *et al*. (2006) illustrated the potential for wine authentication according to cultivar, region and vintage by means of distinguishing between French and German wines. Bottled white wines were

analysed by Coelho *et al*. (2015) and highlighted the potential of fluorescent fingerprints in revealing chemical signatures of oenological and vintage specific treatments, specifically with regards to various sulphur dioxide treatments applied to musts at pressing.

Fluorescent properties of phenolic compounds naturally occurring in wine have been widely studied and the optimal excitation and emission wavelengths of these compounds have been reported throughout literature. Two general regions have been determined within the wine excitation-emission matrix, specifically, excitation between 250 and 290 nm resulting in emission between 300 and 430 nm, while excitation at wavelengths greater than 300 nm results in emission between 360 and 450 nm (Airado-Rodríguez *et al.*, 2009; Airado-Rodríguez *et al*., 2011). Airado-Rodríguez *et al*. (2011) created an integrated depiction (**Figure 2.5**) of the fluorescent regions of polyphenols based on and in accordance with numerous studies. The above research studies all have two important factors in common, namely, sample geometry and the application of chemometrics.

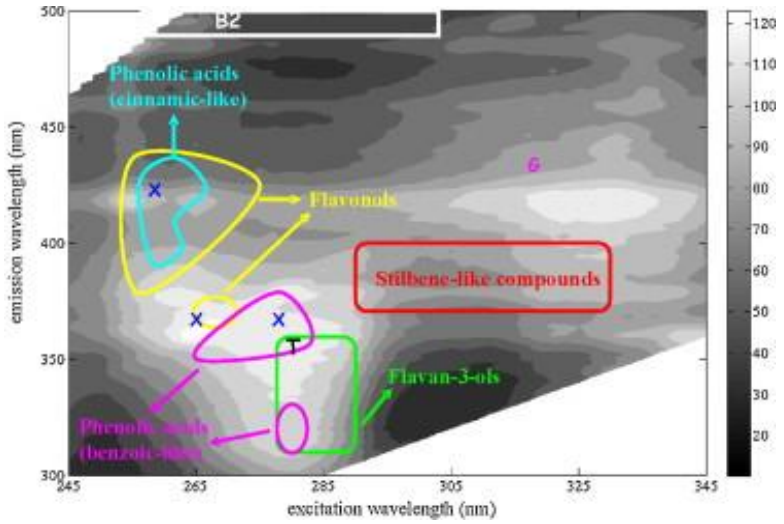


Figure 2.5. Fluorescent landscape of a red wine sample depicting regions of polyphenol fluorescence (X = phenolic aldehydes; G = gentisic acid; T = tryptophan; B2 = vitamin B2). (Airado-Rodríguez *et al*., 2011).

Sample Geometry

Within the sample chamber of spectrofluorometers, one of the most important considerations involves the geometric arrangement of the sample. Fluorescence spectroscopy has traditionally used the conventional right angle technique whereby the incident angle, defined as the angle formed between the excitation beam and the surface perpendicular, is 0° (Airado-Rodríguez *et al.*, 2011). Typically, clear or diluted samples may be analysed in this manner, however, dilution may interfere with the accuracy of results (Airado-Rodríguez *et al.*, 2011) as well as reducing the applicability for food science disciplines due to the inability of analysing

viscous or turbid food samples (Karoui and Blecker, 2011). Owing to the complexity of the wine matrix and the interactions taking place within it, as well as the sensitivity of fluorophores to their surrounding environment, a new technique developed by Parker (1968) overcomes the need for dilution. This is achieved by changing the angle of incidence to around 30° (**Figure 2.6**), thereby reducing scattered light and unwanted spectral noise while analysing samples in their truest form (Lakowicz, 2013; Airado-Rodríguez *et al*., 2011).

Figure 2.6. Right angle geometry (left) versus front-face geometry (right) of cuvettes adapted from literature (Lakowicz, 2013).

2.5 CHEMOMETRICS

Fluorescence spectroscopy generates a fluorescent landscape in the form of an excitation-emission matrix (EEM). This complex three-dimensional fluorescent landscape requires the use of multivariate statistical analysis (chemometrics) to decompose and easily interpret the obtained fluorescence signals (Airado-Rodríguez *et al.*, 2011). Chemometrics may be defined as the extraction of chemically relevant information via mathematical and statistical tools (Varmuza and Filzmoser, 2016). The combination of spectroscopy with chemometrics holds several advantages including a considerable reduction in time of analysis as well as the simultaneous analysis of multiple analytes from a single spectral measurement (Aleixandre-Tudo *et al.*, 2017). The most commonly used multi-way analytical models for fluorescence spectroscopy include parallel factor analysis (PARAFAC) in addition to unfolded (U-PLS) and N-way partial least squares (N-PLS) (Andersen and Bro, 2003; Cabrera-Bañegil *et al.*, 2017). As PARAFAC is most often the tool of choice, the following sections will compare it to modern techniques of machine learning that have previously not been investigated despite their success in complex data analysis.

2.5.1 PARALLEL FACTOR ANALYSIS

Parallel factor analysis (PARAFAC) is a trilinear decomposition model able to identify underlying chemical components within complex data sets (Murphy *et al.,* 2013). It is successfully implemented for fluorescence data due to the fundamental three-way array of

sample x *excitation wavelength* x *emission wavelength* as depicted in the three-dimensional cubed data format below (**Figure 2.7**). Compared to bilinear principle component analysis (PCA) of one score and one loading vector, each PARAFAC component consists of a score vector and two or more loading vectors (Bro, 1997; Airado-Rodríguez *et al.*, 2011). The PARAFAC model is considered to be a simpler and constrained version of bilinear PCA while overcoming the problem of rotational freedom (Bro, 1997). The three-way PARAFAC model formula can be written as:

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

where $\mathbf{X}$ is the three-way data array, $\mathbf{F}$ is the number of components, $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ correspond to the three loading matrices and $\mathbf{e}$ represents the residual error (Airado-Rodríguez *et al.*, 2011; Murphy *et al.,* 2013). Variables $\mathbf{i}, \mathbf{j}$ and $\mathbf{k}$ represent the sample, excitation and emission modes, respectively.

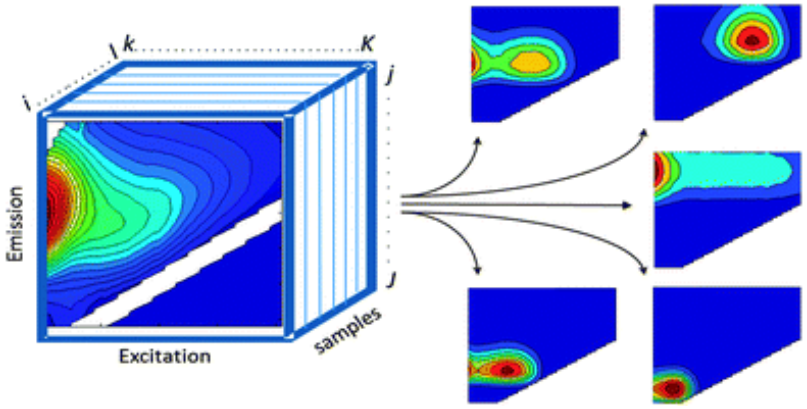


Figure 2.7. A three-way array of EEM data decomposed into five PARAFAC components (Murphy *et al.*, 2013).

Data pre-processing is an important exploratory phase of PARAFAC modelling and involves correcting three main aspects namely instrumental, non-trilinear and intensity distortion. Instrumental distortion requires the removal of any systematic errors or interferences. Due to the uniqueness of each spectrofluorometer, instrumental fluctuations and optical effects may distort excitation and emission spectra. This may be removed by EEM multiplication with correction vectors identified for the particular instrument in use and may be implemented automatically during fluorescence analysis or applied during data pre-processing (Murphy *et al.,* 2013). Non-trilinear distortion requires the removal of any signals unrelated to the fluorescence of the sample such as Rayleigh and Raman scattering, which may easily be removed by inserting missing values within the identified spectral bands (Andersen and Bro,

2003; Airado-Rodríguez *et al.*, 2011; Murphy *et al.,* 2013). Intensity distortion requires normalising datasets with large concentration gradients. Normalisation allocates similar weightings to both high and low concentration samples which may be hindering the discrimination between important chemical variations (Murphy *et al.,* 2013).

Model Exploration

The exploratory phase following data pre-processing involves obtaining the best decomposed dataset for modelling and includes choosing the optimal number of components and the identification and potential removal of outliers (Murphy *et al.,* 2013). Deciding on the optimal number of components may be achieved using the core consistency diagnostic (CORCONDIA) or explained variance which are able to determine the appropriateness of the model (Andersen and Bro, 2003; Murphy *et al.,* 2013). The core consistency is expressed as a percentage, with the most optimal number of components being the closest to 100%, after which the consistency tends to drop significantly once too many components have been selected. However, deciding on the number of components cannot be based solely on mathematical criteria such as CORCONDIA and explained variance. A combination of methods together with a good understanding of the dataset and logical reasoning are essential (Andersen and Bro, 2003; Murphy *et al.,* 2013).

Model Validation

Several steps can be taken to validate the most optimal PARAFAC model. These include examining the visual characteristics of spectral loadings, validation techniques such as Jack-knife or split-half analysis, as well as applying model constraints (Andersen and Bro, 2003; Murphy *et al.,* 2013). Excitation and emission spectral loadings for each component represent the fluorescence activity of responsible analytes and visually examining these outputs may indicate model concerns. Characteristics such as abruptly sharp peaks, large regions of overlap between excitation and emission loadings (> 50 nm) and large negative regions indicate that the model is incorrectly identifying the responsible chemical components and should be refined (Andersen and Bro, 2003). Methods to determine the stability and robustness of the model, such as Jack-knife and split-half analysis, allow for several models to be produced either from a leave-one-out method or dividing the dataset into halves each time while applying model constraints such as non-negativity and unimodality (single peak spectra) may also be useful in improving unstable models (Andersen and Bro, 2003; Murphy *et al.,* 2013).

Once model validation is complete, the loadings and score values for each component may be used in different applications. In relation to wine research previously conducted, score

values for each component may be plotted against each other to determine the classification potential between wine appellations (Airado-Rodríguez *et al.*, 2011), spectral loadings may be correlated with and identified according to those of pure standards and score values may be correlated with concentration calibration sets in order to quantify phenolic compounds (Cabrera-Bañegil *et al.*, 2017). Calibration models such as those built for catechin and epicatechin as well as vanillic acid highlight the potential of creating regression models for quantitative analysis (Cabrera-Bañegil *et al.*, 2017; 2019).

2.5.2 MACHINE LEARNING

Machine learning has allowed for new opportunities to emerge in multi-disciplinary agri-tech applications as a result of its high performance in data intensive scenarios (Liakos *et al*., 2018). The ability of computer based algorithms to learn from and make predictions on the data provided is known as machine learning (Elith, 2019). Machine learning can be distinguished between two broad categories, namely supervised and unsupervised learning. Supervised learning involves training the algorithm on a dataset and indicating the output response required while unsupervised learning aims to identify structure and similarities between inputs without a specified response or target (Elith, 2019). The appeal of machine learning therefore involves the ability for an algorithm to improve with experience over time. The growing success of machine learning has resulted in its multi-disciplinary use and can be seen today in almost every sector including economic use in banking systems, medical diagnosis in health care, in the control and optimisation of manufacturing processes as well as in internet search engines (Alpaydin, 2020). The process of applying machine learning involves data preparation followed by feature, algorithm and parameter selection, and lastly, training and evaluation. Gradient tree boosting is extensively used in the data science field and a new variation of this technique, known as eXtreme Gradient Boosting (XGBoost), has become well recognised in achieving highly effective results (Chen and Guestrin, 2016).

XGBoost

Boosting regression trees are an ensemble technique combining the capabilities of two algorithms; regression trees and boosting. Classification and regression trees experience repeated binary splits, hierarchically splitting the data into regions with the most homogenous response to predictors (Elith *et al*., 2008). Boosting is a sequential, stage-wise technique whereby sequential modelled trees are optimised by minimising the loss of predictive function from a previously sub-optimal model (Elith *et al*., 2008). This binary splitting occurs at each tree's outputs, resulting in hundreds to thousands of possibilities, until specified stopping criteria have been reached. Instead of selecting a single "best fit" model as with more simple

regression techniques, the final output is a linear collection of numerous trees and models (Elith *et al*., 2008). Important parameters of the process include learning rate (the contribution of each tree to the final model), tree complexity (number of iterations fitted) and number of trees required for optimal prediction. **Figure 2.8** illustrates the process of gradient boosting using a classification example, whereby incorrectly predicted features (red circles in the first iteration) are weighted higher in the subsequent tree (larger red circles) until all circles are accurately classified by the third iteration (Zhang *et al*., 2018). A new split is applied in the third iteration in order to correctly classify the remaining blue square and resulting in model F_3 as the sum of T_1, T_2 and T_3.

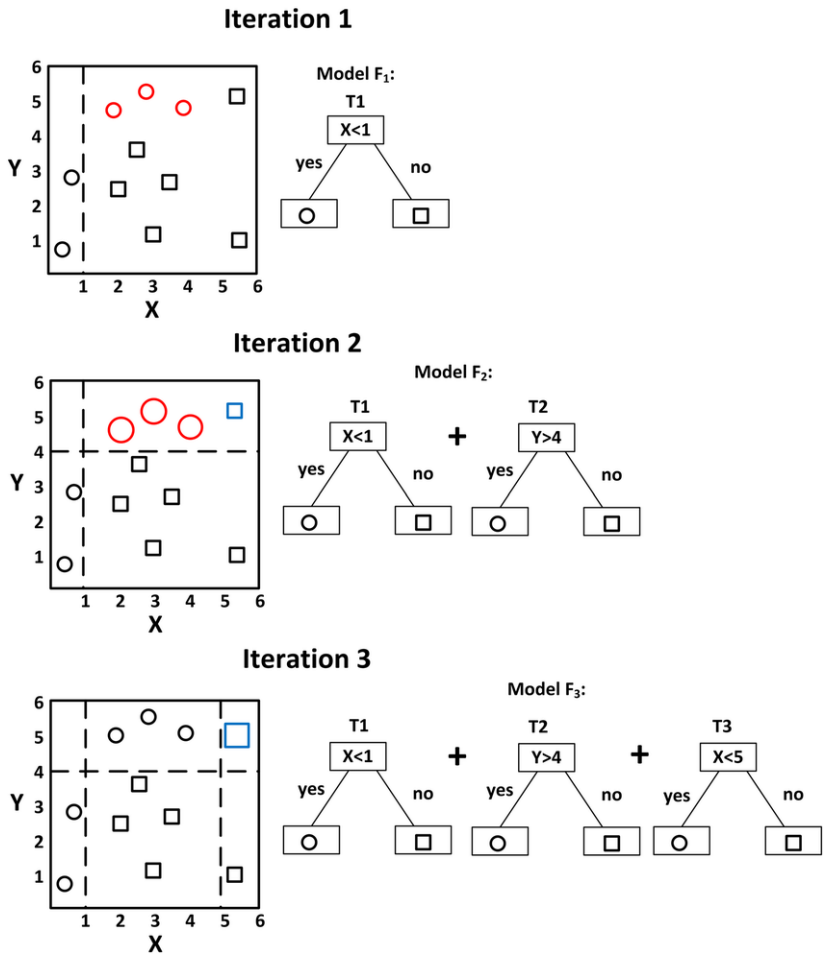


Figure 2.8. A visual example of gradient boosting (Zhang *et al*., 2018).

XGBoost is used for classification, regression and ranking applications. The advantages of this tree boosting algorithm include scalability in all scenarios, built-in handling of sparse data and parallel processing allowing the system to run ten times faster than other popular systems (Chen and Guestrin, 2016). Regularisation methods, shrinkage and scaling of weights as well as column sub-sampling helps to reduce over-fitting which is likely during additive accumulation of the models (Chen and Guestrin, 2016). XGBoost is a highly adaptable

machine learning algorithm, unaffected by dimensionality and non-linear interactions within datasets and presents itself as a novel, present day alternative for complex data analysis. These benefits may be beneficial in handling complex fluorescence matrices and potentially allow for the discrimination of previously unidentified correlations between spectral regions and polyphenols.

Gradient boosting, random forest and ensemble techniques have shown success in predicting skin flavonoid content from red wine grape cultivars (Brillante *et al*., 2015) while Gupta (2018) explored the use of machine learning algorithms, such as neural networks and support vector machines (SVMs), in determining correlations between wine quality attributes and physicochemical characteristics. Machine learning appears to be increasingly popular in wine applications and therefore presents itself as a state-of-the-art alternative for complex technologies (Brillante *et al*., 2015).

2.6 CONCLUSION

The importance of phenolic compounds in red wine due to health benefits as well as quality attributes is well known. To date, conventional analysis has not formed part of routine wine analysis, largely due to the need for numerous reagents, equipment, trained personnel, time of analysis and cost. The need for fast, simple, cost-effective and accurate phenolic analysis methods is steadily increasing. More routine analysis of phenolic compounds may have a positive effect on the decision-making of winemakers in order to maximise the quality of the final red wine product from the initial grape source. Fluorescence spectroscopy, together with chemometrics, shows promise for phenolic analysis and requires further research to determine its potential for phenolic quantification of South African wines. The potential for implementation into on-line systems or optical devices may allow for the monitoring of phenolic compounds and their extraction during winemaking, ultimately improving red wine production and the performance of local wines on the competitive global scene.

2.7 REFERENCES

Adams, D. O., 2006. Phenolics and ripening in grape berries. American Journal of Enology and Viticulture, 57(3), 249–256.

Airado-Rodŕiguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2009. Usefulness of fluorescence excitation-emission matrices in combination with parafac, as fingerprints of red wines. Journal of Agricultural and Food Chemistry, 57(5), 1711–1720.

Airado-Rodríguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2011. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. Journal of Food Composition and Analysis, 24(2), 257–264.

Albani, J.R., 2008. Fluorescence spectroscopy principles: in Principles and applications of fluorescence spectroscopy. John Wiley & Sons. 88-113.

Aleixandre-Tudo, J. L., Buica, A., Nieuwoudt, H., Aleixandre, J. L. and du Toit, W., 2017. Spectrophotometric analysis of phenolic compounds in grapes and wines. Journal of Agricultural and Food Chemistry, 65(20), 4009–4026.

Aleixandre-Tudo, J. L., Nieuwoudt, H., Olivieri, A., Aleixandre, J. L. and du Toit, W., 2018. Phenolic profiling of grapes, fermenting samples and wines using UV-Visible spectroscopy with chemometrics. Food Control, 85, 11–22.

Alpaydin, E., 2020. Introduction to Machine Learning. 4th edition. Cambridge, Massachusetts: MIT press.

Andersen, C. M. and Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. Journal of Chemometrics, 17(4), 200–215.

Bahram, M., Bro, R., Stedmon, C. and Afkhami, A., 2006. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. Journal of Chemometrics: A Journal of the Chemometrics Society, 20(3-4), 99-105.

Beaver, C. W. and Harbertson, J. F., 2016. Comparison of multivariate regression methods for the analysis of phenolics in wine made from two vitis vinifera cultivars. American Journal of Enology and Viticulture, 67(1), 56-64.

Bridgeman, J., Baker, A., Brown, D. and Boxall, J., 2015. Portable LED fluorescence instrumentation for the rapid assessment of potable water quality. Science of the Total Environment, 524, 338–346.

Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade, S.R., Rolle, L. and Tomasi, D., 2015. Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. Computers and Electronics in Agriculture, 117, 186-193.

Bro, R., 1997. PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems, 38(2), 149-172.

Burin, V. M., Arcari, S.G., Luzia, L., Costa, F. and Bordignon-luiz, M. T., 2011. Determination of some phenolic compounds in red wine by RP-HPLC: method development and validation. Journal of chromatographic science, 49(8), 647-651.

Cabrera-Bañegil, M., Hurtado-Sánchez, M., Galeano-Díaz, T.and Durán-Merás, I., 2017. Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples. Food Chemistry, 220, 168–176.

Cabrera-Bañegil, M., Valdés-Sánchez, E., Moreno, D., Airado-Rodríguez, D. and Durán-Merás, I., 2019. Front-face fluorescence excitation-emission matrices in combination with three-way chemometrics for the discrimination and prediction of phenolic response to vineyard agronomic practices. Food Chemistry, 270, 162–172.

Cadot, Y., Miñana-Castelló, M.T. and Chevalier, M., 2006. Anatomical, histological, and histochemical changes in grape seeds from Vitis vinifera L .cv Cabernet franc during fruit development. Journal of Agricultural Food Chemistry, 54(24), 9206–9215.

Chen, T. and Guestrin, C., 2016. XGBoost : A scalable tree boosting system: in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

Cheynier, V., Salas, E., Souquet, J., Sarni-Manchado, P. and Fulcrand, H., 2006. Structure and properties of wine pigments and tannins. American Journal of Enology and Viticulture, 57(3), 298–305.

Christensen, J., Becker, E. M. and Frederiksen, C. S., 2005. Fluorescence spectroscopy and PARAFAC in the analysis of yogurt. Chemometrics and Intelligent Laboratory Systems, 75(2), 201–208.

Christensen, J., Povlsen, V. T. and Sørensen, J., 2003. Application of fluorescence spectroscopy and chemometrics in the evaluation of processed cheese during storage. Journal of Dairy Science, 86(4), 1101–1107.

Cliff, M. A., King, M. C. and Schlosser, J., 2007. Anthocyanin, phenolic composition, colour measurement and sensory analysis of BC commercial red wines. Food Research International, 40(1), 92−100.

Coelho, C., Aron, A., Roullier-Gall, C., Gonsior, M., Schmitt-Kopplin, P. and Gougeon, R., 2015. Fluorescence fingerprinting of bottled white wines can reveal memories related to sulfur dioxide treatments of the must. Analytical Chemistry, 87(16), 8132–8137.

Commission Internationale de l'Eclairage., 1978. CIE recommendations on uniform color spaces, color-difference equations, psychometric color terms, (15th ed., Supplement No. 2). Paris: CIE Publication, Bureau Central de la CIE.

Dambergs, R. G., Mercurio, M. D., Kassara, S., Cozzolino, D. and Smith, P. A., 2012. Rapid measurement of methyl cellulose precipitable tannins using ultraviolet spectroscopy with chemometrics: Application to red wine and inter-laboratory calibration transfer. Applied Spectroscopy, 66(6), 656-664.

Daniel, C., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules, 20(1), 726–737.

De Beer, D., Harbertson, J. F., Kilmartin, P. A., Roginsky, V., Barsukova, T., Adams, D. O. and Waterhouse, A. L., 2004. Phenolics: a comparison of diverse analytical methods. American Journal of Enology and Viticulture, 55(4), 389−400.

Downey, M. O., Harvey, J. S. and Robinson, S. P., 2003. Analysis of tannins in seeds and skins of Shiraz grapes throughout berry development. Australian Journal of Grape and Wine Research, 9(1), 15–27.

Elith, J., Leathwick, J. R. and Hastie, T., 2008. A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802-813.

Fabjanowicz, M., Płotka-Wasylka, J. and Namieśnik, J., 2018. Detection, identification and determination of resveratrol in wine. Problems and challenges. Trends in Analytical Chemistry, 103, 21–33.

Garrido, J. and Borges, F., 2013. Wine and grape polyphenols - A chemical perspective. Food Research International, 54(2), 1844–1858.

Giovenzana, V., Beghi, R., Mena, A., Civelli, R., Guidetti, R., Best, S. and Leòn Gutiérrez, L.F., 2013. Quick quality evaluation of Chilean grapes by a portable VIS/NIR device. Acta Horticulturae, 978, 93-100.

Gishen, M., Dambergs, R. and Cozzolino, D., 2005. Grape and wine analysis - enhancing the power of spectroscopy with chemometrics. Australian Journal of Grape and Wine Research, 11(3), 296-305.

Giusti, M. and Wrolstad, R., 2001. Characterization and measurement of anthocyanins by UV-visible spectroscopy. Current Protocols in Food Analytical Chemistry, (1), F1-2.

Glories, Y., 1984. La couleur des vins rouges, 2eme partie. Connaissance de la Vigne et du Vin, 18, 253−271.

Guilford, J. M. and Pezzuto, J. M., 2011. Wine and Health : A Review. American Journal of Enology and Viticulture, 62(4), 471-486.

Guimet, F., Ferré, J., Boqué, R. and Rius, F. X., 2004. Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation-emission matrix fluorescence spectroscopy. Analytica Chimica Acta, 515(1), 75–85.

Gupta, Y., 2018. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305-312.

Harbertson, J.F., Mireles, M. and Yu, Y., 2015. Improvement of BSA tannin precipitation assay by reformulation of resuspension buffer. American Journal of Enology and Viticulture, 66(1), 95–9.

Harbertson, J. F. and Spayd, S., 2006. Measuring phenolics in the winery. American Journal of Enology and Viticulture, 57(3), 280-288.

Harbertson, J. F., Kennedy, J. A. and Adams, D. O., 2002. Tannin in skins and seeds of Cabernet Sauvignon, Syrah, and Pinot noir berries during ripening. American Journal of Enology and Viticulture, 53(1), 54−59

Ibern-Gómez, M., Andrés-Lacueva, C., Lamuela-Raventós, R.M. and Waterhouse, A.L., 2002. Rapid HPLC analysis of phenolic compounds in red wines. American Journal of Enology and Viticulture, 53(3), pp.218-221.

Iland, P., Ewart, A., Sitters, J., Markides, A. and Bruer, N., 2000. Techniques for chemical analysis and quality monitoring during winemaking (1st ed, 1-111). Campbelltown, Patrick Iland Wine Promotions.

Jensen, J. S., Werge, H. H. M., Egebo, M. and Meyer, A. S., 2008. Effect of wine dilution on the reliability of tannin analysis by protein precipitation. American Journal of Enology and Viticulture, 59(1), 103−105.

Karoui, R. and Blecker, C., 2011. Fluorescence spectroscopy measurement for quality assessment of food systems — a review. Food and Bioprocess Technology, 4(3), 364–386.

Kennedy, J. A., Ferrier, J., Harbertson, J. F. and Peyrot Des Gachons, C., 2006. Analysis of tannins in red wine using multiple methods: correlation with perceived astringency. American Journal of Enology and Viticulture, 57(4), 481−485.

Kennedy, B., 2002. Understanding grape berry development. Practical Winery and Vineyard, (4), 1-5.

Kennedy, J.A., Hayasaka, Y., Vidal, S., Waters, E.J. and Jones, G.P., 2001. Composition of grape skin proanthocyanidins at different stages of berry development. Journal of Agricultural and Food Chemistry, 49(11), 5348–5355.

Lakowicz, J.R., 2013. Instrumentation for fluorescence spectroscopy: in Principles of fluorescence spectroscopy. 3rd edition. Springer Science and Business Media. 27-61.

Letort, A., Laguet, A., Lebecque, A. and Serra, J. N., 2006. Investigation of variety, typicality and vintage of French and German wines using front-face fluorescence spectroscopy. Analytica Chimica Acta, 563, 292–299.

Liakos, K. G., Busato, P., Moshou, D. and Pearson, S., 2018. Machine learning in agriculture : a review, Sensors, 18(8), 2674.

Mercurio, M. D. and Smith, P. A., 2008. Tannin quantification in red grapes and wine: comparison of polysaccharide- and protein-based tannin precipitation techniques and their ability to model wine astringency. Journal of Agricultural and Food Chemistry, 56(14), 5528−5537.

Mercurio, M. D., Dambergs, R. G., Herderich, M. J. and Smith, P. A., 2007. High throughput analysis of red wine and grape phenolics − adaptation and validation of methyl cellulose precipitable tannin assay and modified somers color assay to a rapid 96 well plate format. Journal of Agricultural and Food Chemistry, 55(12), 4651−4657.

Monagas, M., Bartolomé, B. and Gómez-Cordovés, C., 2005. Updated knowledge about the presence of phenolic compounds in wine. Critical reviews in food science and nutrition, 45(2), 85–118.

Murphy, K.R., Stedmon, C.A., Graeber, D. and Bro, R., 2013. Fluorescence spectroscopy and multi-way techniques. PARAFAC. Analytical Methods, 5(23), 6557-6566.

Parker, C.A., 1968. Apparatus and experimental methods. In: Parker, C.A. (Ed.), Photoluminescence of Solutions with Applications to Photochemistry and Analytical Chemistry, 128–302.

Pascual, O., González-Royo, E., Gil, M., Gómez-Alonso, S., García-Romero, E., Canals, J.M., Hermosín- Gutiérrez, I. and Zamora, F., 2016. Influence of grape seeds and stems on wine composition and astringency. Journal of Agricultural and Food Chemistry, 64(34), 6555–6566.

Peleg, H., Gacon, K., Schlich, P. and Noble, A.C., 1999. Bitterness and astringency of flavan-3-ol monomers, dimers and trimers. Journal of the Science of Food and Agriculture, 79(8), 1123–1128.

Pérez-Magariño, S. and González-San José, M. L., 2004. Evolution of flavanols, anthocyanins, and their derivatives during the aging of red wines elaborated from grapes harvested at different stages of ripening. Journal of Agricultural and Food Chemistry, 52(5), 1181–1189.

Quaglieri, C., Jourdes, M., Waffo-Teguo, P. and Teissedre, P.L., 2017. Updated knowledge about pyranoanthocyanins: Impact of oxygen on their contents, and contribution in the winemaking process to overall wine color. Trends in Food Science and Technology, 67, 139-149.

Ribéreau-Gayon, P. and Stonestreet, E., 1965. Le dosage des anthocyannes dans Je vin rouge. Bulletin de La Société Chimique de France, 9, 2649-2652.

Ribéreau-Gayon, P., Glories, Y., Maujean, A. and Dubourdieu, D., 2006. The chemistry of wine. Stabilization and treatments. Handbook of Enology. Volume 2. John Wiley & Sons.

Romera-fernández, M. Berrueta, L. A., Garmón-lobato, S., Gallo, B., Vicente, F. and Moreda, J. M., 2012. Feasibility study of FT-MIR spectroscopy and PLS-R for the fast determination of anthocyanins in wine. Talanta, 88, 303–310.

Sacchi, K. L., Bisson, L. F. and Adams, D. O., 2005. A review of the effect of winemaking techniques. American Journal of Enology and Viticulture, 56(3), 197–206.

Sarneckis, C. J., Dambergs, R. G., Jones, P., Mercurio, M., Herderich, M. J. and Smith, P. A., 2006. Quantification of condensed tannins by precipitation with methyl cellulose: development and validation of an optimized tool for grape and wine analysis. Australian Journal of Grape and Wine Research, 12(1), 39−49.

Schwarz, M., Picazo-Bacete, J.J., Winterhalter, P. and Hermosín-Gutiérrez, I., 2005. Effect of copigments and grape cultivar on the color of red wines fermented after the addition of copigments. Journal of Agricultural and Food Chemistry, 53, 8372–8381.

Singleton, V. L., and Rossi, J. A., 1965. Colorimetry of total phenolics with phosphomolybdic-phosphotungstic acid reagents. American Journal of Enology and Viticulture, 16(3), 144−158.

Somers, T. C. and Evans, M. E., 1974. Wine quality: correlations with colour density and anthocyanin equilibria in a group of young red wines. Journal of the Science of Food and Agriculture, 25(11), 1369−1379.

Somers, T C. and Evans, M. E., 1977. Spectral evaluation of young red wines: anthocyanin equilibria, total phenolics, free and molecular SO2, “chemical age”. Journal of the Science of Food and Agriculture, 28(3), 279−287.

Strasburg, G. M. and Ludescher, R. D., 1995. Theory and applications of fluorescence spectroscopy in food research. Trends in Food Science and Technology, 6(3), 69-75.

Varmuza, K. and Filzmoser, P., 2016. Introduction to multivariate statistical analysis in chemometrics. CRC Press.

Vidal, S., Francis, L., Guyot, S., Marnet, N., Kwiatkowski, M., Gawel, R., Cheynier, V., and Waters, E., 2003. The mouth-feel properties of grape and apple proanthocyanidins in a wine-like medium. Journal of the Science of Food and Agriculture, 83(6), 564–573.

Zhang, Z., Mayer, G., Yves, D., Plazzi, G., and Pizza, F., 2018. Exploring the clinical features of narcolepsy type 1 versus narcolepsy type 2 from European Narcolepsy Network database with machine learning. Scientific Reports, 8(1), 1-11.

Chapter 3

Research Chapter

The direct quantification of red wine phenolics using fluorescence spectroscopy with chemometrics

ABSTRACT

Phenolic compounds are secondary metabolites known to play crucial roles in important chemical reactions impacting the mouthfeel, colour and ageing potential of red wine. Their complexity has resulted in numerous equally complex analytical methods with several drawbacks which often prevent routine phenolic analysis in winemaking. Fluorescence spectroscopy is a rapid, affordable and user-friendly alternative and the combination with chemometrics was investigated for its suitability in directly quantifying phenolic content of red wine and fermenting samples. Front-face fluorescence was optimised and used to build predictive models for total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments. Machine learning algorithms were used for model development and the most successful models were built for total phenols, total condensed tannins and total anthocyanins with coefficient of correlation values (R^2cal) of 0.81, 0.89 and 0.80, respectively. A novel approach for the classification of South African red wine cultivars based on unique fluorescent fingerprints was also successful.

3.1 INTRODUCTION

Phenolic compounds are a diverse group of secondary metabolites found in grapes and wine that can be classified into two families; flavonoids (flavonols, flavan-3-ols and anthocyanins) and non-flavonoids (phenolic acids and stilbenes) (Garrido and Borges, 2013; Aleixandre-Tudo *et al*., 2018). The final phenolic composition of a wine is dependent on numerous factors including viticultural aspects influencing grape berry development and ripening, the grape cultivar and chemical composition at harvest, as well as the winemaking practices implemented throughout fermentation and ageing (Garrido and Borges, 2013). Phenolic compounds have been widely studied for their crucial roles in various chemical reactions that greatly impact important wine attributes, such as mouthfeel, colour and ageing potential (Vidal *et al*., 2003; Monagas *et al*., 2005; Aleixandre-Tudo *et al*., 2018).

The complexity and diversity of red wine phenolic compounds has resulted in numerous analysis methods being developed in order to simplify complex phenolic chemistry into the most relevant phenolic information. Several drawbacks, including expensive equipment and reagent costs as well as the need for trained personnel, prevent the routine analysis of important phenolic parameters during winemaking. The basic spectrophotometric methods most often used are UV-Vis based and rely on the spectral properties of the aromatic ring present in all phenolic compounds, allowing for differentiation between phenolic groups according to characteristic wavelength peaks (Harbertson and Spayd, 2006; Aleixandre-Tudo

et al., 2017). Alternatives such as high performance liquid chromatography (HPLC) are highly sensitive but rarely used outside of research applications while infrared spectroscopies, specifically Fourier transform, have been reported as suitable in phenolic analysis (Romera-fernández *et al*., 2012; Dambergs *et al*., 2012; Daniel, 2015; Aleixandre-Tudo *et al*., 2018). Spectroscopy combined with chemometrics is becoming increasingly investigated in both academic and industry domains to meet growing demands for rapid, accurate, cost-effective and user-friendly analysis techniques that may be applied on site as well as developed into process monitoring, optimisation and control systems.

Fluorescence spectroscopy has been widely used in chemistry and biochemistry disciplines due to its success in analysing the structures, functions and reactivity's of numerous compounds, thereby allowing it to become an important tool in the authentication and quality control of many food science disciplines (Strasburg and Ludescher, 1995). The advantages of fluorescence spectroscopy include being non-destructive, user-friendly, cost effective and highly sensitive when compared to other spectrophotometric methods (Strasburg and Ludescher, 1995; Airado-Rodríguez *et al*., 2011; Karoui and Blecker, 2011). The fluorescent capabilities of the complex wine matrix has been investigated with polyphenols being identified as the largest concentration of naturally occurring fluorophores (Airado-Rodríguez *et al*., 2011). Previous research has been conducted to analyse these fluorescent compounds both qualitatively and quantitatively, with Cabrera-Bañegil *et al*. (2017, 2019) able to quantify pure compounds including catechin, epicatechin, vanillic acid, caffeic acid and resveratrol. Classification tasks have, however, been the focus in wine fluorescence research, with wine authentication according to cultivar, appellation and vintage having been successful (Letort *et al*., 2006; Airado-Rodríguez *et al*., 2011). Understanding the limitations and principles of fluorescence instrumentation is important when conducting analysis, with sample geometry being a major consideration. The conventional right-angled technique traditionally used in fluorescence spectroscopy is used in the analysis of clear or diluted samples. Owing to the complexity of the wine matrix and the chemical interactions taking place within it, as well as the sensitivity of fluorophores to their surrounding environment, a front-face technique developed by Parker (1968) overcomes the need for dilution and allows the analysis of unaltered samples (Airado-Rodríguez *et al*., 2009; Airado-Rodríguez *et al*., 2011; Karoui and Blecker, 2011). Front-face fluorescence therefore presents itself as a potential alternative for the direct and non-invasive analysis of samples during the winemaking process.

Combining spectroscopy with chemometrics (multivariate statistical analysis) holds several advantages including the decomposition and interpretation of complex data sets in a considerably reduced analysis time, its non-destructive nature, and the simultaneous

quantification of several analytes from a single spectral measurement (Gishen *et al*., 2005; Aleixandre-Tudo *et al*., 2018). The most commonly used multi-way techniques in fluorescence analysis have included parallel factor analysis (PARAFAC) as well as unfolded and N-way partial least squares (U-PLS and N-PLS) (Andersen and Bro, 2003; Cabrera-Bañegil *et al*., 2017). Modern machine learning techniques have previously not been investigated in this research area despite their success in complex data handling and ubiquitous use in current technologies.

The need for real-time, rapid, cost-effective and accurate phenolic analysis methods is steadily increasing and routine implementation may aid in the decision-making of winemakers and producers during red wine production. The potential for automation and on-line systems as well as optical portable devices is possible due to the beneficial combination of spectroscopy and chemometrics (Giovenzana *et al*., 2013). The aim of this study was therefore to investigate the suitability of front-face fluorescence spectroscopy to quantify phenolic content of undiluted red wine samples. The five parameters of interest included total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments. Previous wine fluorescence research has, to the best of our knowledge, not investigated the potential of fluorescence spectroscopy to quantify such broad phenolic parameters with a focus on the implications for real-time analysis during the winemaking process. Classification of South African red wine cultivars using fluorescent excitation-emission matrices was also explored for its potential in authentication and quality control.

3.2 MATERIALS AND METHODS

3.2.1 REAGENTS

Ammonium sulphate, hydrochloric acid (HCl 1 M), methyl cellulose, sulphur dioxide (SO_2), ethanol (96%) and sodium metabisulfite (2.5 %) were purchased from Sigma-Aldrich Chemie (Steinheim, Germany). (-)-Epicatechin and malvidin-3-glucoside were purchased from Extrasynthese (Genay, France).

3.2.2 SAMPLES

The collection of 200 fermenting red wine samples took place over the 2019 vintage, following a diverse range of cultivars, vinification practices and terroirs. Both commercial and experimental scale conditions were included, with 91 samples collected from commercial cellars (Stellenbosch University Welgevallen Wine Cellar, Thelema Mountain Vineyards and Kanonkop Wine Estate) and 109 samples collected from the JHN Neethling experimental

cellar at the Department of Viticulture and Oenology (Stellenbosch University). Samples were immediately frozen upon collection. During analysis, samples were thawed and immediately centrifuged at 5000 rpm for 2 min in an Eppendorf 5415D centrifuge (Hamburg, Germany). Additionally, 100 red wine samples from the Agricultural Research Council (ARC Infruitec-Nietvoorbij, Stellenbosch) spanning several vintages (2007-2018) and cultivars were collected, stored at room temperature and centrifuged at 5000 rpm for 2 min on the day of analysis. The cultivars represented in the study, each with varying numbers of samples, included Shiraz (90), Pinotage (49), Cabernet Sauvignon (47), Merlot (36), Malbec (19), Petit Verdot (14), Grenache (9), Pinot noir (9), Tempranillo (5), Cinsaut (4), Arinarnoa (4), a blend (Pinotage, Shiraz and Malbec), Marselan (2), Mourvedre (1), Cabernet Franc (1) and Sangiovese (1).

3.2.3 SPECTROPHOTOMETRIC ANALYSIS

All analyses were conducted with UV-Vis spectroscopy using a Multiskan GO Microplate Spectrophotometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The total phenolics index and total anthocyanin contents were quantified using the methodology reported by Iland *et al*. (2000). One hundred µl of sample supernatant was diluted 50 times with 1 M HCl, vortexed and stored for 1 hour in a dark cupboard before the absorbance between 200-700 nm at 2 nm intervals was recorded. The total phenolics index was calculated as the absorbance at 280 nm multiplied with the dilution factor while total anthocyanin content was calculated in mg/L malvidin-3-glucoside using the absorbance at 520 nm. Total condensed tannin concentration was determined using the methyl cellulose precipitable tannin assay (MCP) protocol developed by Sarneckis (2006) and later modified by Mercurio (2007). In 2 ml microfuge tubes, the treatment involved 50 µl of wine diluted with 600 µl of MCP solution (0.04% w/v), vortexed and left for 2-3 min before 400 µl of ammonium sulphate and 950 µl of distilled water was added. The control tubes contained no MCP solution but rather a total of 1.55 ml distilled water. Both control and treatment stood for 10 min before being centrifuged in an Eppendorf 5415D centrifuge (Hamburg, Germany) at 10 000 rpm for 5 min. The tannin content was then calculated using the difference between control and treatment samples at 280 nm and converted to mg/L using a calibration curve in epicatechin equivalents and a dilution factor of 40. Colour density was determined using the method reported by Glories (1984) whereby 50 µl of wine was analysed against a blank of deionised water and the absorbance recorded at 420 nm, 520 nm and 620 nm. The sum of the three wavelengths was used to determine the colour density of the sample. Polymeric pigments were calculated using the modified Somers assay (Mercurio *et al*., 2007). In 2 ml microfuge tubes, 200 µl of sample supernatant was diluted with 1.8 ml buffer solution (12% v/v ethanol, 0.5 g/L w/v tartaric acid

at pH 3.4) containing 2.5 % sodium metabisulfite, and vortexed. The samples were stored for 1 hour and then analysed at 520 nm. The polymeric or SO_2 resistant pigments were then calculated in absorption units (AU) and a dilution factor of 10.

3.2.4 FLUORESCENCE INSTRUMENTATION

Parameters of a Perkin Elmer LS50B Spectrophotometer were investigated with regards to the intensity, excitation and emission ranges appropriate for wine analysis using diluted samples and conventional fluorescence analysis. A front-face accessory was thereafter investigated to ensure similarly appropriate parameters were obtained, and the optimal angle of incidence was determined as 30 degrees. This calibration from conventional to front-face fluorescence was conducted using a Cabernet Sauvignon wine sample (2018) and validated with a Merlot wine sample (2018) (data not shown).

3.2.5 FLUORESCENCE SPECTROSCOPY

Front-face fluorescence analysis was conducted on all samples at room temperature in a 700 µl quartz cuvette (2 mm width) (Hellma Analytics, Germany) with a 2 cm in diameter aperture fitted in the emission path in order to provide additional filtering of Rayleigh scattering. The excitation-emission matrix (EEM) per sample was recorded as emission spectra between 245 nm and 500 nm at 0.5 nm intervals for excitation wavelengths between 245 nm and 400 nm at 5 nm intervals. Scanning speed was set at 500 nm/min and the excitation and emission slit widths were set at 3 nm and 5 nm, respectively. The UV Winlab instrument software was used for data acquisition.

3.2.6 DATA PRE-PROCESSING

A single, complete dataset containing the combined 289 EEMs was created (11 samples were excluded due to unexplained oversaturation during fluorescence analysis). Once combined, spectral interferences were removed from the EEMs as described by Airado-Rodríguez *et al*. (2011). First and second order Rayleigh scatter were removed by excluding the excitation peaks on the identity line ($\lambda_{ex} = \lambda_{em}$) and at ($2\lambda_{ex} = \lambda_{em}$), respectively. The triangular non-chemical region below the identity line ($\lambda_{ex} > \lambda_{em}$) was set to zero. The software used for data and image processing throughout the study include the open-source web-based user interface JupyterLab (Project Jupyter, USA) using the Python 3 language library scikit-learn (Pedregosa *et al*., 2011) and Matlab version 9.5 (The Mathworks Inc., MA, USA).

3.2.7 CHEMOMETRICS

3.2.7.1 PARALLEL FACTOR ANALYSIS (PARAFAC)

PARAFAC was performed in Matlab using the PLS_Toolbox (The Mathworks Inc., MA, USA) as described in literature (Bro, 1997; Airado-Rodríguez *et al*., 2011; Cabrera-Bañegil *et al*., 2019). The pre-treated EEMs of the 289 samples were stacked in a trilinear arrangement of I x J x K vectors (*samples* x *excitation wavelengths* x *emission wavelengths*) resulting in an initial 289 x 32 x 480 three-dimensional array. Spectral artifacts led to a reduction in EEM size from excitation and emission wavelengths between 245-400 nm and 260-500 nm, to 245-340 nm and 265-500 nm, respectively. The final three-way array of 289 x 20 x 470 was obtained. The appropriate number of components was chosen based on the core consistency diagnostic (CORCONDIA) and explained variance for non-negativity constrained models. Split-half analysis was conducted for model validation. Linear regression was then performed in JupyterLab on the resulting score values to determine univariate calibration models.

3.2.7.2 MACHINE LEARNING

Conventional linear regression in the form of principal component regression and partial least squares regression (PCR and PLSR) were investigated in JupyterLab. The exploration of linear regression included specific region selection based on phenolic fluorescence as found in literature (Airado-Rodríguez *et al*., 2011), data scaling and outlier removal. Machine learning was investigated as a data modelling alternative and an exploration of the optimal pre-processing parameters focused on variance selection, data scaling, spectral region selection and choice of modelling technique. A machine learning pipeline was built in Python consisting of five consecutive steps namely, a column selector used to select for specific columns within the data and allow for spectral region selection between excitation 245-400 nm and emission 245-500 nm, a savgol transform used to apply a Savitzky-Golay filter for data smoothing (Savitzky and Golay, 1964), a pre-processing selector used to find the optimal scaling technique, principal component analysis (PCA) for data decomposition, and the XGBoost regressor to build a tree-based gradient boosted model (Chen and Guestrin, 2016). Bayesian optimisation was used as the framework for automatically tuning the hyper-parameters of the pipeline (Swersky *et al*., 2013; Pelikan *et al*., 1999) and explored over 2 000 iterations and over 160 model configurations per model.

Figure 3.1 is a graphical representation of the machine learning pipeline procedure. Briefly, the data was split into train and test sub datasets, of which 20 samples were retained for model validation. Following this train and test split, a (Synthetic Minority Over-Sampling Technique for Regression) SMOTER algorithm was applied to the training set data. SMOTER makes use of interpolation of target samples identified as extreme cases or within the minority in order to

create synthetic samples that improve upon model training (Torgo *et al*., 2013). A 99% threshold was used, identifying cases within the rare extreme and a k=3 value for k- Nearest Neighbours (KNN) was defined as the interpolation parameter to create the synthetic samples. The training data was thereafter passed through each consecutive step of the pipeline per phenolic parameter, with Bayesian optimisation automatically identifying the best hyper-parameters required for optimal prediction accuracy. Evaluation metrics including coefficient of determination (R^2cal and R^2val), root mean square error (RMSE) and mean absolute error (MAE) were reported for 10-fold cross validation, whereby 10 randomly and equally sized sub datasets were partitioned, retaining 2 samples per sub dataset for internal test validation. RMSE was the key metric used by the Bayesian optimisation algorithm in order to improve upon each new hyper-parameter configuration it explored. The pipeline was repeated a finite number of times and the parameters that resulted in the best cross validated RMSE over all the fits was then used to save a final model configuration. Lastly, the retained 20 sample test dataset was used to evaluate the final model's performance on unseen data.

In order to optimise the pipeline for each phenolic parameter (total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments), four main tests were conducted including running the complete pipeline, the pipeline without synthetic samples, the pipeline with synthetic samples but without region selection and lastly, the pipeline without region selection nor synthetic samples. The optimal pipeline parameters were chosen unique to each phenolic model. Each of the four tests were run several times in order to evaluate the optimal number of components in principal component analysis (PCA). The average train and test scores per number of PCA components were evaluated with a focus on optimal decomposition coupled with model stability. Six components were chosen due to this being consistently optimal for all phenolic models and was thereafter inserted into the pipeline as a fixed hyper-parameter (**Figure 3.1**). Once the optimal parameters were obtained, further model development involved adjusting the phenolic ranges to eliminate minority sample groups from negatively impacting model accuracy, as well as outlier identification and removal.

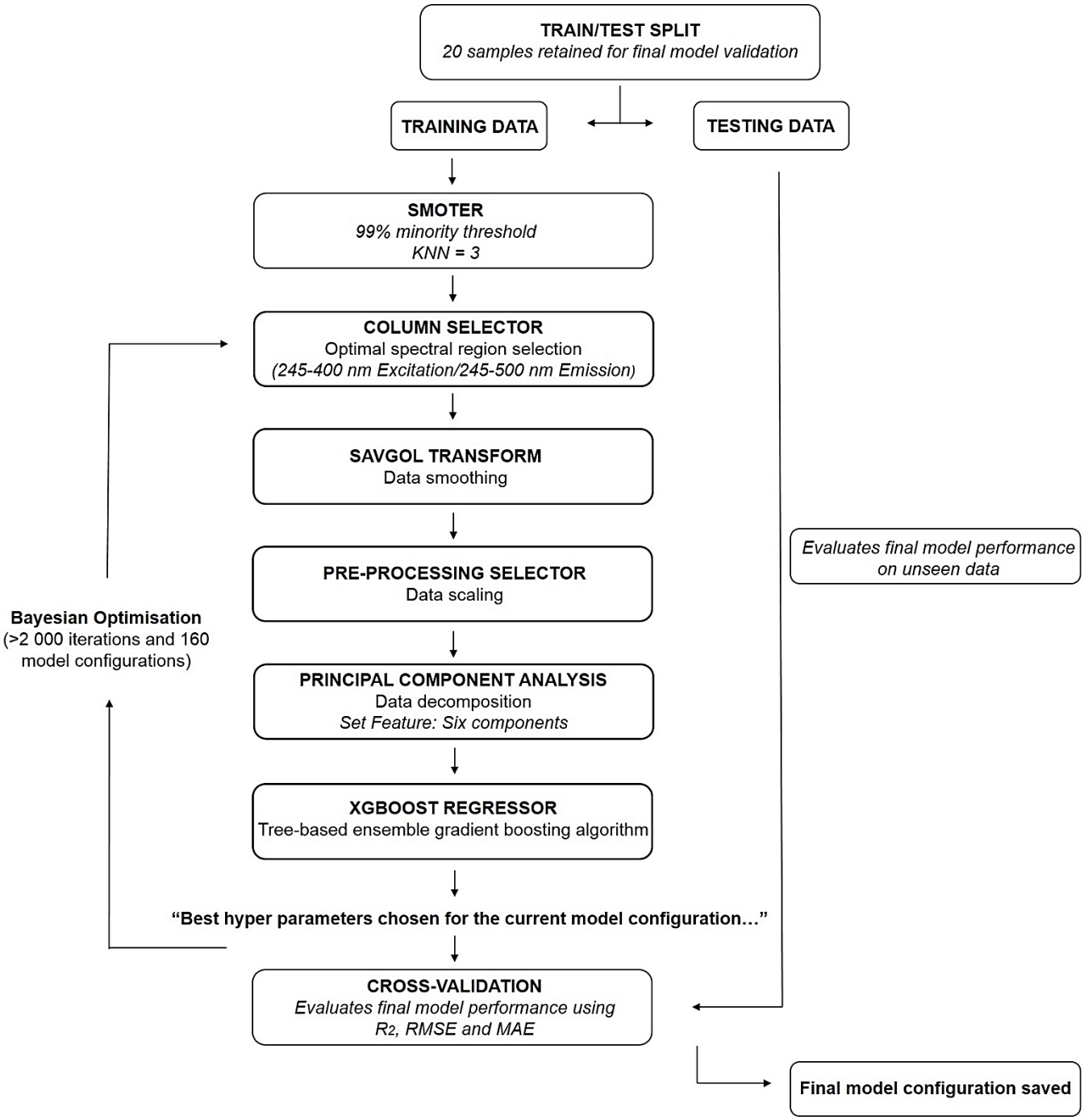


Figure 3.1. Schematic diagram of the machine learning pipeline.

3.2.8 CLASSIFICATION

PARAFAC performed in Matlab, and PCA and neighbourhood component analysis (NCA) performed in Python were the techniques used to evaluate the classification and discrimination abilities of fluorescence spectroscopy. PARAFAC scores obtained per component were plotted against each other (Airado-Rodríguez *et al*., 2011) focusing on the four main cultivar

types included in this study (Cabernet Sauvignon, Merlot, Pinotage and Shiraz) as well the sample state of either fermenting must or wine. PCA was conducted in a similar manner to PARAFAC. NCA was conducted using linear discriminant analysis (LDA) as the linear transformation initialisation method and due to the large variation in number of samples per cultivar, classification was conducted on cultivars with more than or equal to 5, 8, 14 and 20 samples, respectively. NCA was repeated with a focus on classifying according to the sample state of either fermenting must or wine as well as on fermenting musts and wine separately. Leave-one-out cross validation was conducted per set of NCA with score values used to determine classification accuracy.

3.3 RESULTS AND DISCUSSION

3.3.1 WINE EXCITATION-EMISSION MATRICES (EEMS)

Figure 3.2 is an example of a pre-processed EEM belonging to a randomly chosen fermenting Cabernet Sauvignon sample from this study. Two different spectral regions can be observed as a result of the fluorescent properties of wine previously reported in literature (Airado-Rodríguez *et al*., 2009; Airado-Rodríguez *et al*., 2011). Excitation between the more energetic wavelengths of 250 and 290 nm results in emission between 300 and 430 nm, while excitation at wavelengths longer than 300 nm result in emission between 360 and 450 nm (Airado-Rodríguez *et al*., 2009; Airado-Rodríguez *et al*., 2011). **Figure 3.3** is an integrated depiction adapted from literature indicating the characteristic excitation and emission wavelengths of important phenolic compounds (Airado-Rodríguez *et al*., 2011). The non-flavonoid family including phenolic acids (cinnamic-like and benzoic-like), phenolic aldehydes and stilbene-like compounds extends between the ranges of excitation 260-330 nm and emission 320-440 nm. Gentisic acid possesses a unique fluorescence in that it deviates further right of the EEM compared to the rest of the non-flavonoids. The flavonoid family is split into two unique regions with flavonols extending between excitation 260-268 nm and emission 370-422 nm, and flavan-3-ols occurring within excitation 278-290 nm and emission 310-360 nm. Apart from polyphenols, other naturally occurring fluorescent compounds occurring within fermenting musts and wine, such as vitamins and amino acids, have previously been reported (Hoenicke *et al*., 2001; Airado-Rodríguez *et al*., 2009). The fluorescent properties of the amino acid tryptophan, as reported by Christensen *et al*. (2006), have been included. **Figure 3.3** is merely an approximate representation as the excitation-emission regions illustrated below are reported for compounds in dilution measured using the conventional right-angled technique, and spectral shifts, band fluctuations and quantum yield changes may occur when measured within the unaltered wine matrix (Airado-Rodríguez *et al*., 2011).

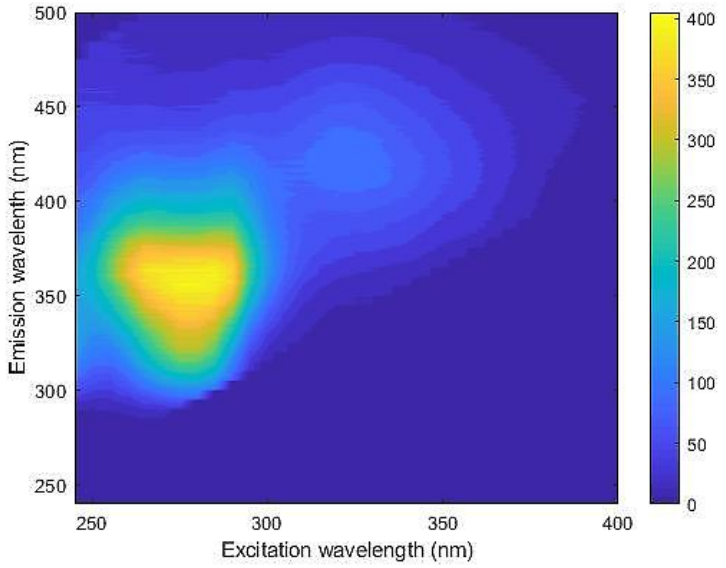


Figure 3.2. Excitation-emission matrix of a fermenting Cabernet Sauvignon sample included in this study (Sample 1) with the scale bar representing fluorescence intensity.

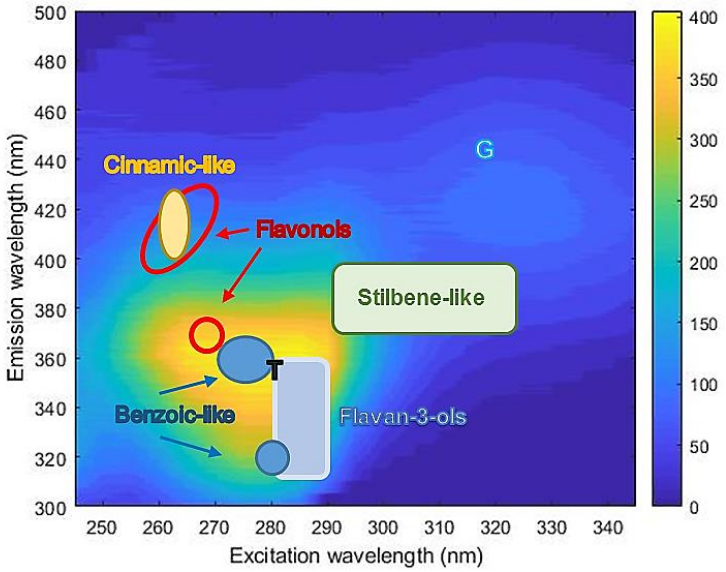


Figure 3.3. Excitation-emission matrix of a fermenting Cabernet Sauvignon sample included in this study (Sample 1) indicating the fluorescent properties of wine fluorophores adapted from literature (Airado-Rodríguez *et al*., 2011). G and T represent gentisic acid and tryptophan, respectively.

Anthocyanins have been reported as weakly fluorescent (quantum yield of 4.1×10^{-3} for malvidin) attributed to their efficient water proton transfer while in the excited state (Agati *et al.*, 2013). Research into anthocyanin fluorescence in red wine has resulted in exponential models being developed for their quantification based on emission ratios of F700/F560 as well as identifying an emission peak of malvidin occurring around 550 nm under excitation 500 nm (Agati *et al*., 2013). A pure malvidin-3-glucoside standard was tested in a model wine dilution series to evaluate the presence of fluorescence within the excitation-emission range chosen

for this study and **Figure 3.4** represents the three-dimensional EEM of pure malvidin-3-glucoside at the highest concentration level (1000 mg/L). The weakly fluorescent intensity below 45 units illustrates the sensitivity of fluorescence spectroscopy but also successfully identifies a potentially overlooked fluorescent region of anthocyanins between excitation 280-300 nm and emission 330-380 nm. It is unclear why this region may not have been explicitly reported in literature, however, it may be a result of limited research into quantitative red wine fluorescence compared to classification and qualitative applications. Alternatively, the UV-Visible absorption properties of anthocyanins involving a characteristic peak between 490 nm and 550 nm (Giusti and Wrolstad, 2001) as well as the accepted grape berry fluorescence indices FERARI (fluorescence excitation ratio anthocyanin relative index) and ANTH_RG (anthocyanin fluorescence index), which involve far red chlorophyll fluorescence (700-780 nm) excited by red light as well as the excitation ratio of red and green light, respectively (Baluja *et al.*, 2012; Pinelli *et al.*, 2018), may not have demanded exploration outside of these accepted wavelengths. Interestingly, Le Moigne *et al.* (2007) obtained good anthocyanin correlations using front-face fluorescence spectroscopy by measuring red grape skins between excitation 250-310 nm at emission 350 nm which correlates well with the fluorescence region identified in **Figure 3.4**.

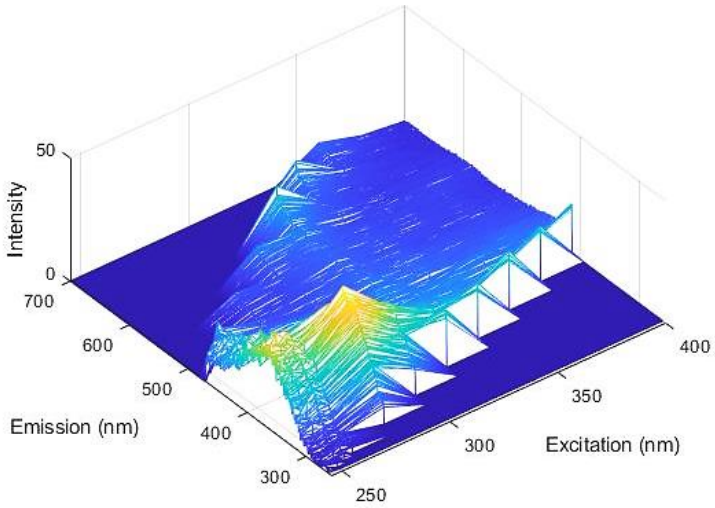


Figure 3.4. Pre-processed excitation-emission matrix mesh plot of pure malvidin-3-glucoside in model wine (1000 mg/L).

Table 3.1. Maximum, minimum, standard deviation and average values per spectrophotometric analysis reference method.

	Total Phenols Index	Total Condensed Tannins (mg/L)	Total Anthocyanins (mg/L)	Colour Density (AU)	Polymeric Pigments (AU)
Maximum	126.10	2912.08	1306.44	42.52	8.09
Minimum	5.11	731.44	9.26	1.89	0.24
Average	44.50	1474.22	350.98	14.01	1.80
Standard deviation	18.02	425.74	194.71	6.06	1.13

Table 3.1 illustrates the phenolic variability achieved during sample collection of both fermenting musts and wine samples. All spectrophotometric methods were performed within a coefficient of variation less than 5%, considered acceptable for reproducibility. The final wine phenolic profile is the result of complex chemical interactions influenced by numerous factors such as those influencing the chemical composition of the grape berry as well as the viticultural and oenological practices implemented throughout processing (Garrido and Borges, 2013). This naturally high variability obtained illustrates the importance of including an extensive dataset during model development in order to sufficiently challenge and train the model on diverse ranges of phenolic levels. Introducing high sample variability aids in building robust calibration models able to make accurate predictions on future samples.

3.3.2. PARALLEL FACTOR ANALYSIS (PARAFAC)

PARAFAC is a trilinear decomposition modelling technique resulting in components (score and loading vectors) that are representative of signals from individual fluorophores. The optimal number of components was chosen to be four, based on the core consistency diagnostic (CORCONDIA) and explained variance for non-negativity constrained models (**Table 3.2**) (Andersen and Bro, 2003) as well as corresponding with results from previous red wine PARAFAC analyses in which components were tentatively correlated with phenolic compounds (Airado-Rodŕiguez *et al*., 2009; Schueuermann *et al*., 2018). Visual inspection of the loadings was performed to confirm the optimal number of components as well as to remove spectral artifacts interfering with the model stability, resulting in a reduced spectral region of 245-340 nm excitation and 265-500 nm emission with a final three-way array of 289 x 20 x 470 (*samples* x *excitation wavelengths* x *emission wavelengths*). Split-half analysis was conducted to validate the uniqueness and stability of the final model.

Table 3.2. Explained variance (%) and core consistency (%) for non-negativity constrained PARAFAC models with one to six components.

	Number of Components					
	1	2	3	4	5	6
Explained Variance (%)	100	95.46	97.74	99.03	99.32	99.44
Core Consistency (%)	87.34	95.39	37.67	88.47	43.70	34.18

Figure 3.5 shows the final model scores (Mode 1) obtained per sample as well as the excitation and emission loadings per PARAFAC component. Score values are estimates of the relative concentrations of the responsible fluorophore and can be used to build univariate calibration models or determine relationships contained within the fluorescence information for potential clustering (Andersen and Bro, 2003; Airado-Rodŕiguez *et al.*, 2009). Components 1 to 4 have been tentatively assigned to their responsible fluorophores in literature by correlating the resulting PARAFAC component excitation and emission peaks with HPLC measurements and bibliographic information (Airado-Rodríguez *et al.*, 2011; Schueuermann *et al.*, 2018). Component 1 is characterised by an excitation peak around 260 nm with an emission shoulder at 370 nm and peak at 390 nm, and has been suggested as representing phenolic aldehydes, benzoic-like acids, myricetin and trans-resveratrol (Airado-Rodríguez *et al.*, 2011) and caffeic acid (Schueuermann *et al*., 2018). Component 2 is characterised by an excitation peak around 280 nm and emission peak around 320 nm. This second component has been consistently matched with monomeric flavan-3-ols, catechin and epicatechin, with high correlations achieved for catechin (R^2 = 0.9221) and epicatechin (R2 = 0.8761) (Airado-Rodŕiguez *et al.*, 2009) as well as the sum of both (R^2 = 0.8468) (Cabrera-Bañegil *et al.*, 2017). Component 3 consists of an excitation peak between 320-330 nm and an emission peak around 420 nm, while component 4 is characterised by an excitation shoulder at 270 nm and peak at 280 nm with an emission peak at 370 nm. Schueuermann *et al.* (2018) proposed cinnamic-like acids, caffeic and p-coumaric, responsible for component 3 while p-coumaric acid, gentisic acid and stilbene-like non-flavonoids were proposed by Airado-Rodŕiguez *et al.* (2009). Component 4 has been suggested as benzoic-like acids as well as tryptophan (Airado-Rodríguez *et al*., 2011; Schueuermann *et al*., 2018). The complexity of the wine matrix results in PARAFAC components most likely corresponding to several different fluorophores or those within the same chemical group rather than individual compounds. No correlations were found between the obtained score values and the reference data per phenolic parameter (Appendix **Figure 3.1**). Despite the potential for component 2 to be well correlated with total condensed

tannins, the best R^2 value obtained after linear regression was 0.21. In the context of this study, PARAFAC was unsuccessful in building calibrations for such broad phenolic parameters such as total condensed tannins versus the successful correlations achieved for pure compounds of catechin or epicatechin (Airado-Rodŕiguez *et al.*, 2009; Cabrera-Bañegil *et al.*, 2017). The structural similarity of the phenolic classes and difficulty in separating them into their singular structures based on their PARAFAC components may be hindering the predictive ability of regression modelling. Conducting PARAFAC on fermenting musts and wine separately did not improve upon results.

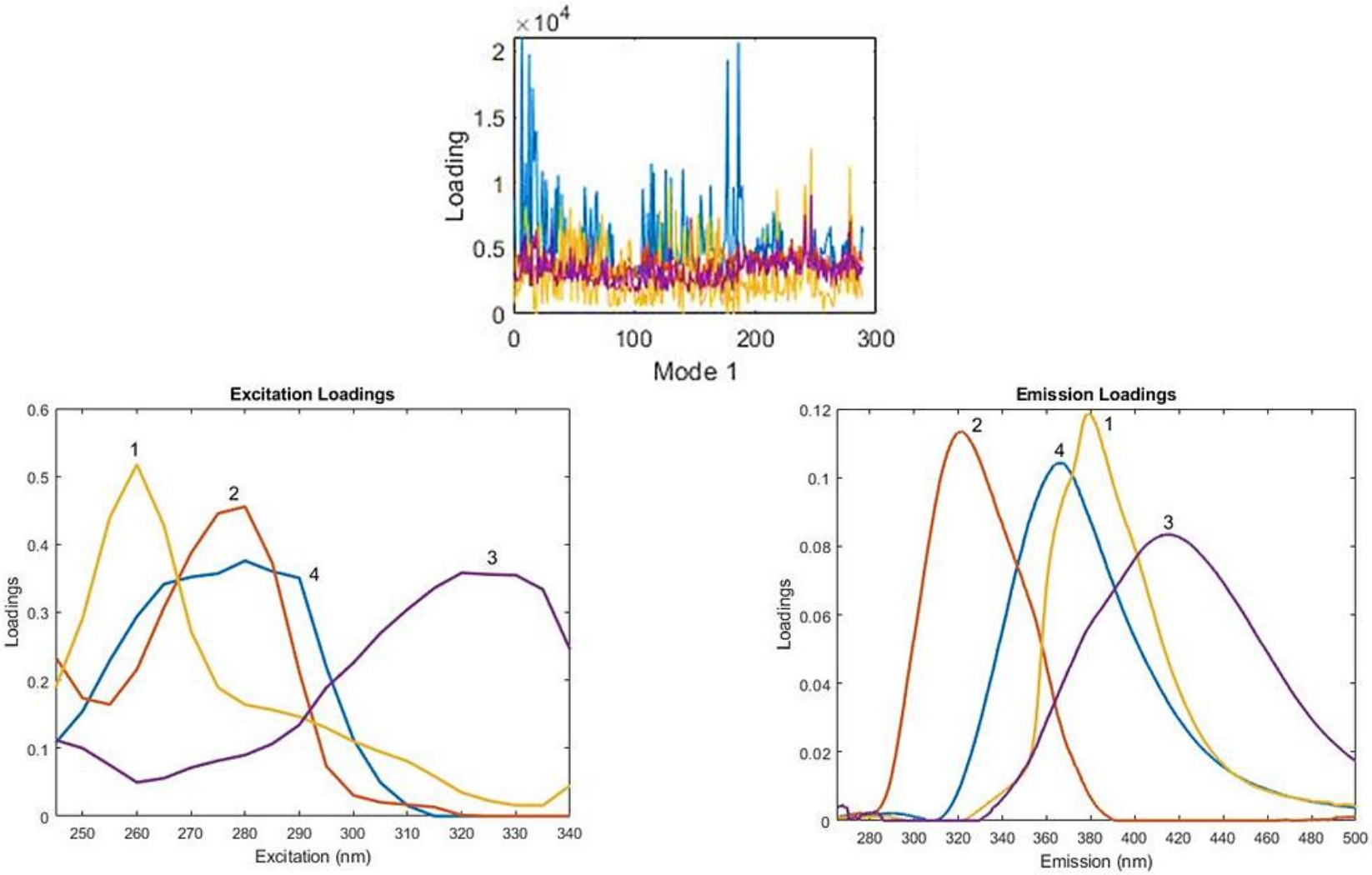


Figure 3.5. Score values per sample (Mode 1), excitation loadings and emission loadings for the four component, non-negativity constrained PARAFAC model.

3.3.3 MACHINE LEARNING

Conventional linear regression in the form of principal component regression (PCR) and partial least squares regression (PLSR) was performed on the fluorescence and reference data. These methods proved unsuccessful despite exploring fluorescent region selection, phenolic range manipulation and outlier removal, with poor calibration and validation scores (data not shown). This suggested a complex dataset requiring more intensive data handling and the exploration of machine learning algorithms. The decision behind using a boosting modelling technique, such as XGBoost, involved the beneficial linear collection of numerous sequentially modelled regression trees rather than a single model of best fit as with simpler regression methods (Chen and Guestrin, 2016). Each successive tree optimises on the residuals of the previous tree's predictions and thereby minimises the loss of predictive ability from previously sub-optimal models (Elith *et al*., 2008; Brillante *et al*., 2015) Gradient boosting is a highly effective technique for classification and regression problems and a favoured option throughout the data science community. This can be seen in the preferred choice of machine learning algorithms used on Kaggle, the largest data science community platform and machine learning competitive scene (Nielsen, 2016).

Briefly, a five-step machine learning pipeline was built consisting of fluorescent region selection, data smoothing and scaling, data decomposition with PCA and lastly, the XGBoost regressor (**Figure 3.1**). The minority over-sampling technique in the form of a SMOTER algorithm applied to the training sub dataset following the train/test split, proved useful in creating a more balanced training model for a widely variable input dataset of fermenting musts and wines. Six principal components showed the most optimal model stability and highest prediction accuracy for all phenolic parameters and was thereafter inserted as a set feature for further model development. Generally, calibration models should be cautiously considered with regards to overly optimistic results. Internal validation in the form of 10-fold cross validation as well as the evaluation of the final model on a retained validation dataset was therefore performed in order to reduce these risks. Each phenolic parameter was individually explored to determine the most optimal pipeline resulting in the highest prediction accuracy and model stability. **Table 3.3** shows the prediction accuracy metrics and characteristics of the best models per phenolic parameter. Once the most optimal pipeline parameters were determined, the pipeline was re-run several times to allow for outlier removal and refinement.

The best total phenols model depicted in **Figure 3.6** (R^2 = 0.81; RMSEV = 7.16; MAEV = 5.39) made use of region selection between 260-360 nm excitation and 370-400 nm emission which overlaps the flavonols and stilbene-like regions as represented above in **Figure 3.3**. Poor

prediction accuracy and unstable models were found when trying to incorporate the entire phenolic region as referenced in literature. Due to a minority of samples with high phenolic values, samples above 80 index units were removed as the model struggled to predict above this threshold.

The best total condensed tannins model (**Figure 3.7**) made use of region selection between 285-340 nm excitation and 290-350 nm emission, overlapping with the flavan-3-ols region depicted in **Figure 3.3**. Samples with tannin levels above 2300 mg/L were removed as the model struggled to predict above this minority group of samples. An R^2 of 0.89, RMSEV of 172.37 and MAEV of 129.14 were obtained. The best total anthocyanins model (**Figure 3.8**) required removing samples with levels above 800 mg/L and made use of region selection between 280-300 nm excitation and 330-380 nm emission which correlates well with the fluorescence of malvidin-3-glucoside identified above (**Figure 3.4**). Prediction scores of R^2 = 0.8, RMSEV = 76.57 and MAEV = 61.57 were obtained. Poorer but stable models were built for colour density (**Figure 3.9**) and polymeric pigments (**Figure 3.10**), the metrics of which are reported in **Table 3.3**. No ideal region could be selected for both models and little improvement was observed with outlier removal and range manipulation. Due to a minority of samples in the higher ranges, samples above 25 absorption units and above 4 absorption units were removed for colour density and polymeric pigments, respectively. The inability to develop a promising regression model for colour density may be due to the characteristics of colour density as a metric. Red wine colour experiences numerous transitions over time as a result of chemical reactions between anthocyanins and other phenolic compounds (Harbertson and Spayd, 2006). The widely used Glories method (1984) is an estimation of total colour by using the sum of absorbances at three wavelengths, namely 420 nm (yellow colouration), 520 nm (red colouration) and 620 nm (blue colouration). The excitation-emission matrix chosen for this study therefore may not have encompassed all responsible compounds provided they possess fluorescent abilities. The poorer prediction accuracy metrics obtained for the polymeric pigments model may also be due to the chosen excitation-emission matrix not encompassing the fluorescent regions of such pigments, as has been identified by the novel fluorescence approach developed using a fluorescence ratio of F700/F560 (Agati *et al*., 2013). However, the unbalanced dataset of 190 fermenting musts and 110 wines may most likely be affecting model calibration due to a minority group of samples with higher polymeric pigment levels (only 40 wine samples with levels above 3 absorption units), the resulting gaps indicated in the regression plot may therefore be negatively affecting the prediction accuracy metrics of an otherwise promising model.

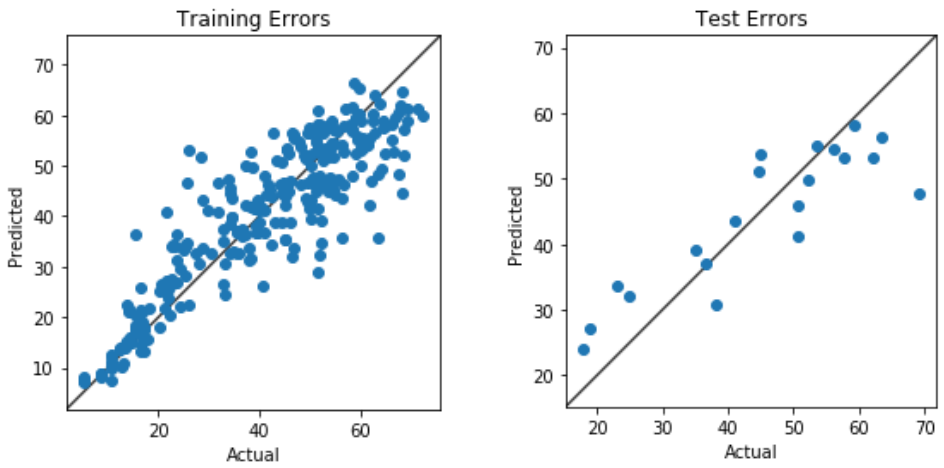


Figure 3.6. Total phenols regression plots, calibration model (left) and validation set (right).

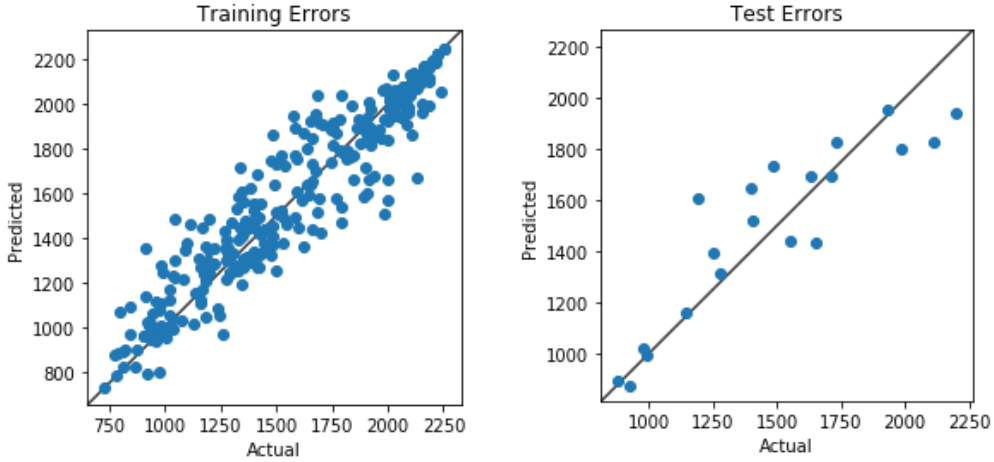


Figure 3.7. Total condensed tannins (mg/L) regression plots, calibration model (left) and validation set (right).

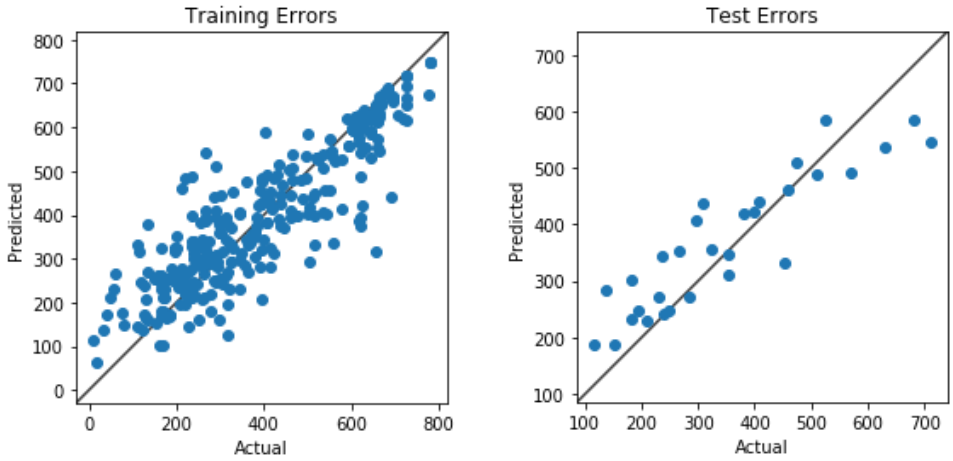


Figure 3.8. Total anthocyanins (mg/L) regression plots, calibration model (left) and validation set (right).

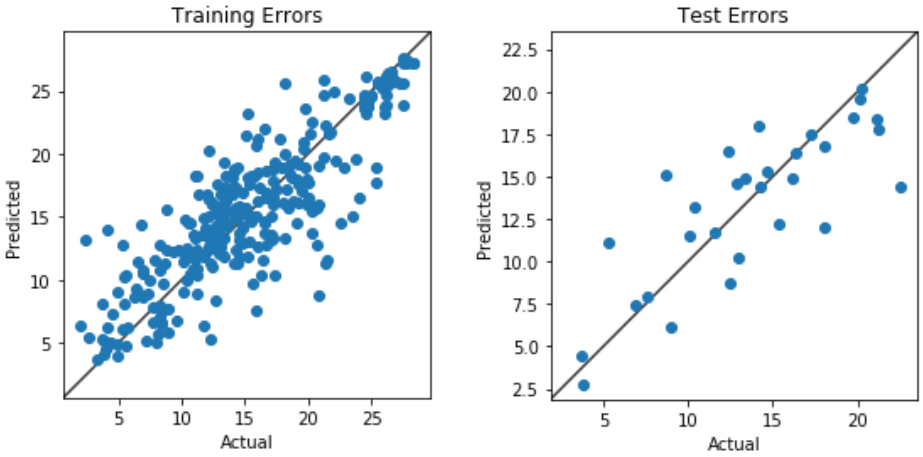


Figure 3.9. Colour density (AU) regression plots, calibration model (left) and validation set (right).

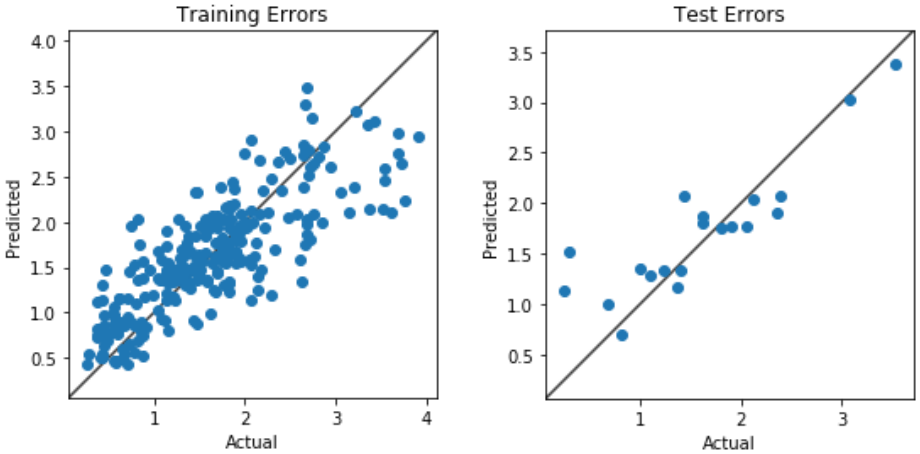


Figure 3.10. Polymeric pigments (AU) regression plots, calibration model (left) and validation set (right).

Cultivar based models were explored per phenolic parameter for the four main cultivars, Cabernet Sauvignon, Shiraz, Merlot and Pinotage. The only model with promising results was built for Cabernet Sauvignon and total condensed tannins with average R^2 train and test scores of 0.78 and 0.81, respectively. This may be a result of high tannin levels characteristic of the cultivar as well as an equally balanced dataset of fermenting musts and wine. Only 45 samples were used in the model and therefore only show promise as to the potential of building a cultivar-based model.

Due to differences in fluorescence between fermenting musts and wine suggested in PCA (**Figure 3.11**), age-based models were explored and the prediction accuracy metrics reported in **Table 3.4**. Overall, models built using only fermenting musts for total phenols, total condensed tannins and polymeric pigments performed slightly better than those built with only finished wines. This could be a result of too few wine samples with too much variability creating large gaps unable to be adequately trained on despite implementing the SMOTER algorithm. The models built using finished wine samples also appear to be more unstable, specifically

with regards to large differences in coefficient of correlation (R^2) between calibration and validation, as seen with total phenols and total condensed tannins (**Table 3.4**). The fermenting-based models for total condensed tannins and polymeric pigments in **Table 3.4** possess slightly better prediction accuracy metrics than the models built using all samples and show potential for classifying other fermenting samples more accurately. Interestingly, the wine-based models built for total anthocyanins and colour density seemed to perform slightly better when looking at RMSE and MAE, however the differences in R^2 should indicate further validation is required. Differences in performance when modelling on fermenting musts and wine separately when compared to the best phenolic models reported in **Table 3.3** may be a result of the random sampling technique used within the machine learning pipeline or the unique behaviour of specialised models built for a specific sub dataset. Overall, the best phenolic parameter models built using all samples may be more promising in terms of generalisability and the ability to predict any sample, regardless of the stage in red wine production, as opposed to more specialised models built for a specific task, such as fermenting or wine based models, which may become over-fitted and perform poorly on unseen data.

Several considerations are important for optimal model development and the acceptance of the subsequently obtained models. Including more samples per cultivar as well as a more balanced dataset of fermenting musts and wine may help in model development. Model considerations include over-fitting and over-validating. Cross validation is incorporated to reduce these risks, however, unidentified noise or influences from the fluorescence spectrophotometer may be fitted on during calibration. Additionally, the retained validation set may potentially be from the same cultivar, the same day of analysis or the same level of fermentation and therefore over confidently validate the model.

Table 3.3. Prediction accuracy metrics (R^2, RMSE and MAE) and pipeline parameters for the best calibration model per phenolic parameter.

	R^2cal	R^2val	RMSEC	RMSEV	MAEV	Excitation/Emission Region (nm)
Total Phenols	0.81	0.77	5.71	7.16	5.39	260-360/370-400
Total Condensed Tannins (mg/L)	0.89	0.80	104.03	172.37	129.14	285-340/290-350
Total Anthocyanins (mg/L)	0.80	0.77	60.67	76.57	61.57	280-300/330-380
Colour Density (AU)	0.68	0.64	2.46	3.10	2.28	-
Polymeric Pigments (AU)	0.64	0.66	0.63	0.49	0.39	-

R^2cal: coefficient of correlation in calibration; R^2val: coefficient of correlation in validation; RMSEC: root mean square error of calibration; RMSEV: root mean square error of validation; MAEV: mean absolute error of validation.

Table 3.4. Prediction accuracy metrics (R^2, RMSE and MAE) for fermenting musts and finished wine calibration models per phenolic parameter.

	R^2cal	R^2val	RMSEC	RMSEV	MAEV
			Total Phenols		
Fermenting	0.70	0.66	6.56	7.45	5.74
Wine	0.74	0.37	3.81	7.77	6.17
			Total Condensed Tannins (mg/L)		
Fermenting	0.82	0.78	95.81	128.24	103.20
Wine	0.69	0.34	122.85	241.13	190.09
			Total Anthocyanins (mg/L)		
Fermenting	0.72	0.77	75.22	89.89	72.18
Wine	0.71	0.55	36.51	60.06	51.28
			Colour Density (AU)		
Fermenting	0.78	0.53	2.65	4.20	3.34
Wine	0.72	0.61	2.03	2.38	2.25
			Polymeric Pigments (AU)		
Fermenting	0.62	0.57	0.27	0.33	0.22
Wine	0.60	0.79	0.49	0.42	0.35

R^2cal: coefficient of correlation in calibration; R^2val: coefficient of correlation in validation; RMSEC: root mean square error of calibration; RMSEV: root mean square error of validation; MAEV: mean absolute error of validation.

3.3.4 CLASSIFICATION

Unique fluorescent fingerprints of wine have been identified for their potential to classify samples based on cultivar type, wine style or appellation (Coelho *et al*., 2015; Airado-Rodríguez *et al*., 2011; Letort *et al*., 2006). The three methods explored in this study for the classification of cultivar type and sample state (fermenting must or wine) included PARAFAC, PCA and NCA. Despite the success of classification in literature, PARAFAC scores in this study were unsuccessful in distinguishing between cultivar or sample state. PCA did not clearly distinguish between cultivars but showed clear distinction between fermenting musts and wine (**Figure 3.11**). NCA was explored due to its success in achieving better classification results compared to other dimensionality reduction techniques, such as PCA and linear discriminant analysis (LDA), as a result of its explicit encouragement of local separation between classes (Goldberger *et al*., 2015). Due to large variation in the number of samples per cultivar, classification was conducted on cultivars with more than or equal to 5, 8, 14 and 20 samples, respectively. Leave-one-out cross validation was conducted per set of NCA analysis with scores reported in **Table 3.5**.

The two best cultivar classification scores were achieved for 9 different cultivars (**Figure 3.12**) and the four main cultivars (**Figure 3.13**). When distinguishing between fermenting musts and wine, the highest cross validation score of 0.82 was achieved for the four main cultivars (**Figure 3.14**), most likely due to a higher number of both sample states compared to other cultivars included in the study. **Figure 3.14** shows wine samples clustered within the lower right-hand side of the plot while the corresponding fermenting musts tend to radiate outwards from this centralised zone. Due to the difference in fluorescence suggested in the stretched appearance of the cultivar classes (**Figure 3.13**) and confirmed with PCA, NCA was conducted on fermenting musts and wines separately. Overall, the cultivar classification ability was stronger for fermenting musts compared to wine (**Table 3.5**). **Figures 3.15** and **3.16** show the best clustering and classification achieved by analysing only fermenting musts. This improved classification for fermenting musts compared to wines highlights the uniqueness of cultivar types before undergoing processing. The final phenolic composition of a wine is a complex chemical matrix influenced by several factors including viticultural practices, different terroirs and various winemaking techniques implemented throughout fermentation and ageing, and therefore clarifies the poorer results for classifying wines purely based on cultivar (Garrido and Borges, 2013; Airado-Rodŕiguez *et al*., 2009). Additionally, the initial composition of grape must may possess higher levels of fluorescent compounds such as vitamins and amino acids before being metabolised by yeast cells during fermentation, while the phenolic composition changes occurring throughout fermentation may also suggest greater fluorescence of monomeric compounds compared to the polymerised compounds found later in wine. Spectral

considerations include a reduced fluorescence intensity from darker samples, the result of which is obtained following increased anthocyanin extraction during fermentation (Hoenicke *et al*., 2001; Airado-Rodríguez *et al*., 2009). Interestingly, the Petit Verdot, Malbec and Shiraz blend (PMS) in **Figure 3.15** is situated relatively central to each of the corresponding pure cultivars included in the fermenting blend and suggests the potential of fluorescence spectroscopy in determining the constituents of blends which may be helpful in authentication and quality control by industry bodies.

Figure 3.16 is an integrated depiction of the highest cross validated cultivar classification for the four main cultivars combined with three-dimensional EEMs of each cultivar. Each sample depicted was chosen based on their phenolic levels to illustrate the unique fluorescent fingerprint per cultivar despite possessing similar phenolic levels (**Table 3.6**). Although showing a similar general three-dimensional fluorescent shape, each cultivar has their own characteristic peak within the EEM and level of fluorescence intensity, with Pinotage having the lowest of the four. Pinotage also exhibits tighter clustering in **Figures 3.12** to **3.16** compared to other cultivars. This may be a result of a particularly unique phenolic composition compared to other cultivars. When investigating the fluorescent intensities of Pinotage samples, more stable fluorescent levels between fermenting musts and wines were observed compared to other cultivars which experienced more extreme variations in fluorescent intensities, the cause of which has not been clearly identified and requires further investigation.

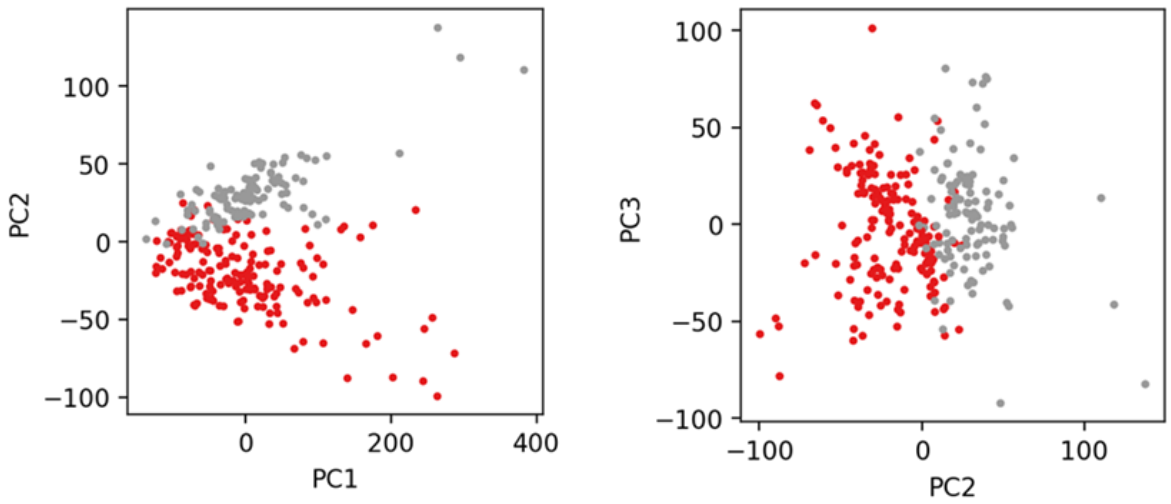


Figure 3.11. Principal Component Analysis (PCA) plot showing fermenting musts (red) and finished wines (grey).

Table 3.5. Leave-one-out cross validation scores per neighbourhood component analysis (NCA) conducted for cultivar classification, sample state classification, fermenting musts and wine classification.

Number of samples per cultivar	Cross Validation Score
Cross validation scores for cultivar classification using all samples	
≥ 5	0.84
≥ 8	0.80
≥ 14	0.72
≥ 20	0.86
Cross validation scores for sample state classification (fermenting musts and wine)	
≥ 5	0.79
≥ 8	0.78
≥ 14	0.77
≥ 20	0.82
Cross validation scores for cultivar classification of fermenting musts only	
≥ 5	0.87
≥ 20	0.93
Cross validation scores for cultivar classification of wine only	
≥ 5	0.76
≥ 20	0.79

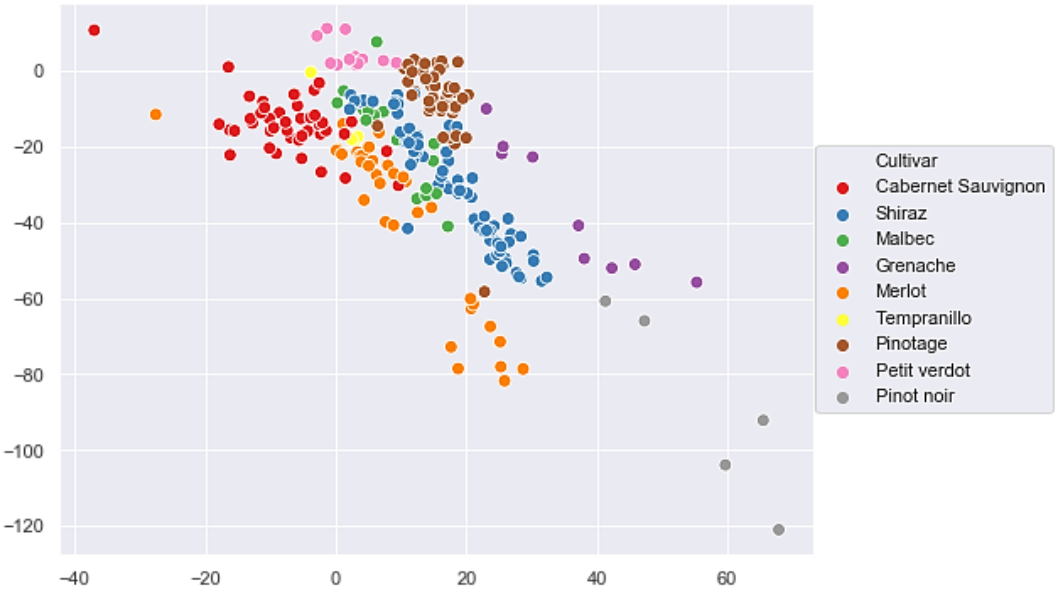


Figure 3.12. Cultivar classification using NCA for cultivars with 5 or more samples (fermenting musts and wine) with a cross validation score of 0.84.

Figure 3.13. Cultivar classification using NCA for cultivars with 20 or more samples (fermenting musts and wine) with a cross validation score of 0.86.

Figure 3.14. Cultivar classification using NCA for the four main cultivars (≥ 20 samples) distinguishing between fermenting musts and wine with a cross validation score of 0.82.

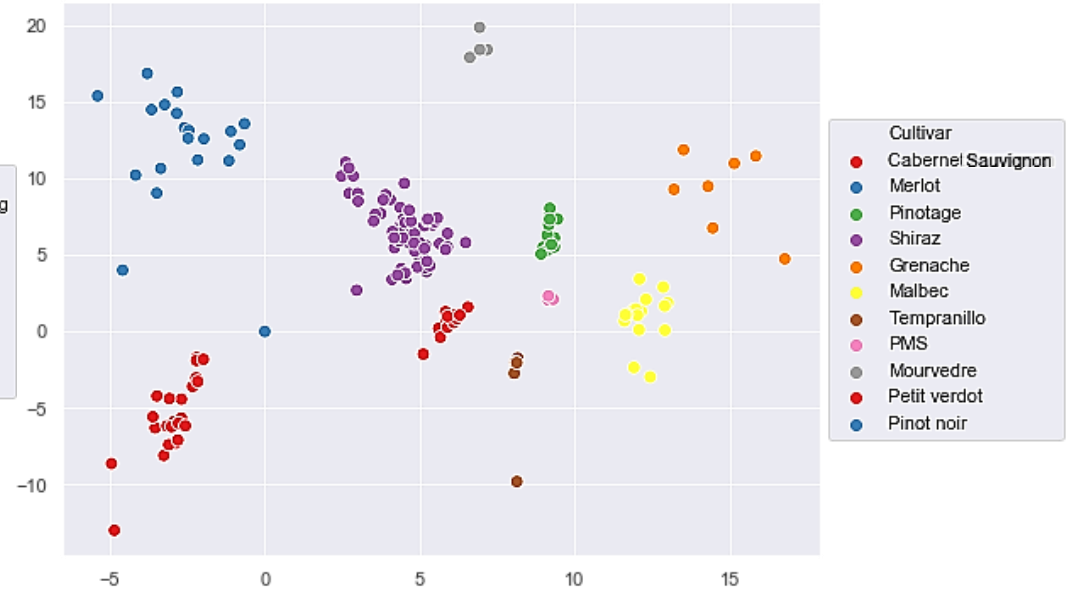


Figure 3.15. Cultivar classification using NCA for cultivars with 5 or more samples on only fermenting musts with a cross validation score of 0.87.

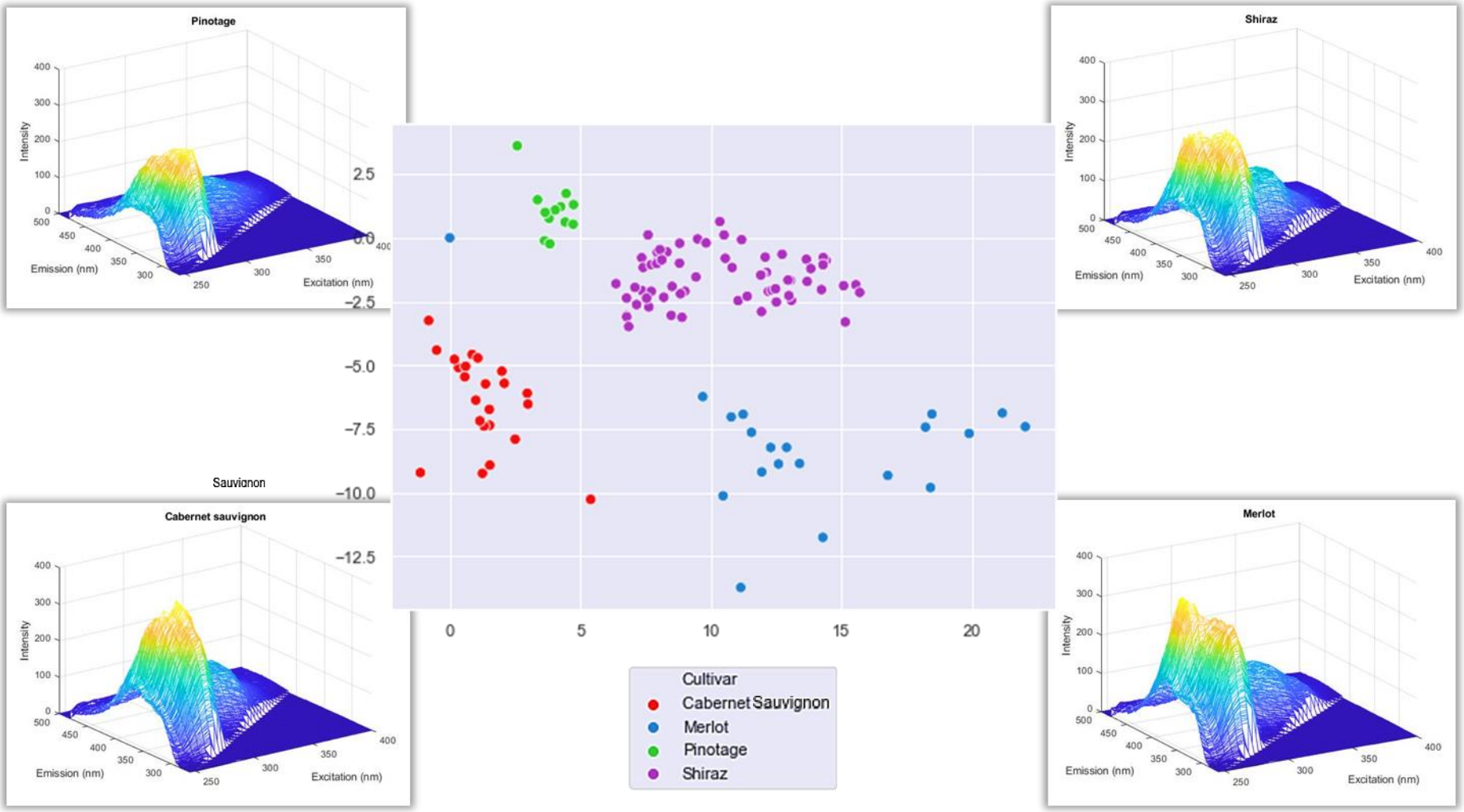


Figure 3.16. Cultivar classification using NCA for the four main cultivars (≥ 20 samples) on only fermenting musts with a cross validation score of 0.93. Three-dimensional excitation-emission matrices of phenolically similar samples corresponding to each cultivar.

Table 3.6. Spectrophotometric analysis measurements showing the phenolic similarity between wines made from different cultivars namely, Merlot, Shiraz, Cabernet Sauvignon and Pinotage (samples 293, 209, 292 and 227).

	Total Phenols	Total Condensed Tannins (mg/L)	Total Anthocyanins (mg/L)	Colour Density (AU)	Polymeric Pigments (AU)
Merlot	59.95	1902.66	304.93	11.02	2.01
Shiraz	59.50	1974.06	324.30	16.50	2.25
Pinotage	59.15	1908.30	313.43	10.71	2.03
Cabernet Sauvignon	60.10	1901.09	231.44	16.67	3.14
Average	59.53	1928.34	314.22	12.74	2.10
Standard deviation	0.33	32.41	7.93	2.66	0.18

3.4 CONCLUSION

Monitoring phenolic extraction throughout fermentation and ageing may aid in decision-making during red wine production. This study showed the potential of front-face fluorescence spectroscopy coupled with chemometrics to quantify important phenolic parameters in fermenting musts and wine. Calibration models built using a gradient boosting technique, XGboost, were successful for the quantification of total phenols, total condensed tannins and total anthocyanins. However, the incorporation of more samples within minority sample groups as well as obtaining a more balanced dataset of different cultivar types, fermenting musts and wines may improve upon model development and therefore the reported results. Additionally, the wide field of chemometrics allows for the use of other statistical analysis methods not explored in this study which may yield better results. The identification of fluorescent regions for each of the phenolic parameters optimises fluorescence analysis for a reduced analysis time and the development of accurate predictive models using front-face fluorescence spectroscopy may allow for their incorporation into future optical portable devices or automated systems, able to analyse samples directly from their fermentation vessels or barrels. Additionally, this study provides a novel approach using NCA for the classification of South African red wine cultivars as well as proposing the potential for analysing and possibly determining the constituents of red wine blends, both of which may be useful in authentication and quality control.

3.5 REFERENCES

Agati, G., Matteini, P., Oliveira, J., de Freitas, V. and Mateus, N., 2013. Fluorescence approach for measuring anthocyanins and derived pigments in red wine. Journal of Agricultural and Food Chemistry, 61(42), 10156-10162.

Airado-Rodŕiguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2009. Usefulness of fluorescence excitation-emission matrices in combination with parafac, as fingerprints of red wines. Journal of Agricultural and Food Chemistry, 57(5), 1711–1720.

Airado-Rodríguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2011. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. Journal of Food Composition and Analysis, 24(2), 257–264.

Aleixandre-Tudo, J. L., Buica, A., Nieuwoudt, H., Aleixandre, J. L. and du Toit, W., 2017. Spectrophotometric analysis of phenolic compounds in grapes and wines. Journal of Agricultural and Food Chemistry, 65(20), 4009–4026.

Aleixandre-Tudo, J. L., Nieuwoudt, H., Olivieri, A., Aleixandre, J. L. and du Toit, W., 2018. Phenolic profiling of grapes, fermenting samples and wines using UV-Visible spectroscopy with chemometrics. Food Control, 85, 11–22.

Andersen, C. M. and Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. Journal of Chemometrics, 17(4), 200–215.

Bro, R., 1997. PARAFAC. Tutorial and applications. Chemometrics and intelligent laboratory systems, 38(2), 149-172.

Cabrera-Bañegil, M., Hurtado-Sánchez, M., Galeano-Díaz, T.and Durán-Merás, I., 2017. Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples. Food Chemistry, 220, 168–176.

Cabrera-Bañegil, M., Valdés-Sánchez, E., Moreno, D., Airado-Rodríguez, D. and Durán-Merás, I., 2019. Front-face fluorescence excitation-emission matrices in combination with three-way chemometrics for the discrimination and prediction of phenolic response to vineyard agronomic practices. Food Chemistry, 270, 162–172.

Chen, T. and Guestrin, C., 2016. XGBoost : A scalable tree boosting system: in proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

Coelho, C., Aron, A., Roullier-Gall, C., Gonsior, M., Schmitt-Kopplin, P. and Gougeon, R., 2015. Fluorescence fingerprinting of bottled white wines can reveal memories related to sulfur dioxide treatments of the must. Analytical chemistry, 87(16), 8132–8137.

Dambergs, R. G., Mercurio, M. D., Kassara, S., Cozzolino, D. and Smith, P. A., 2012. Rapid measurement of methyl cellulose precipitable tannins using ultraviolet spectroscopy with chemometrics: Application to red wine and inter-laboratory calibration transfer. Applied Spectroscopy, 66(6), 656-664.

Daniel, C., 2015. The role of visible and infrared spectroscopy combined with chemometrics to measure phenolic compounds in grape and wine samples. Molecules, 20(1), 726–737.

Garrido, J. and Borges, F., 2013. Wine and grape polyphenols - A chemical perspective. Food Research International, 54(2), 1844–1858.

Giovenzana, V., Beghi, R., Mena, A., Civelli, R., Guidetti, R., Best, S. and Leòn Gutiérrez, L.F., 2013. Quick quality evaluation of Chilean grapes by a portable VIS/NIR device. Acta Horticulturae, 978, 93-100.

Gishen, M., Dambergs, R. and Cozzolino, D., 2005. Grape and wine analysis - enhancing the power of spectroscopy with chemometrics. Australian Journal of Grape and Wine Research, 11(3), 296-305.

Glories, Y., 1984. La couleur des vins rouges, 2eme partie. Connaissance de la Vigne et du Vin, 18, 253−271.

Goldberger, J., Hinton, G.E., Roweis, S. and Salakhutdinov, R.R., 2004. Neighbourhood components analysis. Advances in neural information processing systems, 17, 513-520.

Harbertson, J. F. and Spayd, S., 2006. Measuring phenolics in the winery. American Journal of Enology and Viticulture, 57(3), 280-288.

Karoui, R. and Blecker, C., 2011. Fluorescence spectroscopy measurement for quality assessment of food systems — a review. Food and Bioprocess Technology, 4(3), 364–386.

Letort, A., Laguet, A., Lebecque, A. and Serra, J. N., 2006. Investigation of variety, typicality and vintage of French and German wines using front-face fluorescence spectroscopy. Analytica Chimica Acta, 563, 292–299

Mercurio, M. D., Dambergs, R. G., Herderich, M. J. and Smith, P. A., 2007. High throughput analysis of red wine and grape phenolics − adaptation and validation of methyl cellulose precipitable tannin assay and modified somers color assay to a rapid 96 well plate format. Journal of Agricultural and Food Chemistry, 55(12), 4651−4657.

Monagas, M., Bartolomé, B. and Gómez-Cordovés, C., 2005. Updated knowledge about the presence of phenolic compounds in wine. Critical reviews in food science and nutrition, 45(2), 85–118.

Nielsen, D., 2016. Tree boosting with xgboost - why does xgboost win" every" machine learning competition? Master's thesis, Norwegian University of Science and Technology.

Parker, C.A., 1968. Apparatus and experimental methods. In: Parker, C.A. (Ed.), Photoluminescence of Solutions with Applications to Photochemistry and Analytical Chemistry, 128–302.

Pelikan, M., Goldberg, D.E. and Cantú-Paz, E., 1999. BOA: The Bayesian optimization algorithm. In: Proceedings of the genetic and evolutionary computation conference GECCO-99, 1, 525-532.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnav, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Romera-fernández, M. Berrueta, L. A., Garmón-lobato, S., Gallo, B., Vicente, F. and Moreda, J. M., 2012. Talanta Feasibility study of FT-MIR spectroscopy and PLS-R for the fast determination of anthocyanins in wine. Talanta, 88, 303–310.

Sarneckis, C. J., Dambergs, R. G., Jones, P., Mercurio, M., Herderich, M. J. and Smith, P. A., 2006. Quantification of condensed tannins by precipitation with methyl cellulose: development and validation of an optimized tool for grape and wine analysis. Australian Journal of Grape and Wine Research, 12(1), 39−49.

Savitzky, A. and Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8), 1627-1639.

Schueuermann, C., Silcock, P. and Bremer, P., 2018. Front-face fluorescence spectroscopy in combination with parallel factor analysis for profiling of clonal and vineyard site differences in commercially produced Pinot Noir grape juices and wines. Journal of Food Composition and Analysis, 66, 30–38.

Somers, T. C. and Evans, M. E., 1974. Wine quality: correlations with colour density and anthocyanin equilibria in a group of young red wines. Journal of the Science of Food and Agriculture, 25(11), 1369−1379.

Somers, T C. and Evans, M. E., 1977. Spectral evaluation of young red wines: anthocyanin equilibria, total phenolics, free and molecular SO2, “chemical age”. Journal of the Science of Food and Agriculture, 28(3), 279−287.

Strasburg, G. M. and Ludescher, R. D., 1995. Theory and applications of fluorescence spectroscopy in food research. Trends in Food Science and Technology, 6(3), 69-75.

Swersky, K., Snoek, J. and Adams, R.P., 2013. Multi-task bayesian optimization. Advances in neural information processing systems, 26, 2004-2012.

Torgo, L., Ribeiro, R.P., Pfahringer, B. and Branco, P., 2013. SMOTE for regression. In: Portuguese conference on artificial intelligence, 378-389.

Vidal, S., Francis, L., Guyot, S., Marnet, N., Kwiatkowski, M., Gawel, R., Cheynier, V., and Waters, E., 2003. The mouth-feel properties of grape and apple proanthocyanidins in a wine-like medium. Journal of the Science of Food and Agriculture, 83(6), 564–573.

Chapter 4

Research Chapter

The use of non-invasive fluorescence spectroscopy to quantify phenolic content under real-time fermentation conditions

ABSTRACT

Phenolic compounds play important roles in wine quality attributes such as colour, mouthfeel and ageing potential. The ability to monitor their extraction and implement appropriate vinification techniques relies on accurate phenolic analysis methods. Front-face fluorescence spectroscopy presents itself as a user-friendly, rapid and cost-effective alternative to other spectrophotometric methods and was therefore investigated for its potential in directly measuring phenolic content of red wine samples throughout fermentation. A Cabernet Sauvignon fermentation was monitored using both fluorescence and UV-Vis spectroscopies. Fermentation conditions were explored for their influence on the prediction accuracy of fluorescence based regression models, specifically total phenols, total condensed tannins, total anthocyanins, colour density and polymeric pigments. The stage of fermentation appeared to influence sample fluorescence greater than sample preparation treatment, specifically clean, degassed and unaltered fermenting samples. The coefficient of correlation (R^2cal) for models built using only unaltered samples were above 0.86 for all except colour density. Overall, the ability to analyse unaltered samples directly from the fermentation vessel was possible and holds potential for automated systems or portable device applications. The evolution of fluorescence for Cabernet Sauvignon grape must to final wine was investigated and notable spectral regions identified.

4.1 INTRODUCTION

Red wine production involves alcoholic fermentation taking place in the presence of both solid and liquid phases of the must, resulting in the suspension of grape solids, yeast and various colloidal particles. Phenolic extraction relies on adequate skin-juice contact and various winemaking techniques implemented pre-, post- or during fermentation have been studied for their influence on the resulting red wine phenolic profile (Sacchi *et al*., 2005; Casassa and Harbertson, 2014; Smith *et al*., 2015). These vinification techniques may include the addition of pectolytic enzymes, cap management in the form of pump-overs or punch-downs as well as extended maceration, among others (Sacchi *et al*., 2005). Anthocyanins, flavonols and their subsequently polymerised forms are considered to have the greatest sensory impact on red wine, specifically with regards to important attributes such as mouthfeel, colour and ageing potential (Sacchi *et al*., 2005).

Anthocyanin extraction reaches a maximum early on in fermentation followed by a decline thereafter as a result of co-pigmentation and polymerisation reactions, while condensed tannins experience continued skin-juice extraction with seed tannins increasing linearly

compared to the earlier plateau reached by skin tannins (Canals *et al*., 2005; Sacchi *et al*., 2005; Cadot, *et al*., 2006). Understanding the extraction dynamics of phenolic compounds may aid in implementing timely winemaking practices for the desired effect and therefore requires the routine analysis of these important compounds throughout fermentation. The benefits of fluorescence spectroscopy, including its non-invasive technique, increased sensitivity, rapid and user-friendly action as well as its relative cost-effectiveness when compared to other spectrophotometric methods, have allowed it to become an increasingly popular alternative in various food science disciplines (Karoui and Blecker, 2011; Strasburg and Ludescher, 1995; Airado-Rodríguez *et al.*, 2011). Front-face fluorescence spectroscopy is explored in this study as an alternative to the current spectrophotometric analysis methods used for phenolic analysis.

Understanding fluorescence spectroscopy instrumentation and the factors affecting optimal analysis are essential for collecting accurate and representative spectral information. The electronic transitions taking place during fluorescence analysis, namely the absorption of UV-Visible light, the subsequent redistribution of energy by excited molecules within fluorescent compounds and their detected emitted light, are influenced by several factors such as quenching, the local environment and light scatter phenomena (Strasburg and Ludescher, 1995; Karoui and Blecker, 2011). Higher temperatures during analysis may increase collisional velocity and therefore collisional quenching, resulting in a decreased fluorescence intensity. The local environment including pH changes and sample colour influence the highly sensitive fluorophores, thereby influencing the shape and intensity of the captured fluorescence spectra, and light scatter phenomena such as Rayleigh scattering can be considerably affected in turbid or opaque samples with regards to the optical sampling depth as well as the captured fluorescence signal. The results from analysing diluted samples are not always comparable with those of the original sample, specifically with the matrix of food products significantly affecting intrinsic fluorescent compounds (Airado-Rodríguez *et al*., 2011). The sample geometry of front-face fluorescence eliminates the need for sample dilution as with conventional right-angle fluorescence and allows for the analysis of native samples (turbid, concentrated or solid) owing to the signal captured being independent of the light penetration through the sample (Airado-Rodríguez *et al*., 2011; Karoui and Blecker, 2011). The minimal to no sample preparation required for this technique therefore holds potential for analysing red wine throughout fermentation directly from the fermentation vessel, an application which may be of benefit to the producer in on-line systems or portable devices.

The aim of this study was to investigate the prediction accuracy of the five phenolic regression models built using front-face fluorescence spectroscopy previously in Chapter 3, while exploring the effects of fermentation conditions, specifically the influence of carbon dioxide

and grape solid turbidity, and therefore the required sample preparation in order to successfully analyse red wine samples throughout fermentation directly from the fermentation vessel.

4.2 MATERIALS AND METHODS

4.2.1 REAGENTS

Ammonium sulphate, hydrochloric acid (HCl 1 M), methyl cellulose, sulphur dioxide (SO_2), ethanol (96%) and sodium metabisulfite (2.5 %) were purchased from Sigma-Aldrich Chemie (Steinheim, Germany).

4.2.2 EXPERIMENTAL DESIGN

This study was performed using Cabernet Sauvignon grapes harvested from the 2020 vintage and frozen until processing in the experimental cellar at the Department of Viticulture and Oenology (Stellenbosch University). One crate of grapes was crushed and destemmed into a 20 L plastic bucket and received 50 mg/L sulphur dioxide (SO_2). The must was inoculated with 20 g/hL Zymaflore RX60 (*Saccharomyces cerevisiae*, Laffort, Bordeaux, France) and fermentation took place in a 25 °C temperature controlled room. Two punch-downs were performed per day. Sample collection and analysis took place from the first day of fermentation until the wine had fermented dry (residual sugar < 4 g/L) 12 days later. Although sampled on consecutive fermentation days and UV-Vis spectrophotometric methods conducted daily, fluorescence analysis was performed only on 9 of those days due to logistical reasons.

Following the morning punch-down and homogenous mixing, a representative sample was collected and separated into three 15 ml test tubes. **Figure 4.1** shows the three sample preparation treatments investigated in triplicate, namely clean samples (Treatment A), degassed samples (Treatment B) and unaltered samples (Treatment C). Treatment A involved degassing by vacuum followed by centrifuging at 5000 rpm for 2 min in an Eppendorf 5415D centrifuge (Hamburg, Germany) and subsequently removing the supernatant to inhibit interference of fermentation sediment such as yeast and grape solids. Treatment B involved degassing the samples by vacuum to remove the carbon dioxide (CO_2) within the sample while remaining turbid, and Treatment C experienced no sample preparation, representing sample analysis directly from the fermentation vessel.

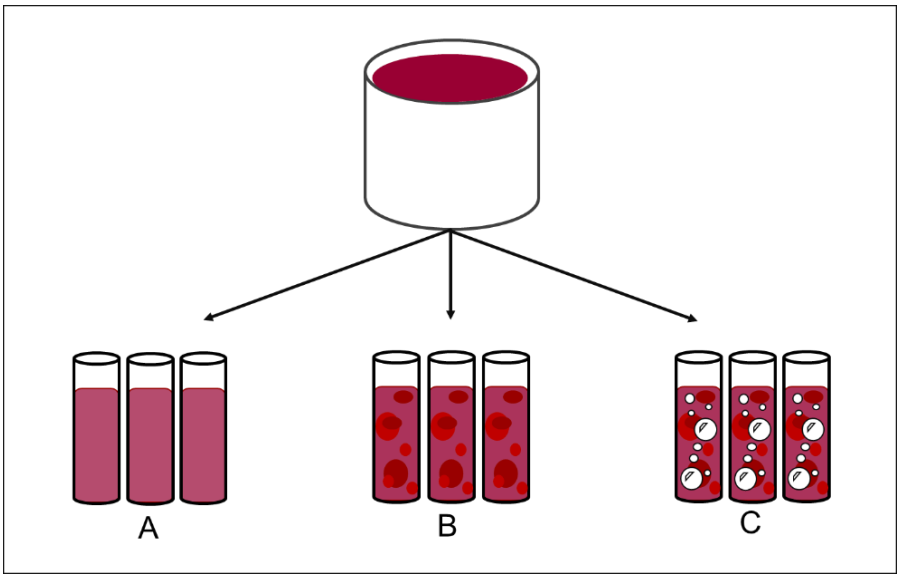


Figure 4.1. Schematic experimental design of sample preparation treatments performed in triplicate. Clean (A), degassed (B) and unaltered (C) samples.

4.2.3 ANALYSIS

4.2.3.1 SPECTROPHOTOMETRIC ANALYSIS

All reference data analysis was performed with UV-Vis spectroscopy using a Multiskan GO Microplate Spectrophotometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The methodology reported by Iland *et al*. (2000) was used to quantify total phenolics and total anthocyanins. One hundred µl of sample supernatant was diluted 50 times with 1 M hydrochloric acid, vortexed and stored in the dark for 1 hour before recording the absorbance at 280 nm and 520 nm, respectively. Total phenolic content was calculated as the absorbance at 280 nm multiplied by the dilution factor while total anthocyanins was calculated in malvidin-3-glucoside equivalents using the absorbance at 520 nm.

The methyl cellulose precipitable tannin assay (MCP) protocol modified by Mercurio (2007) was used to calculate total condensed tannins. The tannin content is calculated using the difference between control and treatment samples and converted into epicatechin equivalents (mg/L) using a calibration curve and dilution factor of 40. The 2 ml microfuge treatment tubes consist of adding 600 µl of MCP solution (0.04% w/v) to 50 µl of wine. After being vortexed and standing for 2-3 min, 400 µl of ammonium sulphate and 950 µl of distilled water are added. The control tubes contain no MCP solution and therefore a total of 1.55 ml distilled water is added. Both control and treatment stand for 10 min before being centrifuged at 10 000 rpm for 5 min and recording the absorbance at 280 nm.

Colour density was calculated as the sum of absorbance at 420 nm, 520 nm and 620 nm wavelengths for a 50 µl sample volume (Glories, 1984). Polymeric pigments were calculated using the modified Somers assay whereby 200 µl of sample supernatant is diluted with 1.8 ml buffer solution (12% v/v ethanol, 0.5 g/L w/v tartaric acid at pH 3.4) containing 2.5 % sodium metabisulfite (Mercurio *et al*., 2007). The samples were stored for an hour before calculating the polymeric pigment content in absorption units (AU) using a dilution factor of 10 and the absorbance at 520 nm.

4.2.3.2 FLUORESCENCE ANALYSIS

Front-face fluorescence analysis of all samples was conducted at room temperature in a 700 µl quartz cuvette (2 mm width) (Hellma Analytics, Germany) using a Perkin Elmer LS50B spectrophotometer. Excitation wavelengths between 245 nm and 400 nm at 5 nm intervals were used to capture emission spectra between 245 nm and 500 nm at 0.5 nm intervals. A 2 cm in diameter aperture was fitted in the emission path for reducing excess light scattering. A scanning speed of 500 nm/min was used and the excitation and emission slit widths were set at 3 nm and 5 nm, respectively. The instrument control and data manipulation software, UV Winlab, was used for data acquisition.

4.2.4 CHEMOMETRICS

4.2.4.1 DATA PRE-PROCESSING

Unwanted spectral signatures were removed using the method described by Airado-Rodríguez *et al.* (2011) whereby first and second order Rayleigh scattering are excluded as the excitation peaks centred on the identity bands ($\lambda_{ex} = \lambda_{em}$) and ($2\lambda_{ex} = \lambda_{em}$), respectively. The triangular region below the identity line ($\lambda_{ex} > \lambda_{em}$) possesses no chemical information and values were therefore inserted as zero. Data and image processing were performed with JupyterLab (Project Jupyter, USA) using the Python 3 language library scikit-learn (Pedregosa *et al*., 2011) and Matlab ver 9.5 (The Mathworks Inc., MA, USA).

4.2.4.2 MODEL VALIDATION

Principal component analysis (PCA) was performed on the dataset to evaluate for differences between sample preparation treatments as well as to determine differences based on the stage of fermentation (early versus late). The regression models built in Python with clean samples and a more extensive sample set for the five phenolic parameters, namely total phenolics, total condensed tannins, total anthocyanins, colour density and polymeric pigments (Chapter 3), were validated using the 81 samples obtained from the above fermentation

experiment. Overall, models built using a well-balanced dataset and large number of both fermenting musts and finished wines may be generally better suited for all applications compared to those built for specific tasks, in this case fermentation-based analysis, which may become over-fitted and predict poorly on new data. Additionally, models built using a more variable dataset may be able to handle the complexity from complex environments such as with degassed or unaltered samples. The data was passed into each phenolic model to determine the prediction accuracy for several different dataset configurations. These sub datasets investigated day of fermentation (all treatment samples for the entire fermentation, day 1-3 treatment samples and day 5-12 treatment samples) and subsequently the three treatments (clean (A), degassed (B) and unaltered (C)). The metrics used to determine prediction accuracy included root mean square error (RMSE) and mean absolute error (MAE). MAE weights all errors equally while RMSE gives errors with larger absolute values more weight than errors with smaller absolute values. Both metrics are regularly used in model evaluation and there is often little consensus when deciding on the most suitable metric, therefore the combination of both allows for improved understanding of different data projections and characteristics of model performance (Chai and Oceanic, 2014).

In order to evaluate the suitability of the hyper-parameters chosen during initial model development as well as the effect of sample preparation on future modelling, the fermentation data was separately passed through the machine learning pipeline and modelled using the previously optimised parameters identified per phenolic parameter model. Briefly, the fermentation data was split into train and test sub datasets, of which 10 samples were retained as the test validation set. Thereafter, the training data was passed through the five consecutive steps of the pipeline including a column selector for optimised spectral region selection, a savgol transform used to apply a Savitzky-Golay filter for data smoothing (Savitzky and Golay, 1964), a pre-processing selector for optimal data scaling, six-component PCA for data decomposition, and lastly, the XGBoost regressor to build a tree-based gradient boosted model (Chen and Guestrin, 2016). The total phenolics model consisted of region selection between 260-360 nm excitation and 370-400 nm emission, the total condensed tannins model made use of region selection between 285-340 nm excitation and 290-350 nm emission and the total anthocyanins model involved region selection between 280-300 nm excitation and 330-380 nm emission, all of which were previously identified as optimal spectral regions (Chapter 3). The metrics used to determine prediction accuracy included coefficient of determination (R^2cal and R^2val), root mean square error (RMSE) and mean absolute error (MAE). Bayesian optimisation was the framework used for the automatic tuning of the other pipeline hyper-parameters such as data scaling and smoothing (Swersky *et al*., 2013; Pelikan *et al*., 1999). Once passed through the pipeline, 2-fold cross-validation was performed due to

the smaller input dataset, with the reported RMSE used as the key metric for Bayesian optimisation and the sequential improvement on previously chosen hyper-parameters. The best final model was evaluated using the previously retained 10 sample test set as a form of external validation.

4.3 RESULTS AND DISCUSSION

4.3.1 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA was conducted on the excitation-emission matrices (EEMs) of the 81 samples collected throughout fermentation. **Figure 4.2** shows the evolution of fluorescence within the fermenting must as the fermentation proceeds, with early fermenting samples (days 1-3) being clustered separately to those of later fermenting samples (days 5-12). This confirms previous findings involving the difference in fluorescence between fermenting musts and wine, while further highlighting the unique fluorescent changes taking place within a single fermentation vessel. Without having fluorescent information for day 4 of the fermentation, the exact moment in which the fluorescence evolves from characteristically being early on versus later in fermentation is unknown. The clear separation between classes, however, may indicate a threshold, potentially the result of maximum plateaued anthocyanin extraction and the subsequent reabsorption of light from a darker sample matrix reached early on in fermentation. **Figure 4.3** shows PCA based on sample preparation treatment, however, no clear distinction between the treatments is found and may indicate that the stage of fermentation has a greater effect on the fluorescent information obtained than sample preparation.

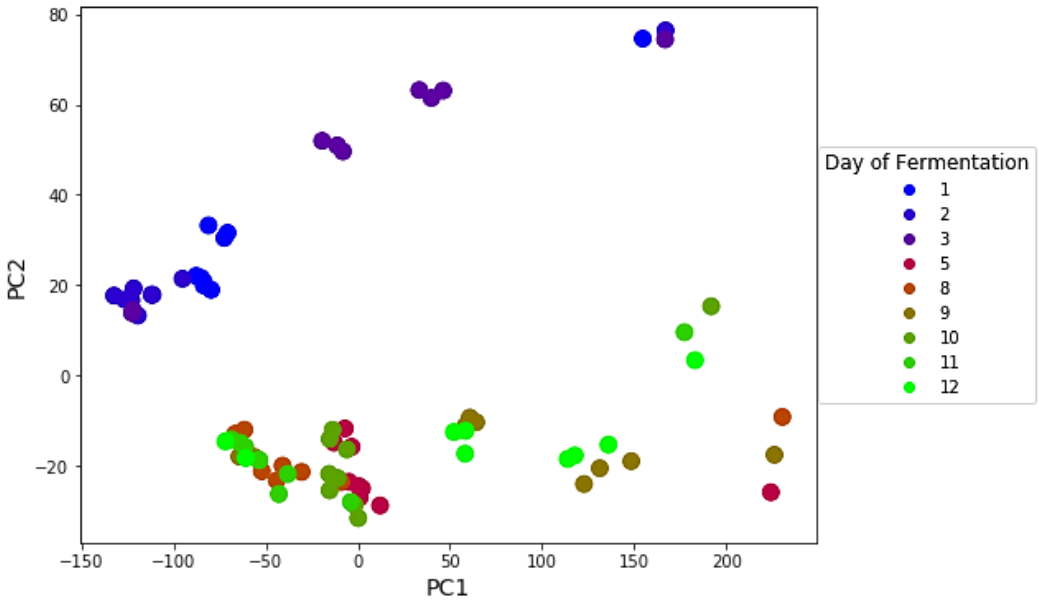


Figure 4.2. PCA plot showing fermenting samples based on day of fermentation.

Figure 4.3 can be considered a visual representation of the benefits of front-face fluorescence spectroscopy, whereby the changed sample geometry allows for the analysis of samples in their natural state in order to retain the influence of the surrounding matrix on highly sensitive fluorophores (Airado-Rodríguez *et al*., 2011; Karoui and Blecker, 2011). The scattered appearance of the samples analysed in triplicate may indicate the heightened sensitivity of fluorescence as a spectrophotometric method (Strasburg and Ludescher, 1995). Although all samples were analysed at room temperature and pipetted into the cuvette as homogenous solutions, other influencing factors must be considered such as the varying rates at which the turbidity settles out in the cuvette over a 25 minute analysis time, the occurrence of which may mimic the turbidity changes occurring naturally during fermentation, as well as the time taken for the analysis of all samples thereby influencing potential instrumental drift or changes in lamp intensity and heating (Andersen and Bro, 2003; Airado-Rodŕiguez *et al*., 2009).

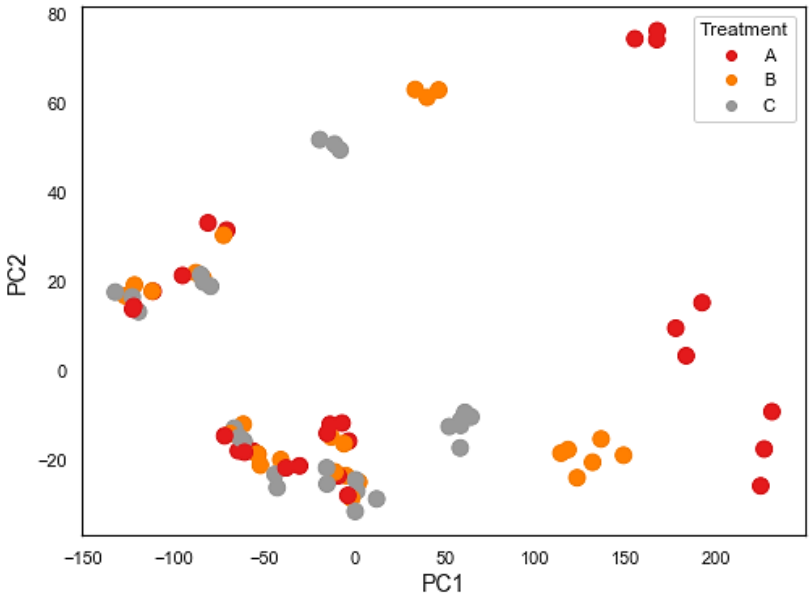


Figure 4.3. PCA plot showing fermenting samples based on sample preparation treatment.

4.3.2 FERMENTATION EXCITATION-EMISSION MATRICES

The three-dimensional EEMs of treatments A, B and C on the first and last day of fermentation are shown in **Figure 4.4**. The fluorescent intensity of the clean sample on day 1 is greater than those of the degassed and unaltered samples which may be attributed to the reduction in fluorescence as a result of turbidity, however, the effect on scattered light is increased for the turbid samples as can be seen in the elevated spectra alongside the removed identity bands ($\lambda_{ex} = \lambda_{em}$) and ($2\lambda_{ex} = \lambda_{em}$) of first and second order Rayleigh scattering, respectively. As fermentation is completed, all treatments experience a reduced fluorescence intensity with treatment A decreasing by roughly 300 units and treatments B and C by roughly 200 units.

This may be a result of the greater fluorescent abilities of monomeric pigments compared to their polymerised counterparts as suggested previously (Chapter 3) as well as colour changes occurring by means of anthocyanin extraction. Darker samples are known to reduce fluorescence intensity due to their increased reabsorption of light (Karoui and Blecker, 2011). Treatments B and C show no major differences between each other and the effect of carbon dioxide (CO_2) during fermentation may not substantially influence fluorescence spectra. The EEMs of treatments B and C in **Figure 4.4** indicate a shouldered peak compared to treatment A, roughly determined as the region 275-295 nm excitation and 320-360 nm emission (**Figure 4.5**, region 1). Treatment A indicates a slightly more prominent fluorescence determined between 255-265 nm excitation and 360-400 nm emission (**Figure 4.5**, region 2). As fermentation proceeds, these regions become more pronounced specifically with region 2 fluorescing more intensely between 320 and 340 nm emission.

The fluorescence in region 2 correlates well with the regions identified as flavan-3-ols, namely catechin, epicatechin and epigallocatechin, as well as polymeric proanthocyanidins (Airado-Rodríguez *et al.*, 2011) and may represent the extraction of condensed tannins during fermentation and their subsequent polymerisation. Although region 1 does not correlate with any phenolic spectral regions previously identified in literature, it does fall within the optimal region previously selected by the machine learning pipeline for the total phenolics model and will be elaborated on below. It can also be seen that the second main fluorescent region identified in wine (excitation greater than 300 nm and resulting emission between 360 and 450 nm) becomes more pronounced by the end of fermentation. Additionally, treatments B and C have slightly greater fluorescent intensities at the end of fermentation which may be a result of light scattering.

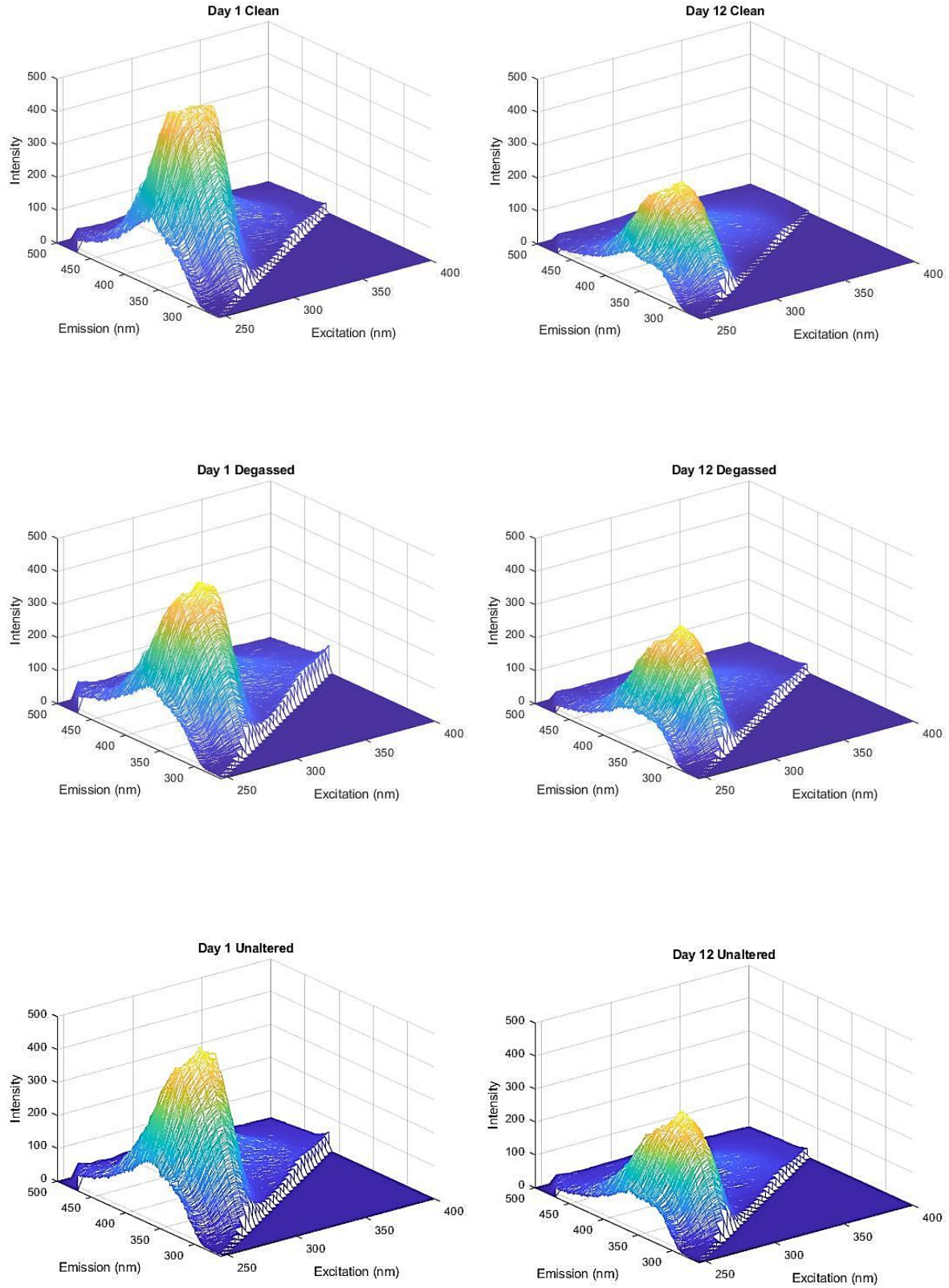


Figure 4.4. Three-dimensional excitation-emission matrices (EEMs) of each sample preparation treatment (clean, degassed and unaltered) on the first and last day of fermentation.

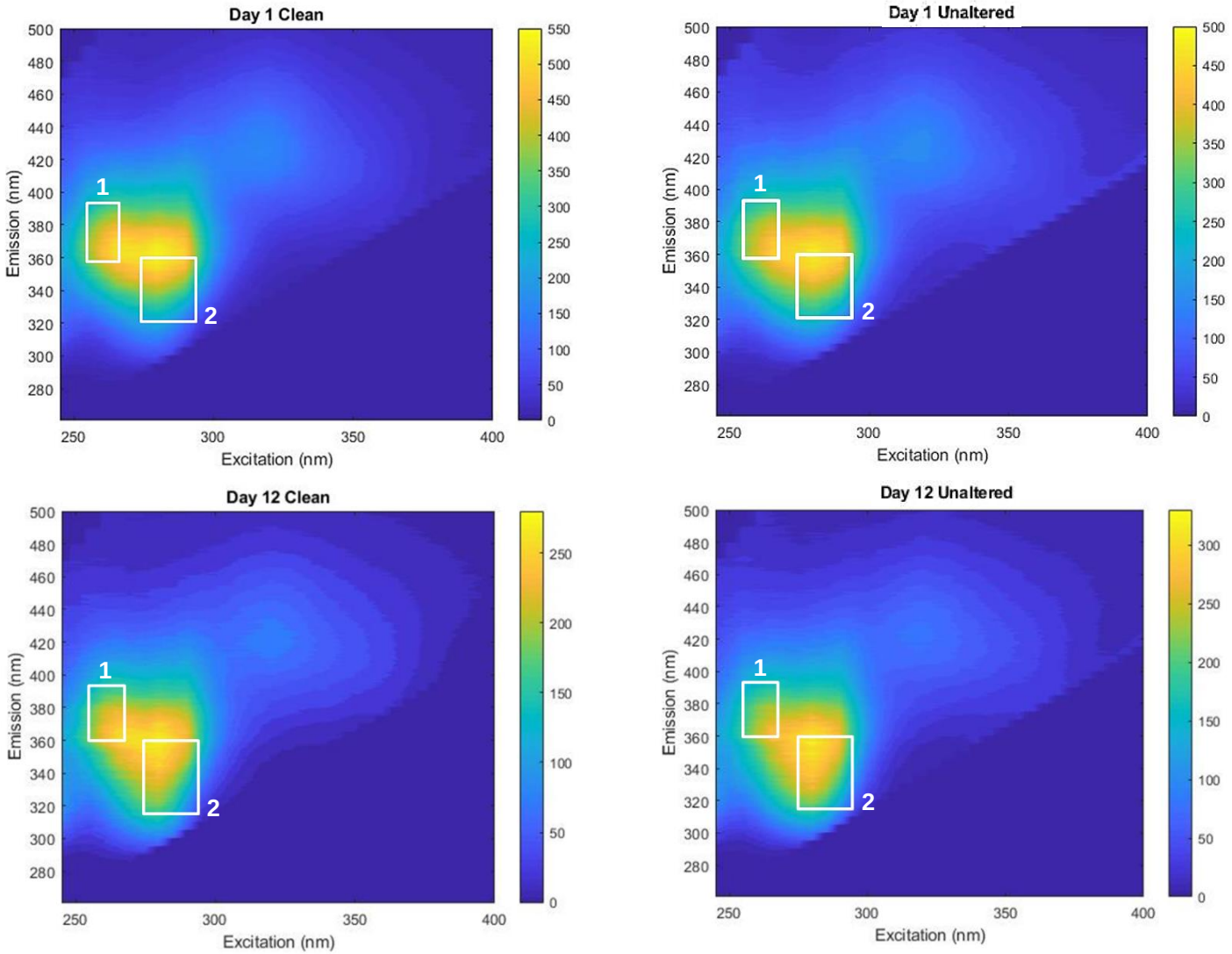


Figure 4.5. Excitation-emission matrices of a clean and unaltered sample on the first and last day of fermentation showing spectral regions of interest.

4.3.3 MODEL VALIDATION

The best models per phenolic parameter obtained through the machine learning pipeline in Chapter 3 were validated using the 81 samples collected throughout fermentation. This model validation involved obtaining the prediction accuracy, by means of root mean square error (RMSE) and mean absolute error (MAE), for various sub datasets (**Appendix Table 4.1**). Each phenolic model has its own unique set of parameters, with the column selector having identified optimal spectral regions for total phenolics (excitation 260-360 nm and emission 370-400 nm), total condensed tannins (excitation 285-340 nm and emission 290-350 nm) and total anthocyanins (excitation 280-300 nm and emission 330-380 nm). The colour density and polymeric pigments models cover the entire EEM obtained during fluorescence analysis. During model development, ten-fold cross validation was incorporated to prevent over-fitting and better understand model stability and performance while internally validating the model. The use of this external validation set of fermenting samples aids in investigating the suitability of the chosen parameters per phenolic model, explores model performance when predicting on unseen data and investigates the influence of sample preparation on prediction accuracy. Due to the differences in fluorescence according to the day of fermentation, three sub datasets were explored including the entire fermentation from day 1 to 12, early fermentation from day 1 to 3 and later fermentation from day 5 to 12. Although PCA did not clearly distinguish between sample preparation treatments, these were included as sub datasets to determine model performance under fermentation conditions, including potential effects from CO_2 and turbidity. The spectrophotometric reference data per phenolic parameter is reported in **Table 4.1** below.

Table 4.1. Maximum, minimum, standard deviation and average values per spectrophotometric analysis reference method.

	Total Phenols Index	**Total Condensed Tannins (mg/L)**	**Total Anthocyanins (mg/L)**	**Colour Density (AU)**	**Polymeric Pigments (AU)**
Minimum	31.63	1002.18	378.03	14.54	1.09
Maximum	52.35	1524.55	549.34	25.65	1.92
Average	42.78	1255.46	479.96	19.82	1.47
Standard deviation	6.40	164.42	56.13	3.26	0.30

Figure 4.6 shows the performance of the total phenols model possessing a calibration RMSE and MAE of 5.71 index values. The model performed better during early fermentation and generally had the greatest prediction accuracy with degassed and unaltered samples (treatments B and C) with the lowest overall RMSE and MAE for treatment C. On average, the model was able to predict the validation samples within 9.55 and 8.68 index units for RMSE and MAE, respectively (**Appendix Table 4.1**). The optimal total phenolic region identified during model development slightly overlaps region 1 as identified in **Figure 4.5** and could potentially be influencing the model's prediction abilities. Unaltered samples seemed to have less intense fluorescence in this region and perhaps building the model on clean samples allowed for over-fitting on regional spectral properties and was therefore able to better predict on samples with greater turbidity.

The total condensed tannins model performance is seen in **Figure 4.7** with no clear effect based on the day of fermentation but rather predicting better on clean samples (treatment A). On average, the model was able to predict the validation samples within 196.41 and 172.43 mg/L when compared with the calibration model's RMSE and MAE of 104.03 mg/L (**Appendix Table 4.1**). When looking at region 2 identified in **Figure 4.5**, the inverse effect of region 1 and total phenols may be occurring, with potential spectral interferences caused by the turbidity of samples reducing the prediction accuracy. The total anthocyanins model was able to predict the fermenting samples on average within 123.13 and 114.21 mg/L when looking at RMSE and MAE, respectively (**Appendix Table 4.1**). The model seemed to perform best during early fermentation (days 1 to 3) and on clean samples as seen in **Figure 4.8**. The optimal spectral region identified during model development slightly overlaps region 2 identified in **Figure 4.5** and may be influenced by the shouldered peak of the turbid samples as described for total condensed tannins.

The colour density model performed most poorly of all the models, showing no clear preference for day of fermentation or sample preparation and on average predicting within 7.419 and 6.810 AU compared to the calibration model's RMSE of 2.46 AU (**Figure 4.9**, **Appendix Table 4.1**). This may be a result of an optimistically cross-validated model as well as the metric of colour density itself. Colour density is an estimation of responsible yellow, red and blue colouring pigments at three UV-Visible spectral regions (Glories, 1984) and therefore the translation of these into the fluorescence EEM may not have been adequately achieved during model development. The polymeric pigments model performed the best, on average predicting within 0.371 and 0.307 AU when compared to the calibration model's RMSE of 0.63 AU (**Appendix Table 4.1**). The best model performance can be seen in **Figure 4.10** during later fermentation (days 5 to 12) and for degassed and unaltered samples (treatments B and

C). This improved prediction accuracy of the external validation set may be a result of cultivar specific benefits or the polymeric pigments range developing throughout fermentation falling within a region of the calibration model better able to predict. Although possessing seemingly poorer accuracy metrics reported previously (Chapter 3), the model as seen in **Figure 4.11** shows a relatively accurate prediction ability when analysing samples below 2 AU and incorporating more samples within the minority group above this threshold may improve upon the model's predictive ability. This illustrates the importance of balanced datasets in modelling.

Prediction models are known to perform better on data used to construct them than new and unseen data, resulting in some expected model depreciation during validation. However, internal validation techniques such as cross-validation or bootstrapping are often optimistically accepted without validating on external data (Bleeker *et al.*, 2003). Within the five best phenolic models previously developed and herein validated, an important consideration includes the variability and balance of the dataset used for model calibration. Certain regions within the models may predict better than others as can be seen with polymeric pigments and although a synthetic dataset was created during model development to offset any data imbalances using a synthetic minority over-sampling technique for regression (SMOTER), gaps may still remain and have implications for prediction accuracy. The results of this external validation also follow a single fermentation of a single cultivar and should therefore be further investigated to determine the prediction accuracy on other cultivars and fermentations.

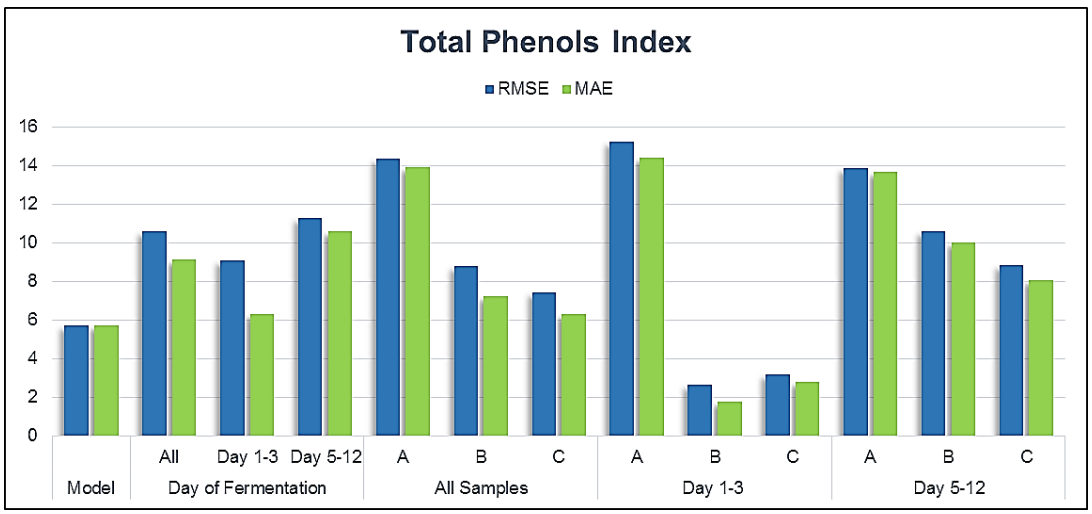


Figure 4.6. Prediction accuracy metrics (RMSE and MAE) for the externally validated total phenolics model.

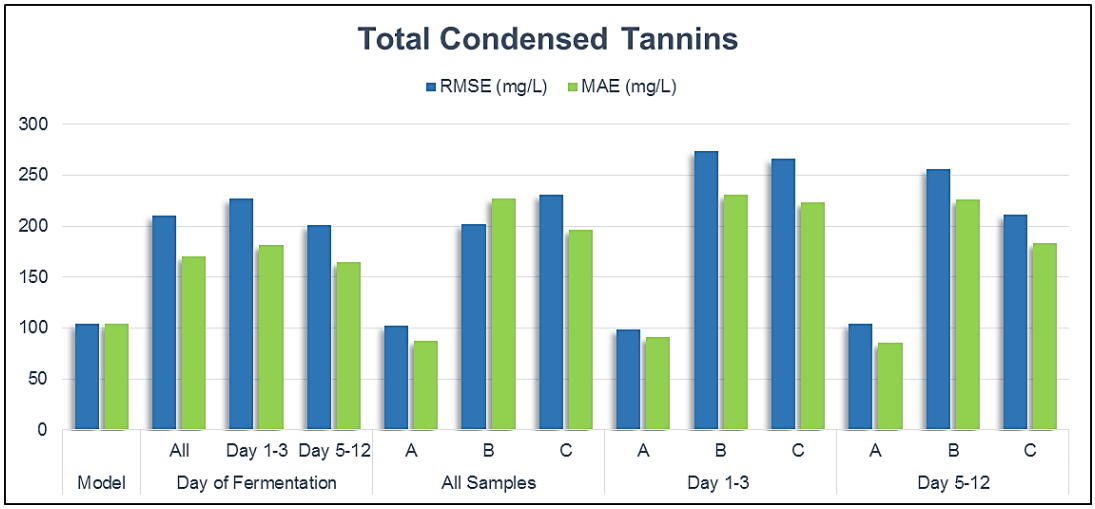


Figure 4.7. Prediction accuracy metrics (RMSE and MAE) for the externally validated total condensed tannins model.

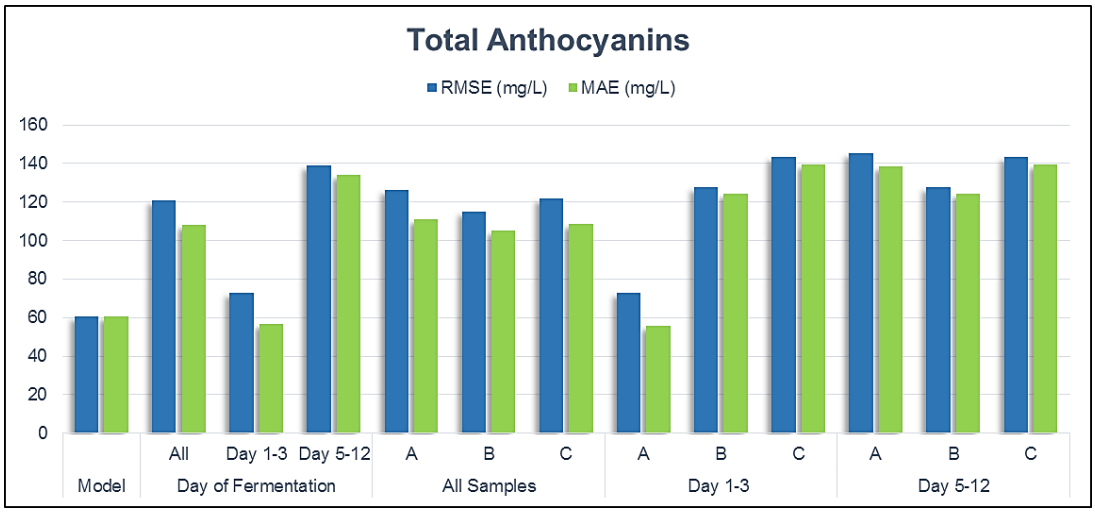


Figure 4.8. Prediction accuracy metrics (RMSE and MAE) for the externally validated total anthocyanins model.

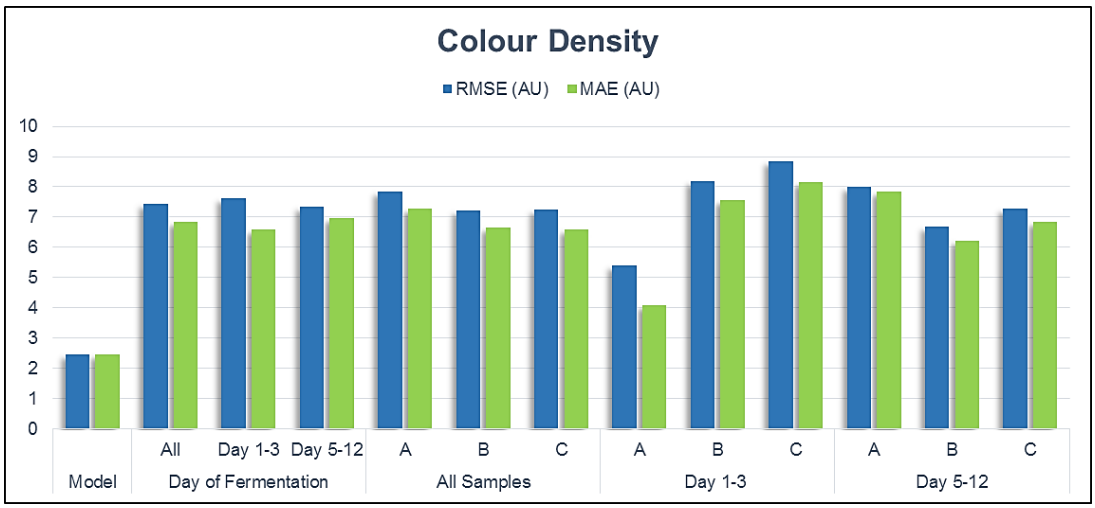


Figure 4.9. Prediction accuracy metrics (RMSE and MAE) for the externally validated colour density model.

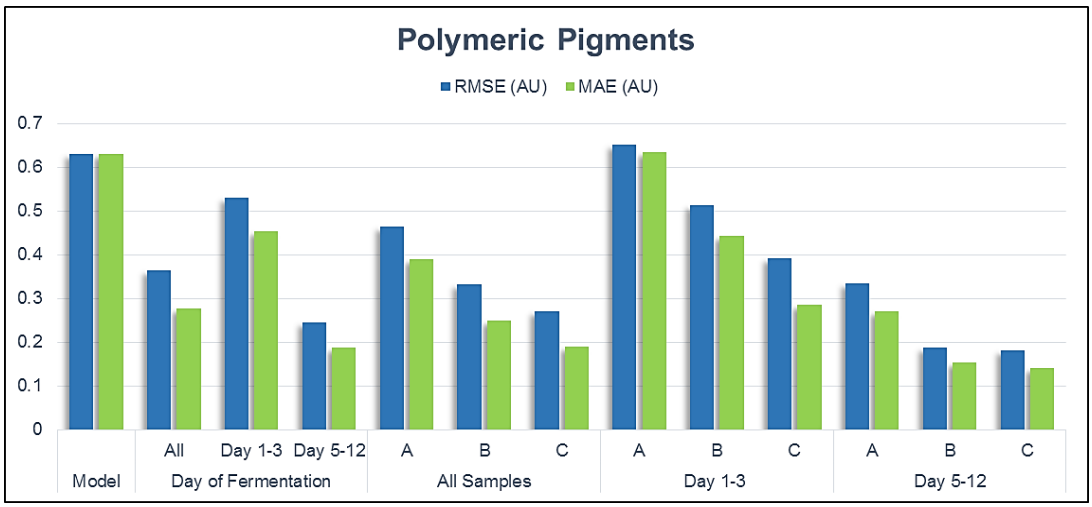


Figure 4.10. Prediction accuracy metrics (RMSE and MAE) for the externally validated polymeric pigments model.

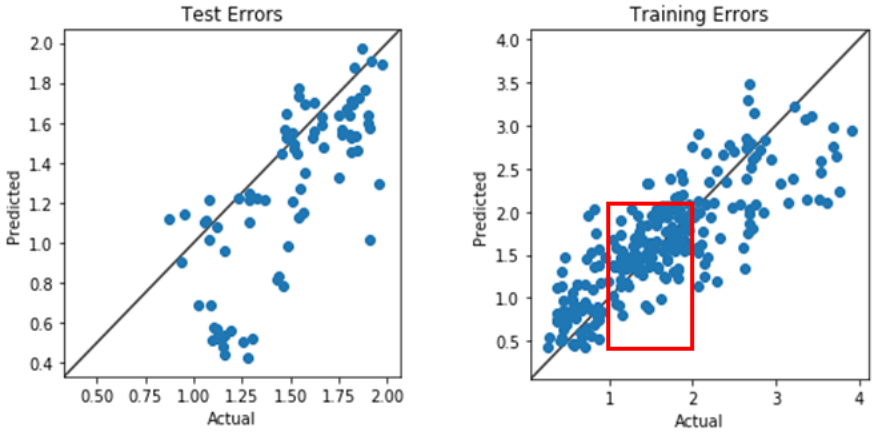


Figure 4.11. The polymeric pigments external validation set (left) against the polymeric pigments model calibration set (right) highlighting the fermentation range being predicted.

4.3.4 INFLUENCE OF SAMPLE PREPARATION ON QUANTIFYING PHENOLIC CONTENT

The fermenting samples were passed through the machine learning pipeline in order to validate the stability and suitability of the model parameters chosen during model development while most importantly determining the influence of sample preparation and the implications for real-time analysis during fermentation. All models passed through the same pipeline steps as in Chapter 3, excluding the SMOTER algorithm due to analysis taking place in triplicate and thereby creating an already well-balanced dataset eliminating the need for synthetic samples. The total anthocyanins and polymeric pigments models required the removal of outliers, specifically samples A1 and B3 and samples B4-B6, respectively. As no other phenolic model possessed outliers from the second day of analysis (B4-B6), it is difficult to identify the cause of such significant difference.

Table 4.2 below shows the predication accuracy metrics obtained per phenolic parameter model throughout fermentation for each sample preparation treatment. The triplicate analysis format conducted on consecutive fermentation days is most likely the cause of such successful results compared to the previously developed calibration models (Chapter 3), however an important consideration is that the effect of sample preparation may not noticeably influence front-face fluorescence spectroscopy, confirming the findings of **Figure 4.3** above. No clear differences can be identified between treatments, except perhaps for colour density where treatment A may have produced slightly better results. However the use of three prediction accuracy metrics, namely R^2, RMSE and MAE, allows for a more holistic evaluation of model performance as in the case of treatments B and C of colour density. Although the data is poorly fitted as reported by R^2, the RMSE and MAE values are not noticeably different to treatment A.

Overall, the obtained models resulted in high correlations and validate the chosen pipeline parameters as well as highlight the potential for building models using unaltered samples, the benefit of which involves the application in analysing samples directly from fermentation vessels. Analysing in triplicate aided in obtaining a well-balanced dataset and should be a consideration in further modelling. Due to only a small number of samples being passed through the pipeline per treatment, the results should primarily be considered as a proof of concept highlighting the potential for building fermentation-based models. The colour density model, although producing promising results in **Table 4.2**, should be approached with caution as model development and external validation performed the most poorly of all the phenolic parameters models and should therefore be further explored with regards to optimal model parameter selection and development. As previously discussed, the characteristics of colour density as a metric may potentially limit the success of modelling in this study due to the fluorescent EEM not encompassing the responsible regions or fluorescence spectral characteristics having not been adequately identified.

When comparing the above models with literature, it was found that the total condensed tannins model performed the best and presents itself as a promising alternative to other spectrophotometric analysis methods such as UV-Vis and infrared spectroscopies. UV-Vis models developed by Aleixandre-Tudo *et al*. (2018a) obtained RMSE scores of 239 and 209 mg/L for calibration and prediction, respectively, and can be compared to the fluorescence model developed previously with a RMSE of 104.03 mg/L and externally validated above to predict on average within 196.409 mg/L. Infrared calibration models built using Fourier transform near infrared (FT-NIR), attenuated total reflectance mid infrared (ATR-MIR) and Fourier transform infrared (FT-IR) spectroscopies show the same trend (Aleixandre-Tudo *et*

al., 2018b). The total condensed tannins model built using unaltered samples possessing R^2cal 0.86, R^2val 0.94, and RMSEC 48.42 mg/L (Table 4.2) is also able to compete while showcasing the potential for building models using fermenting samples analysed directly from the tank, eliminating the need for sample preparation. Although producing slightly less competitive results, the total phenols and total anthocyanins models built in this study show promise and may too present themselves as successful alternatives. When looking at fluorescence spectroscopy in literature, models have previously been built on pure compounds such as catechin and epicatechin rather than broader phenolic metrics such as total condensed tannins (Airado-Rodŕiguez *et al.*, 2009; Cabrera-Bañegil *et al.*, 2017). Raman spectroscopy calibrations for Cabernet Sauvignon wine phenolics have been successfully investigated and although based on competing phenomena, the fermentation models described in **Table 4.2** can be considered comparatively successful in their prediction accuracies (Gallego *et al.*, 2011).

Table 4.2. Prediction accuracy metrics (R^2, RMSE and MAE) of the external validation models obtained per phenolic parameter for sample preparation treatments A (clean), B (degassed) and C (unaltered).

	Treatment	R^2cal	R^2val	RMSEC	RMSEV	MAEV
Total Phenols Index	A	0.90	0.97	1.53	1.22	0.96
	B	0.89	0.87	1.72	2.27	1.73
	C	0.94	0.96	1.26	1.38	1.12
Total Condensed Tannins (mg/L)	A	0.89	0.94	51.70	41.30	34.89
	B	0.86	0.96	57.36	34.44	27.55
	C	0.86	0.94	48.42	42.84	35.95
Total Anthocyanins (mg/L)	A	0.85	0.87	16.35	19.19	14.66
	B	0.89	0.91	13.90	18.40	15.09
	C	0.93	0.89	14.34	20.53	15.17
Colour Density (AU)	A	0.79	0.52	1.28	2.32	1.84
	B	-0.24	0.36	2.41	2.21	2.04
	C	-0.14	-0.24	1.70	1.87	1.64
Polymeric Pigments (AU)	A	0.81	0.61	0.09	0.18	0.12
	B	0.95	0.13	0.04	0.26	0.13
	C	0.95	0.82	0.05	0.14	0.10

R^2cal: coefficient of correlation in calibration; R^2val: coefficient of correlation in validation; RMSEC: root mean square error of calibration; RMSEV: root mean square error of validation; MAEV: mean absolute error of validation.

4.4 CONCLUSION

Monitoring phenolic content during winemaking may aid in the decision making and implementation of vinification practices thereby improving process control and fermentation management. This study validated the potential for phenolic models built using fluorescence spectroscopy and chemometrics as well as the suitability of front-face geometry to quantify phenolics of fermenting musts under fermentation conditions. This may aid in the rapid, cost-effective and accurate monitoring of phenolic extraction throughout fermentation and the implementation of appropriate winemaking practices. Following a Cabernet Sauvignon fermentation allowed for improved understanding of the evolution of fluorescence spectra from juice to wine. The performance of each phenolic parameter model under different conditions, including stage of fermentation as well as sample preparation, was determined and should be considered unique and model specific. The models were adequately validated and show the potential for analysing directly from the fermentation vessel which may allow for phenolic analysis using portable optical devices or on-line automated systems. The potential for building fermentation-based models appears promising and may be beneficial to winemakers in creating cellar specific software able to be expanded on each vintage and used as a tool for optimal red wine production.

4.5 REFERENCES

Airado-Rodŕiguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2009. Usefulness of fluorescence excitation-emission matrices in combination with parafac, as fingerprints of red wines. Journal of Agricultural and Food Chemistry, 57(5), 1711–1720.

Airado-Rodríguez, D., Durán-Merás, I., Galeano-Díaz, T. and Wold, J., 2011. Front-face fluorescence spectroscopy: A new tool for control in the wine industry. Journal of Food Composition and Analysis, 24(2), 257–264.

Aleixandre-Tudo, J. L., Nieuwoudt, H., Olivieri, A., Aleixandre, J. L. and du Toit, W., 2018a. Phenolic profiling of grapes, fermenting samples and wines using UV-Visible spectroscopy with chemometrics. Food Control, 85, 11–22.

Aleixandre-Tudo, J. L., Nieuwoudt, H., du Toit, W. and Aleixandre, J. L., 2018b. Chemometric compositional analysis of phenolic compounds in fermenting samples and wines using different infrared spectroscopy techniques. Talanta, 176, 526–536.

Andersen, C. M. and Bro, R., 2003. Practical aspects of PARAFAC modeling of fluorescence excitation-emission data. Journal of Chemometrics, 17(4), 200–215.

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. E. and Moons, K. G. M., 2003. External validation is necessary in prediction research : A clinical example. Journal of Clinical Epidemiology, 56, 826–832.

Cabrera-Bañegil, M., Hurtado-Sánchez, M., Galeano-Díaz, T.and Durán-Merás, I., 2017. Front-face fluorescence spectroscopy combined with second-order multivariate algorithms for the quantification of polyphenols in red wine samples. Food Chemistry, 220, 168–176.

Cadot, Y., Miñana-Castelló, M.T. and Chevalier, M., 2006. Anatomical, histological, and histochemical changes in grape seeds from Vitis vinifera L .cv Cabernet franc during fruit development. Journal of Agricultural Food Chemistry, 54(24), 9206–9215.

Canals, R., Llaudy, M. C., Vall, J., Canals, J. M. and Zamora, F., 2005. Influence of ethanol

concentration on the extraction of color and phenolic compounds from the skin and seeds of Tempranillo grapes at different stages of ripening. Journal of Agricultural and Food Chemistry, 53(10), 4019-4025.

Casassa, L. F. and Harbertson, J. F., 2014. Extraction, evolution, and sensory impact of phenolic compounds during red wine maceration. Annual review of food science and technology, 5, pp.83-109.

Chai, T. and Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific model development, 7(3), 1247-1250.

Chen, T. and Guestrin, C., 2016. XGBoost : A scalable tree boosting system: in proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794.

Gallego, Á. L., Guesalaga, A. R., Bordeu, E., González, A. S., 2011. Rapid measurement of phenolics compounds in red wine using Raman spectroscopy. IEEE Transactions on Instrumentation and Measurement, 60(2), 507–512.

Glories, Y., 1984. La couleur des vins rouges, 2eme partie. Connaissance de la Vigne et du Vin, 18, 253−271.

Iland, P., Ewart, A., Sitters, J., Markides, A. and Bruer, N., 2000. Techniques for chemical analysis and quality monitoring during winemaking (1st ed, 1-111). Campbelltown, Patrick Iland Wine Promotions.

Karoui, R. and Blecker, C., 2011. Fluorescence spectroscopy measurement for quality assessment of food systems — a review. Food and Bioprocess Technology, 4(3), 364–386.

Mercurio, M. D., Dambergs, R. G., Herderich, M. J. and Smith, P. A., 2007. High throughput analysis of red wine and grape phenolics − adaptation and validation of methyl cellulose precipitable tannin assay and modified somers color assay to a rapid 96 well plate format. Journal of Agricultural and Food Chemistry, 55(12), 4651−4657.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnav, E., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Pelikan, M., Goldberg, D.E. and Cantú-Paz, E., 1999. BOA: The Bayesian optimization algorithm. In: Proceedings of the genetic and evolutionary computation conference GECCO-99, 1, 525-532.

Sacchi, K. L., Bisson, L. F. and Adams, D. O., 2005. A review of the effect of winemaking techniques. American Journal of Enology and Viticulture, 56(3), 197–206.

Savitzky, A. and Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36, 1627-1639

Smith, P. A., Mcrae, J. M. and Bindon, K. A., 2015. Impact of winemaking practices on the concentration and composition of tannins in red wine. Australian Journal of Grape and Wine Research, 21, 601-614.

Strasburg, G. M. and Ludescher, R. D., 1995. Theory and applications of fluorescence spectroscopy in food research. Trends in Food Science and Technology, 6(3), 69-75.

Swersky, K., Snoek, J. and Adams, R.P., 2013. Multi-task bayesian optimization. Advances in neural information processing systems, 26, 2004-2012.

Chapter 5

General conclusions and recommendations

5.1 GENERAL CONCLUSIONS

Phenolic compounds are important secondary metabolites playing crucial roles in red wine characteristics such as colour, mouthfeel and ageing potential, thus largely influencing consumer-perceived quality. The ability to analyse these compounds and their extraction throughout winemaking and ageing may aid in the implementation of appropriate and judicious practices in order to improve upon the final product. To date, the array of spectrophotometric analysis methods available are often as equally complex as their phenolic compounds of interest, resulting in time-consuming, elaborate protocols unable to be easily conducted by inexperienced personnel. As a result, phenolic analysis is not a widespread and routine practice during red wine production. The main aim of this study was therefore to investigate the suitability of fluorescence spectroscopy for the direct quantification of five phenolic parameters, namely total phenolics, total anthocyanins, total condensed tannins, colour density and polymeric pigments, and to subsequently determine the potential for accurate non-invasive analysis during fermentation.

The optimisation of front-face fluorescence spectroscopy in order to analyse undiluted samples was successful. Fluorescence and UV-Vis spectroscopies were conducted on 289 samples, incorporating a diverse range of cultivars for both fermenting and finished wines, and the most optimal chemometric method for developing accurate regression models per phenolic parameter was investigated with a focus on PARAFAC and machine learning algorithms. PARAFAC, although successfully decomposing complex fluorescence data into the responsible components and correlating with results found in literature, was not suitable in developing accurate predictive models for the broader phenolic parameters included in this study. A machine learning pipeline was subsequently built and successfully developed models for total phenolics, total anthocyanins and total condensed tannins which may present themselves as promising alternatives to other spectrophotometric methods such as UV-Vis and infrared spectroscopies. The polymeric pigments model was found to accurately predict below 2 AU and the incorporation of more samples in the minority group above this level may aid in developing a more robust model. The colour density model requires further investigation and does not lend itself as a promising alternative compared to the current analysis methods used.

The validation of these models under different fermentation conditions was performed in order to investigate the potential for analysing unaltered samples directly from the fermentation vessel. The results showed that prediction accuracy of the models was not negatively affected

according to sample matrix conditions throughout fermentation, including the effects of carbon dioxide or turbidity, and therefore demonstrates the suitability of front-face fluorescence in analysing samples in their truest forms. The implications for measuring phenolic content non-invasively throughout fermentation may be beneficial to winemakers in the form of fluorescence based portable devices or in-line systems. Additionally, the validation of the phenolic parameter models under fermentation conditions allowed for external validation and the subsequent inference that, with the exception of colour density, model development and optimisation was successful.

The classification abilities of fluorescent excitation-emission matrices have been widely accepted in food science and agricultural disciplines, and this study investigated NCA as a novel approach for the classification of South African cultivars based on unique, cultivar-specific fluorescent characteristics. Overall, the results from this study illustrate fluorescent differences between fermenting musts and wine, and subsequently the unique fluorescent changes occurring throughout fermentation. These changes may be a result of fluorescent differences between monomeric and polymeric pigments, the metabolism of fluorescent compounds by yeast during fermentation as well as spectral interferences as a result of increased anthocyanin extraction and developing red wine colour. Following a single Cabernet Sauvignon fermentation allowed for a novel understanding of the fluorescent changes and their corresponding spectral regions occurring from grape must to wine. This identified potential influences on the performance of the phenolic parameter models, suggesting that prediction accuracy may be slightly affected by the stage of fermentation and should be a consideration throughout fermentation. Additionally, the potential for developing fermentation-based models was suggested and may present itself as an opportunity for creating cellar specific phenolic parameter models and the ability for expansion with each subsequent vintage.

5.2 FUTURE RECOMMENDATIONS

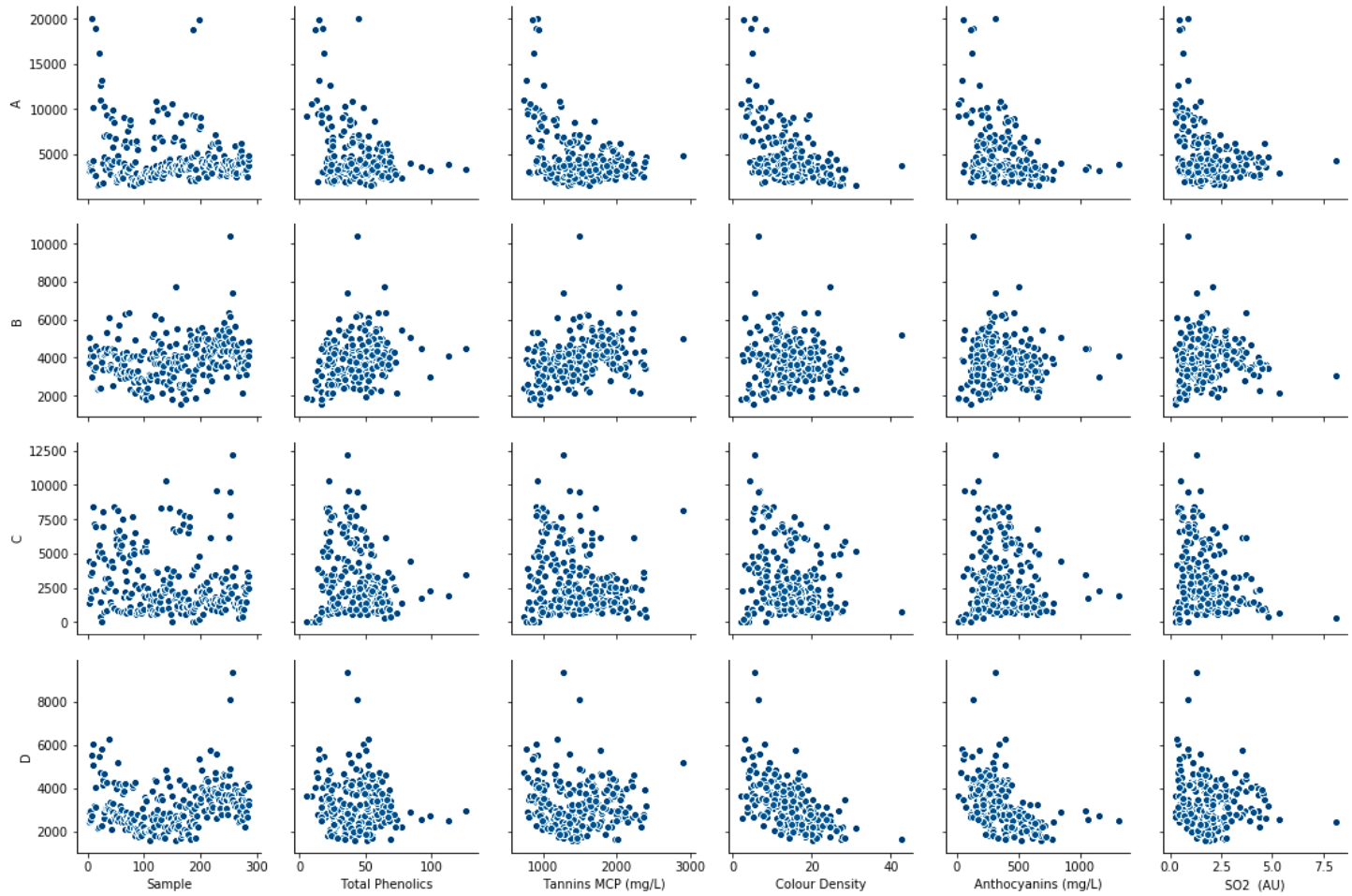
The high natural variability of wine as a result of cultivar type, viticultural factors and different winemaking practices implemented throughout production, presents itself as a major consideration in the development of accurate regression models. The incorporation of a more balanced dataset of fermenting musts and wines as well as increasing the number of samples for minority sample groups, is recommended in order to prevent gaps in model development and therefore prediction accuracy. Additionally, the analysis of samples in duplicate or

triplicate may allow for the current gaps to be filled, which was not possible as a result of fluorescence analysis time per sample and the large number of samples aiming to be included.

Model validation using different cultivars and fermentations is recommended in order to evaluate model performance against different phenolic compositions. The fluorescent changes reported on throughout the externally validated fermentation are unique to Cabernet Sauvignon and should be further investigated in order to understand the unique fluorescence of other cultivars.

The further development of potential portable devices or in-line automated systems should repeat model validation, as the findings reported in this study are unique to the front-face geometry and the fluorescence spectrophotometer optimised in this study and model modification may be necessary when altering sensitive instrumentation.

Appendix Chapter 3



Appendix Figure 3.1. Exploratory analysis of the PARAFAC scores against phenolic UV-Vis reference data.

Appendix Chapter 4

Appendix Table 4.1. Average predication accuracy scores per phenolic parameter model for model validation using a Cabernet Sauvignon fermentation.

		Model	Average	All	Day 1-3	Day 1-5	A	B	C
TP	RMSE	5.71	9.545	10.292	7.536	11.146	14.474	7.332	6.490
(AU)	MAE	5.71	8.678	9.152	6.307	10.574	13.982	6.328	5.723
MCP	RMSE	104.03	196.409	186.563	216.315	193.455	101.862	243.914	236.346
(mg/L)	MAE	104.03	172.426	170.533	181.893	164.852	88.332	228.055	200.891
Anth	RMSE	60.67	123.133	121.032	104.246	138.921	114.877	123.449	136.271
(mg/L)	MAE	60.67	114.207	108.339	94.013	134.077	101.827	117.910	129.075
CD	RMSE	2.46	7.419	7.435	7.515	7.329	7.077	7.372	7.786
(AU)	MAE	2.46	6.810	6.851	6.605	6.974	6.414	6.812	7.203
PP	RMSE	0.63	0.371	0.359	0.522	0.237	0.484	0.344	0.282
(AU)	MAE	0.63	0.307	0.277	0.454	0.188	0.432	0.282	0.206