Research Article

Received: 21 May 2018 Revised: 9 September 2018 Accepted article published: 16 September 2018 Published online in Wiley Online Library: 16 November 2018

(wileyonlinelibrary.com) DOI 10.1002/jctb.5827

NIR spectroscopy applied to the determination of 2-phenylethanol and L-phenylalanine concentrations in culture medium of *Yarrowia lipolytica*

Cristina Quintelas,[a*†] Adelaide Braga,[a†] Daniela P Mesquita,[a†] Antonio L Amaral,[a,b] Eugenio C Ferreira[a] and Isabel Belo[a]

Abstract

BACKGROUND: This work aims at developing a method, based on near-infrared (NIR) spectroscopy, to quantify 2-phenylethanol (2-PE) and L-phenylalanine (L-Phe) concentrations during its microbial production by *Yarrowia lipolytica*. For this purpose, 197 samples obtained from different batch cultures were analyzed using Fourier transform (FT)-NIR transmission spectroscopy in the range of 200–14 000 cm^{-1}.

RESULTS: A principal components analysis was performed for cluster identification and outlier removal. A partial least squares regression was next applied to develop the calibration models, by an iterative method. The predictive ability of the models was confirmed by an external validation procedure with an independent sample set. The obtained results could be considered promising, with coefficients of determination (R^2) of 0.92 for L-Phe and 0.95 for 2-PE, and residual prediction deviation above 3, for the ensemble data.

CONCLUSION: The described methodology, using NIR spectroscopy and chemometrics, can be seen as a promising fast tool to determine both studied flavor compounds during biotechnological processes as an alternative to chromatographic methods.

Supporting information may be found in the online version of this article.

Keywords: NIR spectroscopy; PCA; PLS; flavor compounds; *Yarrowia lipolytica*

INTRODUCTION

Monitoring a bioprocess for the production of flavors and fragrances may be imperative and will permit enhancing the production of these compounds. Typical methods to analyze flavor and fragrances are gas chromatography (GC) and high-performance liquid chromatography (HPLC),[1] presenting obvious constraints for real-time monitoring of biotransformation processes. GC techniques have been applied to amino acid analysis for a long time.[2] Nevertheless, this approach always requires one or several derivatization steps to make the analytes sufficiently volatile.[3] HPLC-based methods often do not need any derivatization;[4] however, no complete chromatographic separation can be achieved in many cases, making it impossible to quantify the compounds. Furthermore, these techniques are often tedious and invasive, requiring sample handling and being difficult to perform in real time. A rapid and more direct method of analysis would yield more timely process information and improved bioprocess control.

In recent years, developments in both chemometrics and instrumentation have resulted in rapid methods relating multivariate spectroscopic and chemical data for predicting the concentration of specific chemical constituents, thus helping to reduce the demand for traditional analysis using chemical reagents. In fact, spectrometry-based analytical methods can be used nowadays to monitor biotransformation processes.[5] The main advantages of near-infrared (NIR) spectroscopy over reference methods are its speed, both non-destructive and non-contaminant nature, and great accuracy.[6] Bioprocesses are usually complex, both from the chemical (ill-defined medium composition) and physical (multiphase matrix) aspects, which poses an additional challenge to the development of robust calibrations.[7] As a result, different studies have been conducted to apply NIR to the monitoring of bioprocesses in general.[8,9] Furthermore, NIR has also been applied to monitoring of: (i) biomass, glucose, lactic and acetic acid

* *Correspondence to: C Quintelas, CEB - Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. E-mail: cquintelas@deb.uminho.pt*

† *These authors contribute equally to this work.*

a *CEB – Centre of Biological Engineering, University of Minho, Braga, Portugal*

b *Instituto Politécnico de Coimbra, ISEC, DEQB, Rua Pedro Nunes, Quinta da Nora, Coimbra, Portugal*

content during aerobic fermentations;[10] (ii) amino acid concentration profile during hydrolysis processes;[11] and (iii) phenylethanol in grapes,[12] red wine[13] and apple wine.[14]

2-Phenylethanol (2-PE) is an aromatic alcohol with a delicate fragrance of rose petals,[15] widely applied in diverse types of products, such as perfumes, cosmetics, pharmaceuticals, foods and beverages.[16] The economic importance of 2-PE is quite significant.[17] With the current available information, the global production of 2-PE is estimated at 10 000 tons per year, being dominated by chemical synthesis.[17,18] Several microorganisms naturally produce 2-PE as part of their amino acid catabolism. In yeasts, the 2-PE biosynthesis is connected to the phenylethylamine (*de novo* synthesis) and Ehrlich pathways (L-phenylalanine (L-Phe) biotransformation).[16,19,20] Among several microorganisms able to produce 2-PE, the yeast *Yarrowia lipolytica* appears to be promising because of its interesting characteristics, such as the Crabtree-negative trait and absence of ethanol production. Taking these properties into consideration, the present work addresses its use for the production of this compound.

In the present work we explored the potential of NIR spectroscopy to determine 2-PE and L-Phe concentrations during biotechnological processes. A chemometric approach was essential in that regard, consisting of a boxplot analysis, employed to identify possible outliers regarding the concentrations of L-Phe and 2-PE, and a principal component analysis (PCA) establishing the interrelationships, regarding the obtained wavelength spectra, among the different culture conditions and *Y. lipolytica* employed strains, as well as possible sample outliers. Finally, a partial least square analysis (PLS) was performed in order to obtain a prediction model suitable for L-Phe and 2-PE monitoring purposes.

EXPERIMENTAL

Microorganism, media and culture conditions

The strains used in this work were *Y. lipolytica* W29 (ATCC 20460), *Y. lipolytica* CBS 2075 and *Y. lipolytica* NCYC 1026. These strains were stored at −80 °C and routinely cultivated on YPDA medium (agar 30 g L^{-1}, glucose 20 g L^{-1}, peptone 20 g L^{-1}, yeast extract 10 g L^{-1}) at 27 °C. Cells were cultivated for 16–17 h in a 500 mL baffled Erlenmeyer flask containing 200 mL YPD medium (glucose 20 g L^{-1}, peptone 20 g L^{-1}, yeast extract 10 g L^{-1}) on a rotary shaker at 200 rpm and 27 °C and further used to inoculate shake flasks for 2-PE production experiments for an initial OD_{600} of 0.5.

The 2-PE production by *de novo* synthesis was carried out in 500 mL shake flasks with 200 mL of the medium containing 10 g L^{-1} yeast extract and 40 g L^{-1} glucose, at 27 °C and 200 rpm,[21] incubated at 200 rpm and 27 °C. Bioconversion of L-Phe to 2-PE was carried out in 500 mL shake flasks with 200 mL modified cultivation medium (based on Cui *et al.*[22]) containing, per liter of deionized water: glucose 60 g, KH_2PO_4 15 g, $MgSO_4.7H_2O$ 0.5 g, yeast nitrogen base without amino acids 0.2 g, thiamine 3 mg, pH 6.5, supplemented with L-Phe 5 g (biotransformation) incubated at 27 °C and 200 rpm. Another strategy was also tested with an initial growth period, without L-Phe supplementation, for 48 h, followed by the addition of 5 g L-Phe to the culture medium at that time (biotransformation + L-Phe). Samples of 2 mL were taken over time, centrifuged (10 000 rpm for 10 min at 4 °C), filtered (Whatman, PES – 0.2 µm) and stored at −20 °C. Prior to analysis, for both HPLC and NIR, the liquid samples were thawed and homogenized by vortexing.

HPLC analysis

The 2-PE and L-Phe concentrations were determined using a Shimadzu ultra-HPLC system equipped with a diode array detector (SPD-M20A). LC separation was carried out with a YMC ODS-Aq (250 mm × 4.6 mm) reverse-phase column at 25 °C. For elution, water (solvent A) and acetonitrile (solvent B) were employed as the mobile phases at a flow rate of 1 mL min^{-1}. A gradient was used whereby the amount of solvent A was increased stepwise: 0 min – 100% A, 10 min – 100% A, 16.7 min – 70% A, 26.7 min – 70% A, 33.3 min – 100% A; 41.7 min – 100% A. A diode array detector was used at a fixed wavelength of 215 nm.[23] The values of R^2 for the calibration curve were 0.99 for both compounds, root mean square error (RMSE) of 4.72% and standard error (SE) of 10 690.2 mV min^{-1} for 2-PE and 8.78% and 89 521.3 mV min^{-1} for L-Phe, and the limits of detection were 3.6 and 22.9 mg L^{-1}, respectively, for 2-PE and L-Phe.

NIR scanning

NIR spectra were recorded on a Fourier-transform NIR spectrometer (FTLA 2000, ABB, Thermo Electron Corporation) equipped with an indium–gallium–arsenide (InGaAs) detector, from 14 000 to 200 cm^{-1}, in transmittance mode, using a flow cell with a 0.7 mm path length. For each sample, 64 scans were obtained with a spectral resolution of 8 cm^{-1} and then averaged. Samples were temperature equilibrated at 23 °C (approximately 3 min) within the spectrometer before scanning. The integration time was adjusted until the peaks at 1100–1200 nm for NIR were close to 60 000 intensity units. Grams/AI software (Thermo Electron Corporation) was used for spectrometer configuration, control, and data acquisition. Distilled water was used as background. A typical NIR spectrum is presented as supplementary material (supporting information, Fig. S1).

Chemometrics and data analysis

The 2-PE and L-Phe concentrations, from samples collected throughout the different experiments in this work, were employed as the Y dataset in the chemometric analyses, whereas the collected NIR spectra, ranging from 14 000 to 200 cm^{-1}, were employed as the X dataset. A number of different chemometric techniques were employed to process these data, namely: (i) boxplot analysis to identify Y dataset outliers; (ii) PCA to identify sample interrelationships (clusters) and X dataset outliers; and (iii) PLS regression to derive the models for each studied compound.

A boxplot analysis returns a box graph, for normally distributed data, with the central mark being the median, the edges being the 25th and 75th percentiles, and the whiskers extending to the most extreme data points not considering the outliers. The maximum whisker length allowed can be defined as a factor of the interquartile distance between the 25th and 75th percentiles, covering a chosen percentage of a normally distributed data. As a result, outliers are plotted individually, outside the box, and can be identified by visual inspection. For the employed boxplot analysis, the maximum whisker length allowed was 1.5, resulting in the identification of the outliers falling outside a 99.3% coverage of normally distributed data.

PCA reduces high-dimensional, and strongly correlated, X datasets by extracting the most relevant information onto latent variables (LVs), representing a linear combination of the original variables. Each new LV accounts for less explanatory power than the previous one, given that this technique aims at maximizing the explained variance for each new orthogonal space. As a result, the

Table 1. Calibration and validation dataset size for the 2-PE and L-Phe studies

	L-Phe estimation		2-PE estimation	
Sample[a]	Calibration	Validation	Calibration	Validation
Ensemble	127	64	131	66
W29	45	23	46	24
CBS 2075	46	24	46	24
NCYC 1026	35	18	38	20
de novo	44	22	43	22
Biotransformation	43	22	44	23
Biotransformation + L-Phe	40	20	43	22

[a] The strain-based experiments encompass the data for the different conditions (*de novo* synthesis, biotransformation and biotransformation + L-Phe) studied for each strain, and the condition-based experiments encompass the three different strains (W29, CBS and NCYC) studied for each condition.

interrelationships between the original X variables and samples, in the new LVs space, can be identified as clusters in the loadings and scores map, respectively.

For the PCA study, as well as for the PLS, three different datasets were employed: [D1] – ensemble dataset containing all the data samples; [D2] – strain-based partial datasets, divided into three groups representing the three different strains studied (W29, CBS and NCYC); and [D3] – conditions-based partial datasets, divided into three groups representing the three different conditions studied (*de novo* synthesis, biotransformation and biotransformation + L-Phe). Hence, in the performed analysis, the total number of samples varied, regarding L-Phe estimation, between 191 for the ensemble dataset and 68 (W29), 70 (CBS 2075), 53 (NCYC 1026), 66 (*de novo* synthesis), 65 (biotransformation), and 60 (biotransformation + L-Phe). With respect to the 2-PE prediction, the number of samples was 197 (ensemble), 70 (W29), 70 (CBS 2075), 58 (NCYC 1026), 65 (*de novo* synthesis), 67 (biotransformation), and 65 (biotransformation + L-Phe). Furthermore, two thirds of the samples were used for modeling (calibration) purposes and one third for validation, as presented in Table 1.

In order to select the most unbiased calibration and validation datasets, a screening of 500 possible random combinations, for these datasets, was employed. In this way, the best overall (calibration + validation) results were chosen, reflecting the most unbiased datasets combination. This procedure ensured that both calibration and validation datasets were distributed along the full range of the 2-PE and L-Phe concentrations.

The next employed PLS analysis shares some common ground with the PCA, also constructing LVs from the original X dataset in new (and orthogonal) spaces, although the main aim here is maximizing the captured predictive power of the X-space with regard to the Y-space. In PLS, a standard normal variate (SNV) methodology is usually employed to remove undesirable the X data matrix variations, alongside cross-validation (CV) techniques to test its predictive significance.

Two different methodologies were employed regarding the PLS approach, one employing the raw dataset [M1] and the other based on an iterative method [M2], first determining the weights of each wavelength for the entire wavelength values, next grouping the wavelength values together according to the weight similarity and, finally, recalculating the PLS with the averaged wavelength values.[24] A procedure was next implemented, for both approaches, to select the most unbiased calibration and validation datasets, by screening at a maximum of 500 possible random combinations for these datasets selection. Thus the best overall (calibration + validation) results were chosen, reflecting the most unbiased dataset combination. For all PLS analyses, the maximum number of PLS components allowed was set at half of the calibration data.

All of the above procedures and calculations were performed in MATLAB 7.11 (MathWorks, Inc. Natick, MA, USA). Further details for these techniques can be found in Einax *et al.*[25]

RESULTS AND DISCUSSION

Analytical data

The minimum, average and maximum values in the ensemble and strain-based samples are presented in Table 2. Each of these samples was then divided into two groups: the calibration (modeling) group with two thirds of the samples and the validation with the remaining one third of the samples.

Table 2. Number of samples and minimum (Min.), maximum (Max.) and average (Avg.) values of the 2-PE and L-Phe compounds

	2-PE (mg L^{-1})				L-Phe (mg L^{-1})			
Sample	No.	Min.	Max.	Avg.	No.	Min.	Max.	Avg.
Ensemble	197	0	1057.9	114.8	191	0	4789	1112.5
W29 strain	70	0	1057.9	137.0	68	0	4655.1	1083.3
CBS strain	70	0	818.4	103.5	70	0	4789	1229.2
NCYC strain	58	0	819.3	101.3	53	0.7	4539.3	995.8
de novo synthesis	65	0.3	37.0	5.1	66	9.0	819.3	147.5
Biotransformation	67	0	1057.9	244.8	65	13.0	4712.8	1885.8
Biotransformation + L-Phe	65	0	676.7	90.4	60	0	4789	2336.3

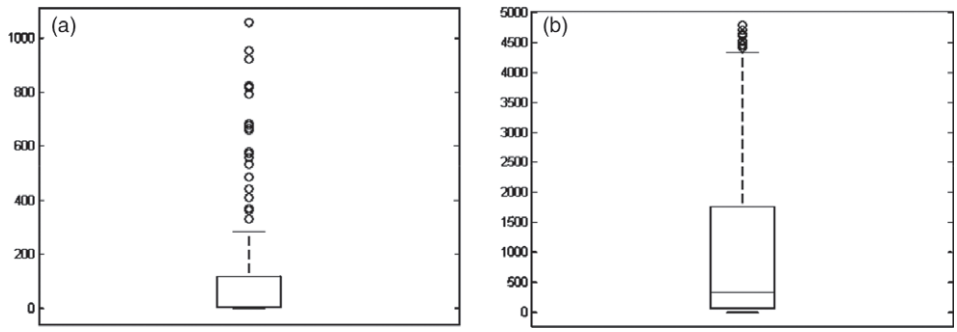

Figure 1. Boxplot analysis of the Y datasets: (a) L-Phe; (b) 2-PE.

Boxplot analysis

In order to identify possible Y data (L-Phe and 2-PE concentrations) outliers, a boxplot analysis was performed and is presented in Fig. 1. For this boxplot analysis, the maximum whisker length allowed was 1.5, i.e. 1.5 times the interquartile distance between the 25th and 75th percentiles. This resulted in the identification of samples falling outside a 99.3% coverage of a normally distributed data. Taking into consideration the results obtained, roughly 10% of the L-Phe dataset and 15% of the 2-PE dataset fell outside this limit. Given the fact that these values are significant, a Kolmogorov–Smirnov test was employed in order to determine whether the Y datasets were normally distributed. In fact, the Kolmogorov–Smirnov test for both the L-Phe and 2-PE datasets allowed confirmation that both distributions were not normal and thus the obtained results for the boxplot analysis had to be put into perspective. In that sense, and given that no singular values appeared to be quite apart from the rest, it was decided to proceed with further chemometric analysis without any outlier exclusion.

PCA analysis

A PCA analysis was further performed on the X dataset (wavelengths values), as depicted in Fig. 2 depicting the different 2-PE production conditions, namely the *de novo* synthesis, biotransformation, and biotransformation with the later addition of L-Phe. Analyzing Fig. 2(a), no distinction between the different conditions was noticeable; thus no interrelationship between the wavelength values (X dataset) and the different production conditions was apparent.

Figure 2(b) depicts the different *Y. lipolytica* strains (W29, CBS and NCYC). In both cases, the first, second and third principal components (or latent variables, LV) explained, respectively, 21.2%, 17.4% and 15.6% of the X dataset variability, for a total of 54.2%. The analysis of Fig. 2(b) led to the identification of strain-based sample interrelationships, regarding the X dataset, as a number of clusters presenting the homogeneous sample markers are evident. A closer analysis allowed us to establish that the stream (cluster) of gray circles (CBS strain) further apart from the remaining clusters (higher in the PC1 and PC3 axes) belonged to the samples representing experiment times of 96 h and above (thus with potential larger 2-PE and smaller L-Phe values). A second cluster, of white circles (NCYC strain), was also found to be quite apart from the remaining clusters (higher in the PC2 and PC3 axes) and represented the *de novo* synthesis and biotransformation conditions from 96 h and beyond in the experiment. However, it should be noted that the biotransformation + L-Phe samples, from 96 h and beyond, did not fall into this cluster. This may be because half of the points beyond 96 h were not available regarding the NCYC samples in the biotransformation + L-Phe conditions, and only two

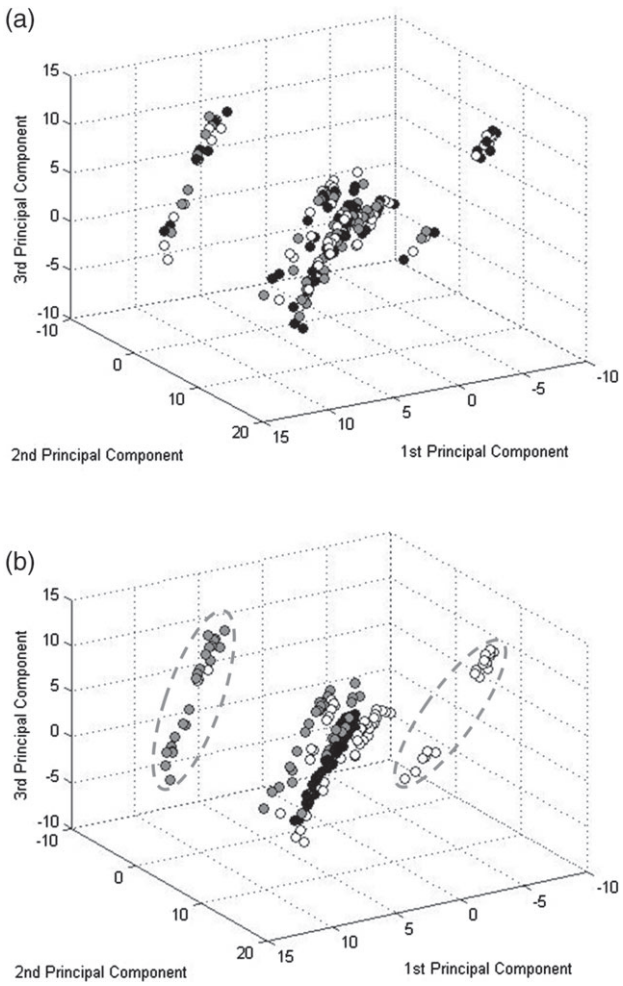

Figure 2. PCA analysis of the X dataset, showing the first (LV1), second (LV2) and third (LV3) latent variables. (a) Black circles represent the de novo synthesis dataset, gray circles represent the biotransformation + l-Phe dataset, and white circles represent the biotransformation dataset. (b) Black circles represent the W29 dataset, gray circles represent the CBS dataset, and white circles represent the NCYC dataset.

points (at 168 h) presented L-Phe values below 3000 mg L^{-1}, being, therefore, quite apart from the corresponding *de novo* synthesis and biotransformation samples (always below 1300 mg L^{-1}).

Although the PCA results pointed clearly towards a strain-based X dataset dependency, both the different conditions and strains dependencies were further studied in the subsequent PLS analysis, with three different datasets employed: [D1] – ensemble

Table 3. Equation (Eq.), coefficient of determination (R^2), root mean square error (RMSE, %), number of PLS components (*n*) and residual predictive deviation (RPD) of the PLS analysis for the studied model results regarding the L-Phe prediction

		Eq. (tr + val)	R^2 (tr + val)	RMSE (tr + val)	RMSE (val)	RPD (tr + val)	RPD (val)	*n*
D1	M1	$y = 0.933x$	0.8115	13.61	23.68	2.27	1.31	32
	M2	$y = 0.9355x$	0.7974	13.82	23.51	2.24	1.32	17
D2	M1	$y = 0.9749x$	0.8425	12.25	21.34	2.53	1.45	13 + 22 + 13
	M2	$y = 0.9621x$	0.8293	12.86	20.66	2.41	1.50	10 + 12 + 10
D3	M1	$y = 0.9935x$	0.9203	8.78	15.16	3.52	2.04	17 + 17 + 22
	M2	$y = 0.9992x$	0.8834	10.85	16.53	2.85	1.87	10 + 11 + 13

Tr-training (calibration); val- validation.

Table 4. Equation (Eq.), coefficient of determination (R^2), root mean square error (RMSE, %), number of PLS components (*n*) and residual predictive deviation (RPD) of the PLS analysis for the studied model results regarding 2-PE prediction

		Eq. (tr + val)	R^2 (tr + val)	RMSE (tr + val)	RMSE (val)	RPD (tr + val)	RPD (val)	*n*
D1	M1	$y = 0.9797x$	0.8069	9.18	15.95	2.20	1.26	30
	M2	$y = 0.9521x$	0.7802	9.74	16.50	2.07	1.22	19
D2	M1	$y = 1.0124x$	0.8893	6.99	12.12	2.89	1.66	14 + 17 + 14
	M2	$y = 0.9762x$	0.8712	7.56	11.86	2.67	1.70	12 + 10 + 10
D3	M1	$y = 1.0071x$	0.9491	4.72	8.23	4.27	2.45	17 + 19 + 20
	M2	$y = 0.9714x$	0.8978	6.60	8.86	3.06	2.28	11 + 10 + 8

Tr-training (calibration); val- validation.

dataset containing all the data samples; [D2] – strain-based partial datasets, divided into three groups representing the three different strains studied (W29, CBS and NCYC); and [D3] – conditions-based partial datasets, divided into three groups representing the three different conditions studied (*de novo* synthesis, biotransformation and biotransformation + L-Phe). Moreover, in Fig. 2, no significant sample outliers could be found and, therefore, the entire original data were employed (apart from 2-PE or L-Phe missing data points).

PLS regression

PLS regressions were performed on the L-Phe and 2-PE data obtained by HPLC analysis as the Y dataset and the FT-NIR spectra (14 000 – 200 cm^{-1}) as the X dataset. In order to determine the best model results for L-Phe and 2-PE, the obtained equation, coefficients of determination (R^2) and RMSE values (as a relative percentage of the sample range) were obtained and are presented in Tables 3 and 4.

Regarding L-Phe prediction, the best results were obtained by the raw dataset [M1] methodology with an R^2 value of 0.9203 and an RMSE value of 8.78% for the overall samples (and 0.7308 and 15.16%, respectively, for the validation samples), for the conditions-based [D3] dataset. In fact, the [M1] methodology could be considered, in the present case, to be slightly better than the iterative [M2] methodology, especially when concerning the overall samples. Furthermore, when comparing the ensemble [D1], strain-based [D2] and conditions-based [D3] datasets, the last one presented the best prediction results, whereas the ensemble [D1] attained the worst.

The same overall conclusions could be drawn for the 2-PE prediction, although with best prediction abilities. Again, the best results were also obtained by the raw dataset [M1] methodology, with an R^2 value of 0.9491 and an RMSE value of 4.72% for the overall samples (and 0.8391% and 8.23%, respectively, for the validation samples), for the conditions-based [D3] dataset. Once again, the [M1] methodology presented slightly better results than the iterative [M2] methodology, concerning the overall samples, although for the validation data this was not always the case. Confirming the previous L-Phe results, the conditions-based [D3] dataset presented the best prediction ability and the ensemble [D1] attained the worst. As such, the final methodology took into account the individual PLS analyses for each experimental condition, further assembled as a final single prediction.

The residual predictive deviation (RPD) was also calculated and is presented in Table 3. This parameter is defined as the ratio between the standard deviation (SD) of a population and the SE of cross-validation (SECV) for a prediction. An RPD value greater than 3 is considered fair and recommended for screening purposes.[26] The RPD values calculated for the best model regarding the L-Phe prediction ensemble data was 3.52 (2.04 for the validation data), whereas for the 2-PE prediction ensemble data it was 4.27 (2.45 for the validation data). Although falling short of the value 3 regarding the validation data, when considering the ensemble data the potential of the developed models for L-Phe and 2-PE prediction was confirmed.

The best prediction model results, regarding the raw dataset [M1] methodology, for the L-Phe and 2-PE concentrations, are presented in Fig. 3. In both cases there seems to be an overfitting of the calibration data, which can be explained by the shear amount of the X dataset original variables (3578 different wavelengths in both cases). This being the case, it can occur that the selected wavelengths for the chosen LVs regarding the calibration dataset may not be the best regarding the validation dataset. In fact, as the iterative [M2] methodology employs only a fraction of the X dataset variables (X dataset original variables transformed prior to fewer averaged wavelength values), it comes as no surprise that, regarding the validation RMSE values, better results were obtained, in most cases, for this methodology. Figure 4 presents an

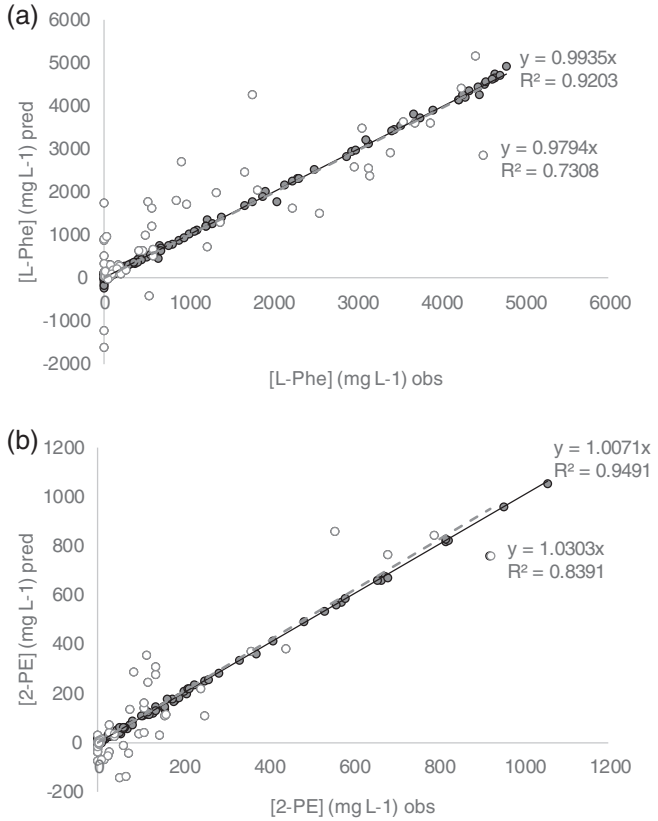

Figure 3. Best model results for l-Phe (a) and 2-PE (b) prediction. Gray circles represent the calibration data and white circles represent the validation data.

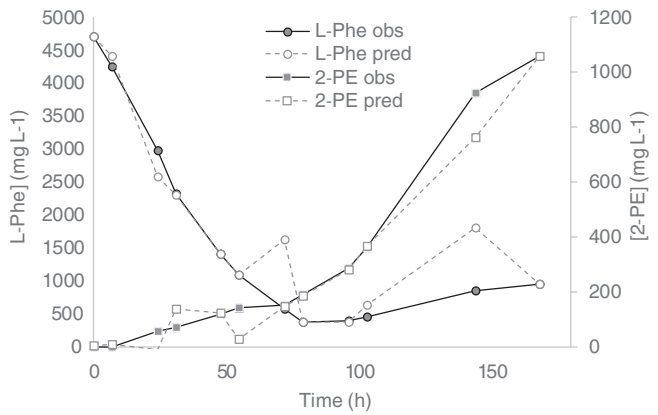

Figure 4. l-Phe (a) and 2-PE (b) concentrations for the W29 strain biotransformation experiment. Gray markers represent the observed values, white the predicted values, round l-Phe and square 2-PE.

example of the L-Phe and 2-PE concentration monitoring by HPLC and NIR for the W29 strain biotransformation experiment.

Over the past decade, many studies have explored the potential of NIR spectroscopy for bioprocesses monitoring. However, these studies' main focus relied on the quantitative monitoring of substrate consumption and biomass concentration.[5,9] The quantification of 2-PE and L-Phe by FT-NIR, in fermentative processes, is not usual, owing to the diversity and complexity of the fermentative matrix. Lorenzo *et al.*[13] tested the determination of fermentative volatile compounds in aged red wines by this technique and obtained an R^2 value of 0.36 for 2-PE estimation, for 240 samples and full cross-validation. Comparing these authors' results with those obtained by the current methodology, a large improvement is obtained regarding 2-PE determination. In fact, an R^2 value of 0.95 (considering both the training and validation samples) was obtained, which compares quite favorably with the 0.36 value of Lorenzo *et al.*[13] Another interesting study was developed by Ye *et al.*[14] These authors tested the detection of volatile compounds, including 2-PE, in apple wines using FT-NIR spectroscopy; the calibration results, using 42 different samples, presented an R^2 value of 0.84, to which, again the obtained value in the current study compares favorably.

The quantification of amino acids, as L-phenylalanine, using FT-NIR spectroscopy was also tested by a number of authors in different matrices. Escuredo *et al.*[27] studied the amino acid profile of the quinoa (*Chenopodium quinoa* Willd.) using NIR spectroscopy and chemometric techniques. Twelve amino acids (arginine, cystine, isoleucine, leucine, lysine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine and valine) were analyzed. For L-Phe a coefficient of determination (R^2) of 0.78 was obtained. Another work, developed by Shen *et al.*,[28] can be highlighted. These authors evaluated the amino acid content in Chinese rice wine by FT-NIR spectroscopy and found an R^2 value of 0.89 for L-Phe. Taking into consideration that the current study obtained an R^2 value of 0.92, for L-Phe determination – higher than that obtained by the previous authors – this reinforces the higher performance of the developed methodology.

It should be emphasized that this methodology presents several advantages over classical analytical techniques, offering a practical alternative to time-consuming methods such as liquid chromatographic techniques. Nowadays, NIR can be taken into consideration as a versatile technique, with no sample preparation, decreased costs and analysis time, and the ability to sample through glass and packaging materials.

CONCLUSIONS

The potential of NIR transmission spectroscopy was tested for the quantification of 2-PE and L-Phe concentrations during its production with *Y. lipolytica*. A chemometric approach was used employing first a PCA analysis for outlier removal and cluster identification. Next, a PLS analysis was performed in order to obtain a prediction model suitable for L-Phe and 2-PE monitoring purposes. This procedure resulted in relatively high coefficients of determination (R^2), and low RMSE, for the prediction ability of both compounds. Furthermore, the RPD was above three, showing its adequacy towards these compounds monitoring. Therefore, we believe this methodology to be of future practical implementation, for fast 2-PE and L-Phe monitoring in bioprocesses, given further robustness improvement, namely in dealing with overfitting issues.

ACKNOWLEDGEMENTS

The authors thank the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469 unit, COMPETE 2020 (POCI-01-0145-FEDER-006684) and BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 – Programa Operacional Regional do Norte. The authors also acknowledge financial

support to Cristina Quintelas through a postdoctoral grant (SFRH/BPD/101338/2014) provided by FCT – Portugal.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Supporting Information

Supporting information may be found in the online version of this article.

REFERENCES

1 Begnaud F and Chaintreau A, Good quantification practices of flavours and fragrances by mass spectrometry. *Philos Trans R Soc A Math Phys Eng Sci* **374**:20150365 (2016).
2 Husek P, Rapid derivatization and gas chromatographic determination of amino acids. *J Chromatogr A* **552**:289–299 (1991).
3 Calder AG, Garden KE, Anderson SE and Lobley GE, Quantitation of blood and plasma amino acids using isotope dilution electron impact gas chromatography/mass spectrometry with U-(13)C amino acids as internal standards. *Rapid Commun Mass Spectrom* **13**:2080–2083 (1999).
4 Qu J, Wang Y, Luo G, Wu Z and Yang C, Validated quantitation of underivatized amino acids in human blood samples by volatile ion-pair reversed-phase liquid chromatography coupled to isotope dilution tandem mass spectrometry. *Anal Chem* **74**:2034–2040 (2002).
5 Do Nascimento RJA, De Macedo GR, Dos Santos ES and De Oliveira JA, Real time and in situ near-infrared spectroscopy (NIRS) for quantitative monitoring of biomass, glucose, ethanol and glycerine concentrations in an alcoholic fermentation. *Braz J Chem* **34**:459–468 (2017).
6 López A, Arazuri S, García I, Mangado J and Jarén C, A review of the application of near-infrared spectroscopy for the analysis of potatoes. *J Agric Food Chem* **61**:5413–5424 (2013).
7 Tamburini E, Marchetti MG and Pedrini P, Monitoring key parameters in bioprocesses using near-infrared technology. *Sensors (Switzerland)* **14**:18941–18959 (2014).
8 Sandor M, Rüdinger F, Solle D, Bienert R, Grimm C, Groß S *et al.*, NIR-spectroscopy for bioprocess monitoring & control. *BMC Proc* **7**(Suppl 6):P29 (2013).
9 Cervera AE, Petersen N, Lantz AE, Larsen A and Gernaey KV, Application of near-infrared spectroscopy for monitoring and control of cell culture and fermentation. *Biotechnol Prog* **25**:1561–1581 (2009).
10 Tosi S, Rossi M, Tamburini E, Vaccari G, Amaretti A and Matteuzzi D, Assessment of in-line near-infrared spectroscopy for continuous monitoring of fermentation processes. *Biotechnol Prog* **19**:1816–1821 (2003).
11 Wu Z, Peng Y, Chen W, Xu B, Ma Q, Shi X *et al.*, NIR spectroscopy as a process analytical technology (PAT) tool for monitoring and understanding of a hydrolysis process. *Bioresour Technol* **137**:394–399 (2013).
12 Ripoll G, Vazquez M and Vilanova M, Ultraviolet–visible–near infrared spectroscopy for rapid determination of volatile compounds in white grapes during ripening. *Cienc Tec Vitivinic* **32**:53–61 (2017).
13 Lorenzo C, Garde-Cerdán T, Pedroza MA, Alonso GL and Salinas MR, Determination of fermentative volatile compounds in aged red wines by near infrared spectroscopy. *Food Res Int* **42**:1281–1286 (2009).
14 Ye M, Gao Z, Li Z, Yuan Y and Yue T, Rapid detection of volatile compounds in apple wines using FT-NIR spectroscopy. *Food Chem* **190**:701–708 (2016).
15 Burdock GA, *Flavor Ingredients*. 6th edn. CRC Press, Boca Raton, FL (2010).
16 Carlquist M, Gibson B, Yuceer YK, Paraskevopoulou A, Sandell M, Angelov AI *et al.*, Process engineering for bioflavour production with metabolically active yeasts: a mini-review. *Yeast* **32**:123–143 (2015).
17 Hua D and Xu P, Recent advances in biotechnological production of 2-phenylethanol. *Biotechnol Adv* **29**:654–660 (2011).
18 Angelov AD and Gotcheva V, Biosynthesis of 2-phenylethanol by yeast fermentation. *Sci Work UFT* **59**:490–495 (2012).
19 Etschmann M, Bluemke W, Sell D and Schrader J, Biotechnological production of 2-phenylethanol. *Appl Microbiol Biotechnol* **59**:1–8 (2002).
20 Albertazzi E, Cardillo R, Servi S and Zucchi G, Biogeneration of 2-phenylethanol and 2-phenylethylacetate important aroma components. *Biotechnol Lett* **16**:491–496 (1994).
21 Celińska E, Kubiak P, Białas W, Dziadas M and Grajek W, Yarrowia lipolytica: the novel and promising 2-phenylethanol producer. *J Ind Microbiol Biotechnol* **40**:389–392 (2013).
22 Cui Z, Yang X, Shen Q, Wang K and Zhu T, Optimisation of biotransformation conditions for production of 2-phenylethanol by a Saccharomyces cerevisiae CWY132 mutant. *Nat Prod Res* **25**:754–759 (2011).
23 Eshkol N, Sendovski M, Bahalul M, Katz-Ezov T, Kashi Y and Fishman A, Production of 2-phenylethanol from l-phenylalanine by a stress tolerant Saccharomyces cerevisiae strain. *J Appl Microbiol* **106**:534–542 (2009).
24 Genisheva Z, Quintelas C, Mesquita DP, Ferreira EC, Oliveira JM and Amaral AL, New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR). *Food Chem* **246**:172–178 (2018).
25 Einax J, Zwanziger H and Geiss S, *Chemometrics in Environmental Analysis*. Wiley-VCH, Weinheim, Germany (1997).
26 Fearn T, Assessing calibrations: SEP, RPD, RER and R2. *NIR news* **13**:12–14 (2002).
27 Escuredo O, González Martín MI, Wells Moncada G, Fischer S and Hernández Hierro JM, Amino acid profile of the quinoa (Chenopodium quinoa Willd.) using near infrared spectroscopy and chemometric techniques. *J Cereal Sci* **60**:67–74 (2014).
28 Shen F, Niu X, Yang D, Ying Y, Li B, Zhu G *et al.*, Determination of amino acids in Chinese rice wine by Fourier transform near-infrared spectroscopy. *J Agric Food Chem* **58**:9809–9816 (2010).

