Lutein, violaxanthin, and zeaxanthin spectrophotometric quantification: a machine learning approach

Victor Pozzobon[1] ✉ **and Cristobal Camarena-Bernard**[1]

[1]LGPM, CentraleSupélec, Université Paris-Saclay, SFR Condorcet FR CNRS 3417, Centre Européen de Biotechnologie et de Bioéconomie (CEBB), 3 rue des Rouges Terres 51110 Pomacle, France

A machine learning workflow was used to develop spectrophotometric equations quantifying chlorophyll *a*, *b*, lutein, violaxanthin, and zeaxanthin simultaneously. Microalgae samples were extracted in methanol following a classical procedure, and no chromatographic separation was applied. To do so, numerous samples with various pigments concentrations (obtained by HLPC) were gathered with their associated visible spectra. The data collected were used to calibrate a machine learning model based on partial least square regression. The best quantification (trade-off between accuracy and over-fitting) was obtained with a 7-feature model (one absorbance and six absorbance derivatives). From a practical perspective, the proposed model is not only calibrated but also validated. Therefore, the equations can readily be used for quantifying lutein, violaxanthin, and zeaxanthin (if high enough). They would significantly shorten the delay in obtaining samples' carotenoids concentrations compared to liquid chromatography while retaining adequate accuracy (below 10 %). Furthermore, the workflow is presented step-wisely so that other scholars may adapt it to their needs (e.g., producing a simpler model focusing only on one pigment). Finally, the data and source files are available in an online repository.

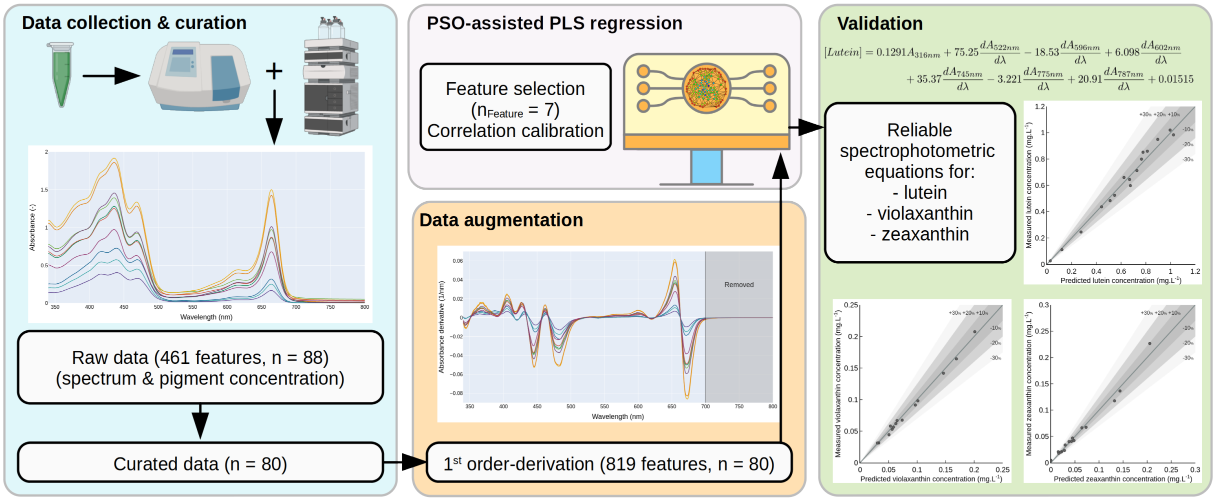
Pigments | Quantification | Spectrophotometry | Machine learning | Partial least square

Correspondence: ***victor.pozzobon@centralesupelec.fr***

1. Introduction

Nowadays, microalgae are regarded as small cell factories capable of producing molecules covering a wide range of applications (from pharmaceuticals to food and feed) (1). Among them, the pigments produced by microalgae are of great interest. They hold numerous beneficial properties ranging from proven antioxidant effects to alleged anti-cancer properties. While astaxanthin and β-carotene are the flagship carotenoid molecules, other carotenoids such as lutein and zeaxanthin, have demonstrated health benefits, a growing market, and an acknowledged need for production process improvement (2).

Indeed, lutein and zeaxanthin are one of the rare molecules capable of crossing the blood-brain barrier (3), which explains how they can access and positively affect the eye. For a long time, scientists were unsure of the additional systemic benefits associated with these carotenoids. Very recent studies (observational, interventional, and meta-analyses) have demonstrated them (4, 5). For example, lutein intake is associated with improved cognitive performance in all stages of life (from infant to elderly). The hypothesized mechanism is the following: plasma lutein crosses the blood-brain barrier and accumulates in the brain, where it acts as an antioxidant and prevents neuron damage. Consequently, neuronal efficiency increases, which translates into better memory, higher verbal fluency, and slowing down of Alzheimer's disease. Besides, systemic benefits, such as lower cardiovas-


Data collection & curation
+
Absorbance (-)
Wavelength (nm)
Raw data (461 features, n = 88)
(spectrum & pigment concentration)
Curated data (n = 80)
PSO-assisted PLS regression
Feature selection
($n_{Feature}$ = 7)
Correlation calibration
Data augmentation
Absorbance derivative (1/nm)
Removed
Wavelength (nm)
1st order-derivation (819 features, n = 80)
Validation
$[Lutein] = 0.1291 A_{316nm} + 75.25\frac{dA_{522nm}}{d\lambda} - 18.53\frac{dA_{596nm}}{d\lambda} + 6.098\frac{dA_{602nm}}{d\lambda} + 35.37\frac{dA_{745nm}}{d\lambda} - 3.221\frac{dA_{775nm}}{d\lambda} + 20.91\frac{dA_{787nm}}{d\lambda} + 0.01515$
Reliable spectrophotometric equations for:
- lutein
- violaxanthin
- zeaxanthin
+30% +20% +10%
-10%
-20%
-30%
Measured lutein concentration (mg.L^{-1})
Predicted lutein concentration (mg.L^{-1})
Measured violaxanthin concentration (mg.L^{-1})
Predicted violaxanthin concentration (mg.L^{-1})
Measured zeaxanthin concentration (mg.L^{-1})
Predicted zeaxanthin concentration (mg.L^{-1})

Graphical abstract

cular disease risk (6, 7), have been proven in animal models, and anti-cancer properties are being investigated *in vitro* (8–10).

Currently, lutein is extracted from marigold flowers, and zeaxanthin is produced by microbial biotechnology (11). The lutein production process is regarded as sub-optimal in various ways: it requires arable lands and large quantities of water, the production is seasonal, the associated work is tedious, and the produced lutein is esterified which lowers its bioavailability. Therefore, the relevance of its production by microalgal biotechnology is of interest (2). Various approaches have been investigated for lutein production by microalgae: photoautotrophy (12) and chemoheterotrophy (13–15), 1-stage and 2-stage (16–19), cultivations with light stress (20–22), nitrogen stress (23, 24), temperature stress (25), salinity stress (26), pH stress (27), or oxidative stress (28). In any case, a swift lutein quantification method is required. It is especially true in the context of actual industrial production, where cells should be harvested as soon as they reach the desired lutein content to avoid unnecessarily lengthy cultivations. Concomitantly to lutein production, microalgae also express zeaxanthin and its epoxidized form, violaxanthin. Therefore, a light stress cultivation strategy could promote lutein and zeaxanthin production (via VAZ cycle induction (29)). Nevertheless, such processes are expensive (energy cost of light) and require tight control. Therefore, there is a need for a rapid and reliable method to monitor these pigments' expression by microalgae.

Few techniques exist to quantify carotenoids once extracted from a sample. The first one is HPLC analysis. It is considered the reference as it separates the chemicals on a chromatographic column before detecting them. Carotenoids being a broad family of very similar molecules (30) and microalgae containing several of them, targeted pigments isolation is paramount. Sadly, HPLC techniques do not yield immediate results (a classical delay is between 30 minutes, if the apparatus is free, to several hours if a sequence is in progress). Furthermore, they require high-end equipment requiring sizable capital expenditure and expertise.

Alternatively, spectrophotometric methods exist and are well-established within microalgal biotechnology laboratories (31). While efficient, inexpensive, and easy to deploy, they can be regarded as suboptimal for two reasons. First, they only quantify the total carotenoid content. If lutein is of interest, this drawback can be circumvented by assuming that most of the carotenoids are lutein. One should note that this assumption only holds in particular cases (for example, unstressed green microalgae (32, 33)). Second, the wavelengths corresponding to pigments are selected beforehand based on obvious peaks in the absorbance spectra of each species of interest. This method raises the question of the relevance of the choice of the wavelength, especially for molecules exhibiting very similar spectra, such as carotenoids. To overcome these problems, authors have deployed several strategies in recent studies (not necessarily applied to microalgae). For example, when only a few carotenoid species are present, one can separate non-polar (*e.g.*, β-carotene) and polar (*e.g.*, lutein) carotenoids using hexane and dimethylformamide, respectively (34). Alternatively, if the sample matrix is relatively simple, it is possible to evaluate pigments' concentrations by comparing the sample spectrum to convoluted standards spectra (35). Nevertheless, none seems suited for microalgal samples as they host complex carotenoid mixtures with potentially nonnegligible matrix effects.

Therefore, this work aimed to apply tabletop spectrophotometers absorbance measurements to quantify lutein, violaxanthin, and zeaxanthin from microalgal samples, individually. The use of absorbance spectra for dissolved molecules quantification has been of attention in various fields of science over many years (36–38). Consequently, the scientific community produced streamlined workflows and mathematical background to support this technique. Classically, mixtures with known concentrations of the species of interest are created, and the corresponding spectra are acquired. Together, they constitute a dataset linking spectra with concentrations. Then a numerical model computing the correlations between the two types of data has to be chosen. Among the candidate models, such as Principal Component Analysis or MultiLinear Regression, Partial Least Square (or PLS) regression algorithm is of note (39) and has also proven successful in microalgal biotechnology applications (40, 41). Its particularity is that it uses principal component decompositions to create a set of components (linear combination of variables) associated with both input and output variables (through maximization of covariance between the scores). In this way, the most meaningful information is retained, making it a robust model (low sensibility to the training data) that handles well colinear inputs (when multiple variables provide the same information, for example, two neighboring wavelengths in the case of a spectrum) (42). Still, in the case of spectrophotometric readings processing, two metaparameters remain to be optimized by the operator: the number of components (boiling down to the number of wavelengths taken into account in the correlations) and the selected wavelengths themselves.

This article presents a machine learning-based technique producing a correlation linking visible spectra measurements to lutein, violaxanthin, and zeaxanthin concentrations. For the sake of completeness, even though it has been extensively covered by other authors, chlorophyll *a* and *b* concentrations were also derived. In this work, spectra and pigments' concentrations were obtained from various cultures. The strains used for these experiments were *Chlorella vulgaris* and *Scenedesmus almeriensis*. They have been grown under various modes: photoautotrophy, chemoheterotrophy, and mixotrophy. It yielded a rich dataset over a wide range of pigments' concentrations at different stages of the culture. The generated data were then used to power the machine learning workflow linking spectra and pigments' concentrations. Finally, the source files associated with this work are freely available in an online repository for anyone to download. This way, the interested reader could deploy the workflow and obtain correlations suiting her/his need (e.g., focusing on one pigment only, producing simpler correlations, ...).

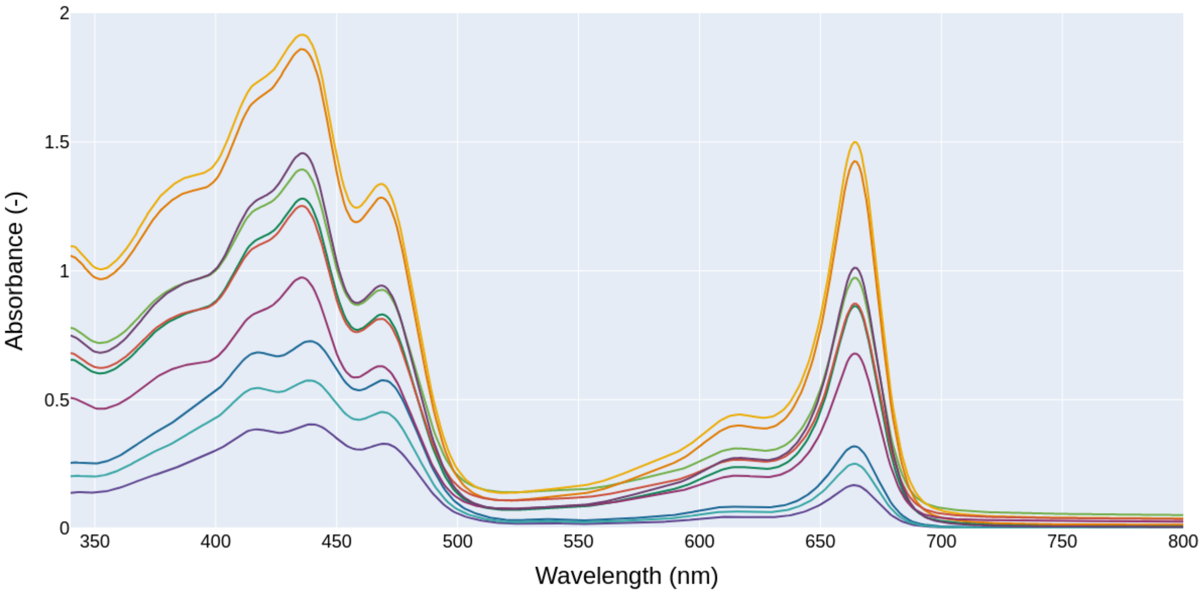


Fig. 1. Absorbance spectra of pigment extraction form 10 different biological samples (drawn randomly from the dataset)

2. Materials and methods

2 1. Cell cultivations

Samples coming from several experiments with different strains were used for this study. This choice was made for two reasons. First, it shortened the time required to agglomerate a large sample bank to conduct this study. Second, it allowed capturing of diverse pigment profiles, increasing the robustness of the obtained equations.

A first set of experiments was carried out using *Chlorella vulgaris* (CV 211-11b) (SAG Culture Collection, Germany). Cells were cultivated in photoautotrophy on B3N medium (43) under various light intensity (25 to 200 µmolPhotoPAR/m^2/s) and culture conditions (salt stress, nitrogen starvation, cold stress, ...). A second set of experiments was led using *Scenedesmus almeriensis* (kindly supplied by Pr. Gabriel Acien, from University of Almeria, Department of Engineering). This microalga was cultivated using B3N medium over different modes: photoautotrophy, chemoheterotrophy, and mixotrophy.

A total of 88 samples were produced throughout the different experimental campaigns.

2 2. Pigment spectra acquisition and HPLC quantification

For each sample, cells were washed twice by centrifugation (4 °C, 11000 rpm, 10 minutes). Biomass was then frozen and freeze-dried (1-day primary drying, 1-day secondary drying, Christ alpha 1-2 LD +). Biomass powder was stored in the dark at -20 °C before being used for pigment assays.

To quantify cell pigment content, 1 mg of freeze-dried microalgae powder was homogenized in 5 ml pure methanol using MP Biomedicals FastPrep42 bead miller. The suspension was cooked for 20 minutes at 60 °C (shaded from light) (44). This temperature was chosen as a trade-off between extraction enhancement and carotenoid degradation. On the one hand, the higher temperature, the faster the extraction. On the other hand, above a temperature of 60 °C, carotenoids are likely to undergo sizable oxidation (45). Furthermore, one should note that for microalgae that are not as recalcitrant to extraction as *Chlorella vulgaris*, the extraction could surely be conducted at room temperature. Additional measures could also be taken to prevent potential carotenoid degradation, such as operation under nitrogen or argon or the addition of an antioxidant such as BHT (46). The liquid containing the pigments was then filtered (0.22 µm), and its absorbance over the visible spectrum (340 - 800 nm, 1 nm resolution) was recorded (1 mL quartz cuvettes, Shimadzu UV-1800) (Fig. 1). The same liquid was stored in dark vials at 4 °C while waiting for its presentation to the HPLC analyzer for quantification.

Quantification of pigments was carried out on an Ultima 3000 HPLC (Thermo Fisher Scientific) coupled with a UV Detector. Separation was achieved on an Acclaim Polar Advantage II C18 column (4.6 × 150 mm, 3 µm, 120 Å) from Thermo Fisher Scientific. The column temperature was maintained at 30 °C. Pure methanol was the mobile phase. The flow rate was 0.5 mL/min, and the elution was set in isocratic mode. Injection volume was 5 µL, and the total run analysis was 40 minutes. Compounds were identified by comparing their retention time and their UV-Vis spectra with standard solutions. UV-Vis spectra were recorded from 200 nm to 700 nm. Absorbance was recorded at 400, 450, 500, and 650 nm. Pigments quantifications were led using the area of the peaks in external calibration for the most sensible of the recorded wavelength. External calibration concentrations ranged from 0.25 to 5 mg/l. Pigment standards and methanol were purchased from Sigma-Aldrich. Standards had a purity greater than 97 %. For each sample, the five pigments of interest (chlorophyll *a*, *b*, lutein, violaxanthin, and zeaxanthin) were reported systematically. 'N.A.' was used whenever one of them could not be detected or quantified. An example of the resulting chromatograph is a available as Supplementary Material.

2 3. Data management

The generated dataset consisted of 88 data points. These data points were composed of an absorbance spectrum from 340 to 800 nm (1 nm resolution), resulting in 461 input variables and measurements of chlorophyll *a*, *b*, lutein, violaxanthin, and zeaxanthin concentrations, resulting in 5 output variables.

As a first step, manual curation of the data was undergone. 7 data points were excluded as they exhibited a strong baseline deviation. 1 other data point was discarded because of doubt about the associated spectrum file name. The second step was to transform data so that all the values could be processed by the PLS algorithm. Indeed, pigment concentrations below the quantification limit were reported as 'N.A.'. The question of data replacement in the case of a value below the quantification limit for the PLS algorithm was already investigated in depth by other scholars (47). Their conclusions are clear when the actual value is known. One can use it to replace the machine reading. Otherwise, replacing the value with 0 is a safe procedure as it does not induce a bias and limits variance. In this case, the known value could not be accessed because of this biological origin. Thus, Schisterman *et al.* advice was followed and values below the detection limit were replaced with 0.

The dataset was shuffled and split into two subsets, one for training (80 % of the total, randomly drawn, n = 64) and one for validation (complementary 20 %, n = 16).

2 4. Partial least square calibration

As presented in the introduction, the partial least square model features two metaparameters to optimize. The first one is the number of components which, in this case, is equivalent to the number of wavelengths to be used to predict the concentrations (*e.g.* 6, 7, ... wavelengths in total). The second is the list of the particular wavelengths to be retained (*e.g.* 470 nm, 680 nm, ...).

2.4.1. General methodology. The first step was determining the number of wavelengths included in the final correlations. Given the number of concentrations to be predicted (5 in total), one could expect 6 wavelengths to be an adequate choice. Indeed, this would allow the algorithm to pick one wavelength per pigment and one to assess the noise level. Still, this approach may also be deemed too simplistic as a sample would contain many more molecules than the 5 quantified the HPLC analysis. Therefore, algorithms with 1 to 10 wavelengths were tested. Obviously, 1 wavelength is not enough, and 10 may lead to over-fitting, *i.e.* irrelevant wavelengths yielding marginal gain on the training dataset but inducing additional error for on-field data.

Once the number of wavelengths to include has been selected, the challenge is to determine which wavelengths are relevant (48). An option is to select the wavelengths manually. A naive approach would be to choose one for each absorption peak of the species. While it might be relevant for chlorophylls, it seems impossible to distinguish carotenoids this way. In addition, the question of selecting additional wavelengths in the case of correlation featuring more than 6 wavelengths remains intact. Another option is to use a numerical optimizer to select the most suitable set of wavelengths. In this case, second option was favored.

Finally, a scree plot (performance vs. number of wavelengths) was used to determine the best compromise between the number of wavelengths and potential over-fitting. One last criterion is the usability of the correlation. Indeed, 3-wavelength spectrophotometric correlations are common and easy to implement in a tabletop spectrophotometer user interface. Nevertheless, correlation with more than 8 wavelengths would require a computer to be processed on the fly, as the spectrophotometer used for this study could admit up to 8 wavelengths in the user-programmed correlation. While not problematic from a scientific point of view, the reader interested in reproducing and implementing the proposed approach should bear this point in mind and adapt it to its own lab equipment.

2.4.2. Wavelength selection optimization. The wavelength selection was led using a Particle Swarm Optimizer (PSO) (49). Indeed, the very high number of possible combinations (choosing 10 among 461 wavelengths leads to a bit more than 10^{20} possible combinations), the brute force approach laid out of the scope. The inclusion, or not, of a wavelength being a boolean value and considering the large dimensionality of the problem, gradient-based methods do not seem appropriate either. Stochastic methods, on the contrary, have been shown to cope well with such configurations. Among them, Particle Swarm Optimization is of note, as it is rather easy to implement, deploy on parallel architecture, and capable of browsing considerable search spaces. The technical details associated with this optimizer are described in the next section.

The PSO algorithm was implemented canonically (code available on the repository). The swarm social parameter (attraction towards the location of the best value of the swarm) was set at 0.6. The particle cognitive parameter (attraction towards the location of the best value encountered by the particle) was chosen as 0.6, and the inertia (autonomous exploration capability of the particle) was modeled using a random chaotic function (inertia = $0.1r_1+2r_2(1\text{-}r_2)$, with r_1 and r_2 random numbers). The swarm comprised a variable number of particles: 5000 × nb. of wavelengths. This choice widened the search for high dimensionality configurations while limiting the computational cost of low dimensionality configurations. Still, one should note that the size of the search space is a power function of dimensionality, not a linear. Therefore, the search space size increased faster than the size of the swarm. Still, it was observed that this drawback was counterbalanced by a qualitatively longer one exploration for high dimensionality configurations. Finally, swarm exploration was stopped after 50 iterations, for which the best swarm loss function (f) value did not diminish.

As the procedure aims to predict up to 5 concentrations, special care had to be taken in choosing the loss function

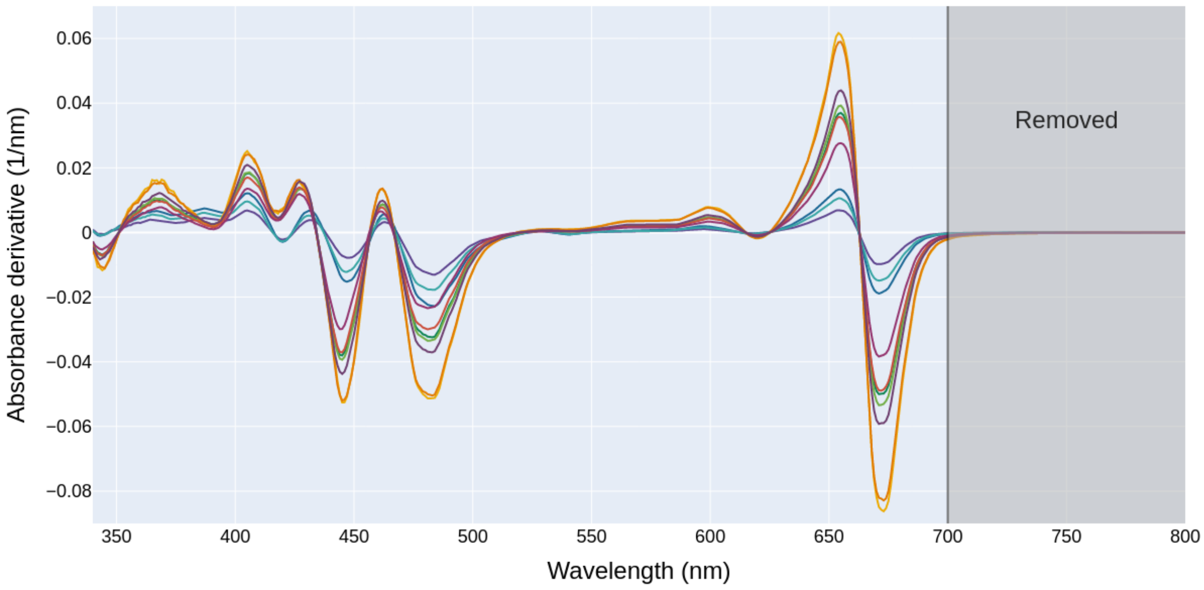


Fig. 2. Derivatives of absorbance spectra of pigment extraction form 10 different biological samples (drawn randomly from the dataset, same as previous figure)

associated with the optimizer. All concentrations being expressed with the same unit (mg/mL), it is likely that using a classical metric (e.g., Sum of Squared Error, Mean Square Error, Absolute Error, ...) would foster the prediction of highly concentrated pigments (chlorophylls) to the detriments of the more diluted ones (carotenoids). Therefore, the retained loss function weighted the squared deviation of a prediction for each pigment by the average value of its concentration over the dataset (Eq. 1).

$$f = \sum_{i=1}^{5}\sum_{j=1}^{64} \frac{(C_{pred,i,j} - C_{exp,i,j})^2}{\overline{C_{exp,i}}} \quad \textbf{(1)}$$

Where i represents the pigments (1 to 5 for chlorophyll *a*, *b*, lutein, violaxanthin, zeaxanthin), j the training set data point (from 1 to 64), and $\overline{C_{exp,i}}$ the average value of pigment i concentration over the dataset.

2 5. Preliminary results and data augmentation

The presented procedure was led successfully using raw spectra. It suggested using 8 wavelengths and yielded satisfactory results. However, the construction of the equations was troubling (Eq. 2). They featured surprising high coefficients with opposite signs for two sets of successive wavelengths (about 12 for A_{470nm}/A_{474nm} and 53 for A_{742nm}/A_{755nm}). These observations led us to believe that the algorithm was trying to access the derivative of the spectrum.

$$\begin{aligned}[Lutein] = 2.385A_{453nm} + 11.61A_{470nm} - 13.05A_{474nm} \\ - 5.047A_{607nm} + 1.329A_{610nm} - 0.04464A_{721nm} \\ - 51.80A_{742nm} + 55.90A_{755nm} + 0.02893 \quad \textbf{(2)}\end{aligned}$$

The use of the first derivative is a known preprocessing technique, especially when individual species spectra overlap (50, 51). The data set was augmented by providing the spectra derivatives in addition to the raw spectra. The derivation was led using a forward differencing with a 3 nm step (the best compromise between step size and curve smoothness). Figure 2 presents derivated spectra. As one can see, the region extending from 700 to 800 nm is essentially flat. Therefore, it was removed from the dataset to limit the number of features. Overall, the augmented dataset reached 819 input features per data point.

3. Results

3 1. Optimal number of wavelengths

Figure 3 presents the values of the loss function for the 100 PSO repetitions for the tested models (1 to 10 wavelengths features). As one can see, the spread of the results increases with the number of features used to build the correlations. This can be explained by the increase in the search space size. Indeed, each additional feature makes the task of the PSO optimizer harder and sometimes gets it to be trapped in a sub-optimal minimum. In addition, the results switch from a monomodal to a bimodal distribution if one includes 8 features or more. The second mode exhibits a higher average value of the loss function (about 3 mg^2/L^2 versus 1.75 mg^2/L^2). This behavior is a token of the PSO's difficulties in converging to a robust optimum. Furthermore, it is associated with the appearance of over-fitting. From a qualitative point of view, the appearance of the second mode raises serious doubts about the reproducibility of the proposed approach on other datasets.

In this context, it is complex to choose the appropriate number of features to include objectively. Including 7 features seems to be a relevant trade-off regarding accuracy and risk of over-fitting. To support this opinion, the loss function values were treated as random variables and analyzed statistically. An ANOVA test (p = 0.000) followed by Tukey's Honestly Significant Difference test (α = 0.05) was run to

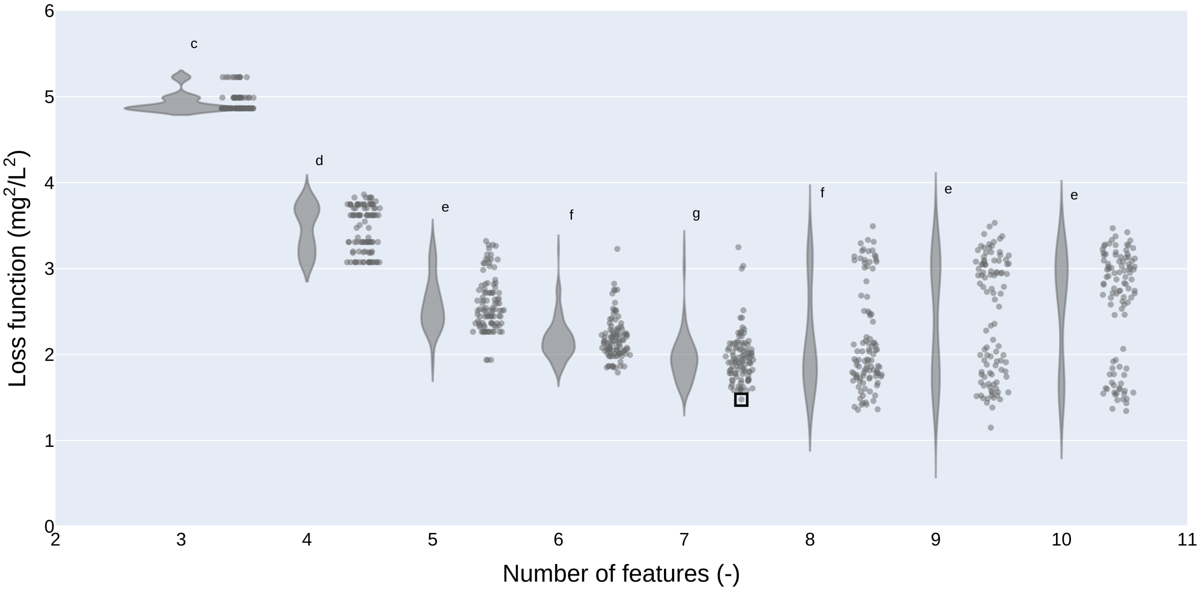


Fig. 3. Violin plot of the 100 PSO repetitions for the tested models (1 to 10 wavelengths features), zoomed onto 3 to 10-feature models. Dots - individual repetition of the optimization algorithm. Round shapes - observed density functions associated with the 100 PSO repetitions. Square - 7-feature model retained for the study. Compact letter display - ANOVA test (p = 0.000) followed by Tukey's Honestly Significant Difference test (α = 0.05). Average loss function value for 1-wavelength models: 23.18 mg^2/L^2. Average loss function value for 2-wavelength models: 13.05 mg^2/L^2

assess the statistical relevance of the observed difference between the models. As one can see, the 7-feature model is statistically different from the other, making its combination of low loss function value and moderate spread statistically significant. Therefore, the best-performing 7-feature model was retained for the rest of the study.

3 2. Analysis of the selected features

On top, figure 4 illustrates the input features (raw spectrum and its derivative) for a random sample. The numbers correspond to features retained by the model. On the bottom, figure 4 displays the frequency of appearance of the 819 features over the 100 calibrated 7-feature models. The first comment is that most retained features belong to the first-order spectrum derivative. This observation aligns well with the guess drawn from the analysis of the preliminary results. Among the selected features, 4 belong to the red peaks of the chlorophylls (numbers 1, 5, 6, and 7). 3 features (numbers 2, 3, and 4) belong to the blue part of the spectrum, where chlorophylls and carotenoids contribute to the absorbance. Surprisingly, no feature can be directly associated with pure noise (e.g., absorbance between 730 and 800 nm). Still, it can be hypothesized that by combining the absolute absorbance value (feature 1) and the derivatives, the model can reconstruct an estimate of the noise. Alternatively, the intrinsic noise level on the samples was low enough to be neglected, as could be asserted from raw spectra examination (fig. 1).

The distribution of the selected features is also of interest. The most striking comment is the importance of the absorbance derivative near the carotenoid peak around 470 nm (feature 3). This feature appears in 80 % of the calibrated models. Its pivotal role is not a surprise though as it is at the same wavelength as the classically used to quantify total carotenoids in the 3-wavelength models (31).

Apart from the surprising persistence of this feature, further insights can be derived from the analysis of the frequency graph. 6 areas of interest to the models can be identified. The 3 first ones can be deemed relatively minor. Indeed, given the reported frequencies, the models might have picked one among them. Those regions are the raw spectrum in the carotenoid region (around 460 nm, features index $\simeq$ 120), in the chlorophylls red region (around 650 nm, features index $\simeq$ 320), and in the far-red region (around 800 nm, features index $\simeq$ 460) - indicating that some models directly measured noise -. The 3 last ones can be deemed of major importance, as the models might have picked several features within them. Those regions are located in the derivative section of the features. They are the wide carotenoids region (features index $\simeq$ 520 to 650), the chlorophyll *b* region (features index $\simeq$ 720 to 760), and the chlorophyll *a* region (features index $\simeq$ 760 to 819). This comment further highlights the relevance of including the spectra derivative as input features for the calibration algorithm.

3 3. Correlations and validation

Once the model has been calibrated and analyzed, its internal coefficients can be retrieved to produce spectrophotometric correlations. The equations are provided hereinafter (Eq. 3 to 7). The computed pigment concentrations are in mg/mL. One can note that given the low magnitude of the derivative signal, the associated coefficients exhibit large values (hundreds or thousand). While surprising at first, it does not hinder the equations' usability.

Figures 5 compares the predicted and measured concentrations on the validation dataset for 5 pigments of interest. As one can see, most of the predictions fall within a ± 10 % interval around the measured value, few within a ± 20 % interval, and almost none lying more than 20 % away from the

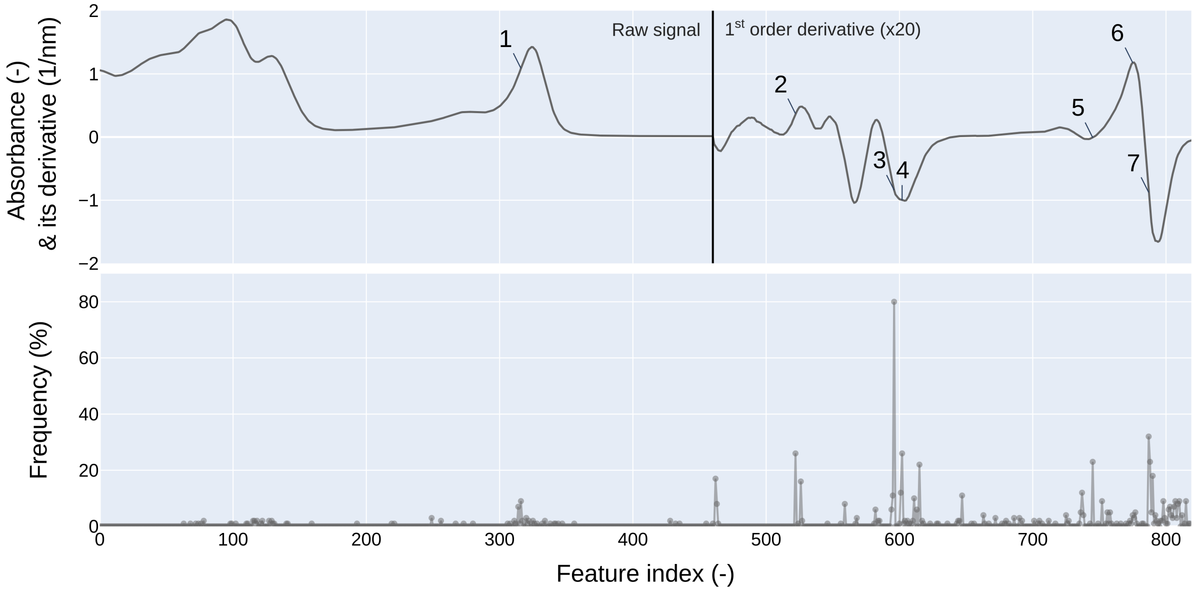


Fig. 4. Top - features originating from a randomly selected sample. On the left is the raw spectrum, and on the right is its first-order derivative (multiplied by 20 for the sake of readability). Number - features selected by the algorithm for the best 7-feature model. Bottom - feature apparition frequency throughout the 100 PSO repetitions for the 7-feature model

measured value. Furthermore, the spread around the first bisector is constant. In addition to being a token of the quality of the model, this means that the proposed model is capable of dealing indifferently with samples produced in photoautotrophy, chemoheterotrophy, or mixotrophy. These can be considered very satisfactory results. This comment is to be moderated for zeaxanthin. Indeed, as reported levels are very low (most below 0.05 mg/mL), a small absolute error leads to a significant relative error.

$$\begin{aligned}[Chlorophyll\ a] = -17.74A_{316nm} - 835.3\frac{dA_{522nm}}{d\lambda} \\ -211.8\frac{dA_{596nm}}{d\lambda} - 225.0\frac{dA_{602nm}}{d\lambda} - 3492\frac{dA_{745nm}}{d\lambda} \\ +191.9\frac{dA_{775nm}}{d\lambda} - 344.1\frac{dA_{787nm}}{d\lambda} - 0.8285\end{aligned} \quad \textbf{(3)}$$

$$\begin{aligned}[Chlorophyll\ b] = -9.609A_{316nm} - 447.8\frac{dA_{522nm}}{d\lambda} \\ -167.0\frac{dA_{596nm}}{d\lambda} - 59.44\frac{dA_{602nm}}{d\lambda} - 1336\frac{dA_{745nm}}{d\lambda} \\ -169.8\frac{dA_{775nm}}{d\lambda} - 535.7\frac{dA_{787nm}}{d\lambda} - 0.2848\end{aligned} \quad \textbf{(4)}$$

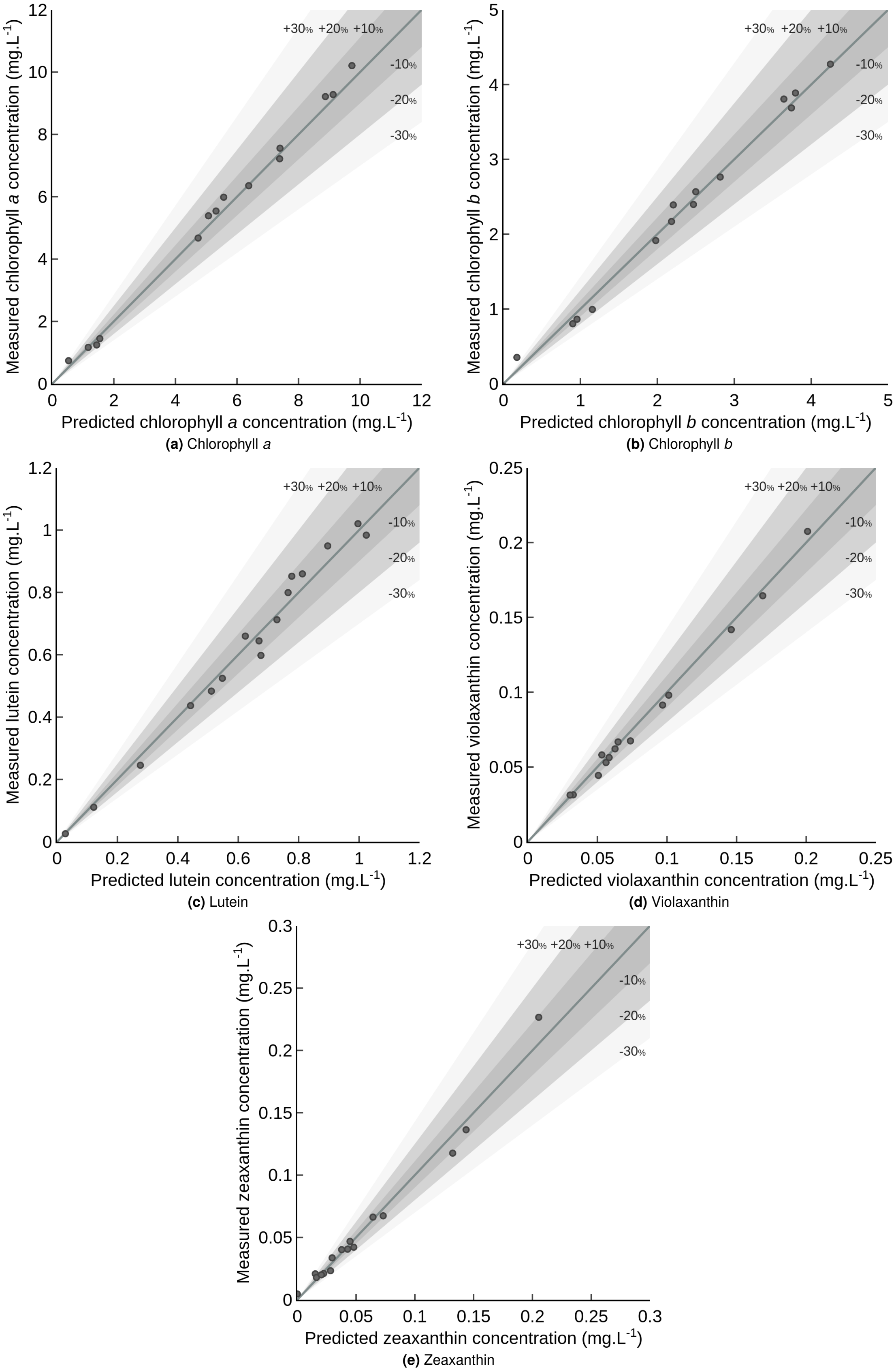
$$\begin{aligned}[Lutein] = 0.1291A_{316nm} + 75.25\frac{dA_{522nm}}{d\lambda} \\ -18.53\frac{dA_{596nm}}{d\lambda} + 6.098\frac{dA_{602nm}}{d\lambda} + 35.37\frac{dA_{745nm}}{d\lambda} \\ -3.221\frac{dA_{775nm}}{d\lambda} + 20.91\frac{dA_{787nm}}{d\lambda} + 0.01515\end{aligned} \quad \textbf{(5)}$$

$$\begin{aligned}[Violaxanthin] = -0.1085A_{316nm} - 1.99\frac{dA_{522nm}}{d\lambda} \\ -34.87\frac{dA_{596nm}}{d\lambda} + 15.49\frac{dA_{602nm}}{d\lambda} - 15.54\frac{dA_{745nm}}{d\lambda} \\ +2.383\frac{dA_{775nm}}{d\lambda} + 12.17\frac{dA_{787nm}}{d\lambda} - 0.0466\end{aligned} \quad \textbf{(6)}$$

$$\begin{aligned}[Zeaxanthin] = 0.2406A_{316nm} + 25.69\frac{dA_{522nm}}{d\lambda} \\ +53.37\frac{dA_{596nm}}{d\lambda} - 49.12\frac{dA_{602nm}}{d\lambda} - 62.46\frac{dA_{745nm}}{d\lambda} \\ -19.9\frac{dA_{775nm}}{d\lambda} - 3.703\frac{dA_{787nm}}{d\lambda} + 0.02645\end{aligned} \quad \textbf{(7)}$$

Once satisfactory agreement had been acknowledged, the next step was to investigate the residual error for each prediction. The size of the validation dataset (n = 16) was too small to assert with certainty that the errors followed a Gaussian distribution. However, they were centered on 0 and evenly spread on both sides. Therefore, it could be concluded that the proposed correlations are robust.

Finally, the quality of the prediction was compared to established models. Sadly, such models do not yield individual carotenoid pigments but the sum of xanthophyll and carotenes. Therefore, the comparison was led by summing the concentration of lutein, violaxanthin, and zeaxanthin obtained by HPLC, on one side, and our equations, on the other. Wellburn's correlations (Eq. 8, 9 and 10), for methanol as solvent, were used as classical spectrophotometric equations (31). As shown in Figure 6, both the proposed correlations and Wellburn's correlations yield values dispersed around the HPLC readings. The averaged dispersion is 48 % for the classical correlations and 9.7 % for the proposed ones. Still, one should note that the spectra resolution can partly explain the



(a) Chlorophyll *a*

(b) Chlorophyll *b*

(c) Lutein

(d) Violaxanthin

(e) Zeaxanthin

Fig. 5. Comparisons of the predicted and measured pigment concentrations on the validation dataset. Line: first bisector. Shaded areas: ±10, ±20 and ±30 % deviation

discrepancy obtained with the classical equations. Because of the low spectral resolution of the acquisition (1 nm), it was only possible to use the less accurate version of the correlations. With a higher spectral resolution (typically 0.2 nm), the classical correlations should perform better (by a factor 2, according to Wellburn's comment in his article).

$$Chl_a = 15.65\ A_{666nm} - 7.34\ A_{653nm} \tag{8}$$

$$Chl_b = 27.05\ A_{653nm} - 11.21\ A_{666nm} \tag{9}$$

$$Car_{x+c} = (1000\ A_{470nm} - 2.86\ Chl_a - 129.2\ Chl_b)/221 \tag{10}$$

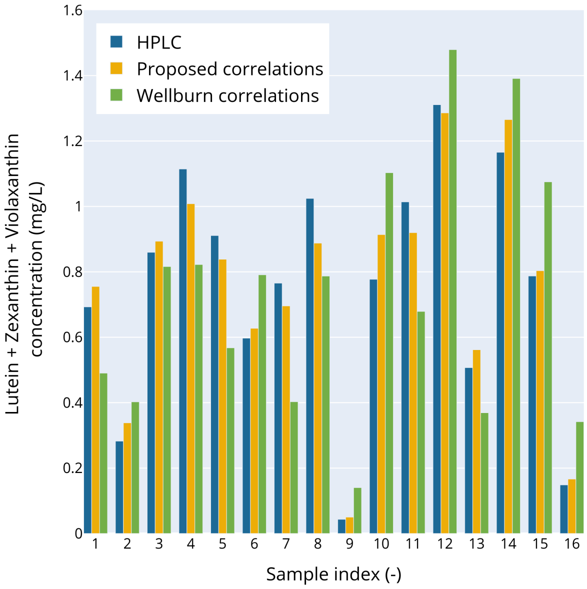


Fig. 6. Comparison between the sum of lutein, violaxanthin, and zeaxanthin concentrations on the validation dataset. Values were obtained by summing the concentrations of the 3 pigments from HPLC (blue), the proposed correlations (yellow), and Wellburn's equations for methanol as solvent (directly providing total carotenoids as xanthophyll + carotenes)

3 4. Practical applicability

Notwithstanding the quality of the proposed correlations, one should keep in mind that if chlorophyll *a*, *b*, and total carotenoids are sufficient information, then classical 3-wavelength models should be used. They are easier to handle and can be directly implemented into most spectrophotometers' built-in interfaces. If equipped with an integrating sphere, one could also use the extraction-free method elegantly proposed by Ritchie and Sma-Air (52). The proposed correlations are of interest whenever the focus is set on discriminating lutein, violaxanthin, and zeaxanthin. One should also bear in mind that zeaxanthin levels should be high (typically higher than 0.05 mg/mL) to claim a 10 % accuracy. If such is the goal, then one could extract pigments at room temperature under inert atmosphere to minimize the risk of carotenoid degradation. Finally, in terms of practical implementation, most spectrophotometers' software offer modules where they can compute spectra derivatives and combine features to yield concentrations. The proposed correlations are therefore usable for laboratory rapid and reliable quantification of lutein, violaxanthin, and zeaxanthin.

4. Conclusion

This article presented a machine learning workflow allowing the construction of spectrophotometric equations to simultaneously quantify chlorophyll *a*, *b*, lutein, violaxanthin, and zeaxanthin from microalgae sample. The pigments were extracted in methanol following a classical procedure, and no chromatographic separation was required. The quantification is based on seven features (one absorbance and six absorbance derivatives). From a practical perspective, the proposed model is not only calibrated but also validated. Therefore, the equations can readily be used for quantifying lutein, violaxanthin, and zeaxanthin (if high enough). They would significantly shorten the delay in obtaining samples' carotenoids concentrations compared to liquid chromatography while retaining adequate accuracy (below 10 %). Furthermore, the workflow is presented step-wisely so that other scholars may adapt it to their needs (e.g., producing a simpler model focusing only on one pigment). Finally, the data and source files are available in an online repository.

Funding

This study was carried out in the Centre Européen de Biotechnologie et de Bioéconomie (CEBB), supported by Région Grand Est, Département de la Marne, Greater Reims and the European Union. In particular, the authors would like to thank Département de la Marne, Greater Reims, Région Grand Est and European Union with European Regional Development Fund (ERDF Champagne Ardenne 2014-2020) for their financial support to the Chair of Biotechnology of CentraleSupélec.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Pr. Gabriel Acien, from University of Almeria, Department of Engineering, for kindly supplying the *Scenedesmus almeriensis* used in our laboratory.

PLS implementation was directly drawn from Scikit Learn 1.1.2 machine learning framework (53).

Availability of data, material and code

The datasets and codes generated during and/or analysed during the current study are available in a GitHub repository, https://github.com/victorpozzobon/pigmentQuantificationSpectrophotometry

Authors' contribution

CC and VP initiated and designed the study. CC and VP led the experimental work. VP led the numerical work.

All the authors critically interpreted the results. VP drafted the manuscript, CC corrected it. All authors approved the manuscript.

References

1. Wendie Levasseur, Patrick Perré, and Victor Pozzobon. A review of high value-added molecules production by microalgae in light of the classification. *Biotechnology Advances*, page 107545, April 2020. ISSN 0734-9750. .
2. Mario Ochoa Becerra, Luis Mojica Contreras, Ming Hsieh Lo, Juan Mateos Díaz, and Gustavo Castillo Herrera. Lutein as a functional food ingredient: Stability and bioavailability. *Journal of Functional Foods*, 66:103771, March 2020. ISSN 1756-4646. .
3. Diego Gazzolo, Simonetta Picone, Alberto Gaiero, Massimo Bellettato, Gerardo Montrone, Francesco Riccobene, Gianluca Lista, and Guido Pellegrini. Early Pediatric Benefit of Lutein for Maturing Eyes and Brain—An Overview. *Nutrients*, 13(9):3239, September 2021. ISSN 2072-6643. . Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.
4. James M. Stringham, Elizabeth J. Johnson, and B. Randy Hammond. Lutein across the Lifespan: From Childhood Cognitive Performance to the Aging Eye and Brain. *Current Developments in Nutrition*, 3(7), July 2019. . Publisher: Oxford Academic.
5. Silvio Buscemi, Davide Corleo, Francesco Di Pace, Maria Letizia Petroni, Angela Satriano, and Giulio Marchesini. The Effect of Lutein on Eye and Extra-Eye Health. *Nutrients*, 10(9): 1321, September 2018. . Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.
6. Barbara Demmig-Adams, Marina López-Pozo, Jared J. Stewart, and William W. Adams. Zeaxanthin and Lutein: Photoprotectors, Anti-Inflammatories, and Brain Food. *Molecules*, 25(16):3607, January 2020. ISSN 1420-3049. . Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.
7. Fatemeh Hajizadeh-Sharafabad, Zohreh Ghoreishi, Vahid Maleki, and Ali Tarighat-Esfanjani. Mechanistic insights into the effect of lutein on atherosclerosis, vascular dysfunction, and related risk factors: A systematic review of in vivo, ex vivo and in vitro studies. *Pharmacological Research*, 149:104477, November 2019. ISSN 1043-6618. .
8. Jimi Kim, Jeonghee Lee, Jae Hwan Oh, Hee Jin Chang, Dae Kyung Sohn, Oran Kwon, Aesun Shin, and Jeongseon Kim. Dietary Lutein Plus Zeaxanthin Intake and DICER1 rs3742330 A > G Polymorphism Relative to Colorectal Cancer Risk. *Scientific Reports*, 9 (1):3406, March 2019. ISSN 2045-2322. . Number: 1 Publisher: Nature Publishing Group.
9. Wesam Mostafa Omar, Amr E. Ahmed, Mai Raslan, Khalid El-Nesr, Mamdouh Moawad Ali, Mohamed De Abdelmaksoud, and Dina El Dahshan. Effect of Lutein-Rich Extract on Human Cancer Cells. *Middle East Journal of Cancer*, 12(1):147–150, January 2021. ISSN 2008-6709. . Publisher: Shiraz University of Medical Sciences.
10. Yogendra Prasad Kavalappa, Sowmya Shree Gopal, and Ganesan Ponesakki. Lutein inhibits breast cancer cell growth by suppressing antioxidant and cell survival signals and induces apoptosis. *Journal of Cellular Physiology*, 236(3):1798–1809, 2021. ISSN 1097-4652. . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcp.29961.
11. Yating Zhang, Zhen Liu, Jianan Sun, Changhu Xue, and Xiangzhao Mao. Biotechnological production of zeaxanthin by microorganisms. *Trends in Food Science & Technology*, 71: 225–234, January 2018. ISSN 0924-2244. .
12. Jina Heo, Dong-Sik Shin, Kichul Cho, Dae-Hyun Cho, Yong Jae Lee, and Hee-Sik Kim. Indigenous microalga Parachlorella sp. JD-076 as a potential source for lutein production: Optimization of lutein productivity via regulation of light intensity and carbon source. *Algal Research*, 33:1–7, July 2018. ISSN 2211-9264. .
13. Hyun-Sik Yun, Young-Saeng Kim, and Ho-Sung Yoon. Effect of Different Cultivation Modes (Photoautotrophic, Mixotrophic, and Heterotrophic) on the Growth of Chlorella sp. and Biocompositions. *Frontiers in Bioengineering and Biotechnology*, 9, 2021. ISSN 2296-4185.
14. Yibo Xiao, Xi He, Qi Ma, Yue Lu, Fan Bai, Junbiao Dai, and Qingyu Wu. Photosynthetic Accumulation of Lutein in Auxenochlorella protothecoides after Heterotrophic Growth. *Marine Drugs*, 16(8):283, August 2018. ISSN 1660-3397. . Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.
15. Xian-Ming Shi, Xue-Wu Zhang, and Feng Chen. Heterotrophic production of biomass and lutein by Chlorella protothecoides on various nitrogen sources. *Enzyme and Microbial Technology*, 27(3):312–318, August 2000. ISSN 0141-0229. .
16. Jianhua Fan, Jianke Huang, Yuanguang Li, Feifei Han, Jun Wang, Xinwu Li, Weiliang Wang, and Shulan Li. Sequential heterotrophy–dilution–photoinduction cultivation for efficient microalgal biomass and lipid production. *Bioresource Technology*, 112:206–211, May 2012. ISSN 0960-8524. .
17. Liliana Flórez-Miranda, Rosa Olivia Cañizares-Villanueva, Orlando Melchy-Antonio, Fernando Martínez-Jerónimo, and Cesar Mateo Flores-Ortíz. Two stage heterotrophy/photoinduction culture of Scenedesmus incrassatulus: potential for lutein production. *Journal of Biotechnology*, 262:67–74, November 2017. ISSN 0168-1656. .
18. Hamdy Elsayed Ahmed Ali, Eman A. El-fayoumy, Wessam E. Rasmy, Ramadan M. Soliman, and Mohd Azmuddin Abdullah. Two-stage cultivation of Chlorella vulgaris using light and salt stress conditions for simultaneous production of lipid, carotenoids, and antioxidants. *Journal of Applied Phycology*, 33(1):227–239, February 2021. ISSN 1573-5176. .
19. Chun-Yen Chen, I-Chia Lu, Dillirani Nagarajan, Chien-Hsiang Chang, I-Son Ng, Duu-Jong Lee, and Jo-Shu Chang. A highly efficient two-stage cultivation strategy for lutein production using heterotrophic culture of Chlorella sorokiniana MB-1-M12. *Bioresource Technology*, 253:141–147, April 2018. ISSN 0960-8524. .
20. J. F. Sánchez, J. M. Fernández, F. G. Acién, A. Rueda, J. Pérez-Parra, and E. Molina. Influence of culture conditions on the productivity and lutein content of the new strain Scenedesmus almeriensis. *Process Biochemistry*, 43(4):398–405, April 2008. ISSN 1359-5113. .
21. Shih-Hsin Ho, Ming-Chang Chan, Chen-Chun Liu, Chun-Yen Chen, Wen-Lung Lee, Duu-Jong Lee, and Jo-Shu Chang. Enhancing lutein productivity of an indigenous microalga Scenedesmus obliquus FSP-3 using light-related strategies. *Bioresource Technology*, 152: 275–282, January 2014. ISSN 0960-8524. .
22. Dengjin Li, Yizhong Yuan, Dujia Cheng, and Quanyu Zhao. Effect of light quality on growth rate, carbohydrate accumulation, fatty acid profile and lutein biosynthesis of Chlorella sp. AE10. *Bioresource Technology*, 291:121783, November 2019. ISSN 0960-8524. .
23. Antonio Molino, Sanjeet Mehariya, Angela Iovine, Patrizia Casella, Tiziana Marino, Despina Karatza, Simeone Chianese, and Dino Musmarra. Enhancing Biomass and Lutein Production From Scenedesmus almeriensis: Effect of Carbon Dioxide Concentration and Culture Medium Reuse. *Frontiers in Plant Science*, 11, 2020. ISSN 1664-462X.
24. Lisa M. Schüler, Tamára Santos, Hugo Pereira, Paulo Duarte, N. Gangadhar Katkam, Cláudia Florindo, Peter S. C. Schulze, Luísa Barreira, and João C. S. Varela. Improved production of lutein and β-carotene by thermal and light intensity upshifts in the marine microalga Tetraselmis sp. CTP4. *Algal Research*, 45:101732, January 2020. ISSN 2211-9264. .
25. Mengyue Gong and Amarjeet Bassi. Investigation of Chlorella vulgaris UTEX 265 Cultivation under Light and Low Temperature Stressed Conditions for Lutein Production in Flasks and the Coiled Tree Photo-Bioreactor (CTPBR). *Applied Biochemistry and Biotechnology*, 183(2):652–671, October 2017. ISSN 1559-0291. .
26. Elisabeth Bermejo, María C. Ruiz-Domínguez, María Cuaresma, Isabel Vaquero, Adrian Ramos-Merchante, José M. Vega, Carlos Vílchez, and Inés Garbayo. Production of lutein, and polyunsaturated fatty acids by the acidophilic eukaryotic microalga Coccomyxa onubensis under abiotic stress by salt or ultraviolet light. *Journal of Bioscience and Bioengineering*, 125(6):669–675, June 2018. ISSN 1389-1723. .
27. Weiqi Fu, Giuseppe Paglia, Manuela Magnúsdóttir, Elín A. Steinarsdóttir, Steinn Gudmundsson, Bernhard Ø Palsson, Ólafur S. Andrésson, and Sigurður Brynjólfsson. Effects of abiotic stressors on lutein production in the green microalga Dunaliella salina. *Microbial Cell Factories*, 13(1):3, January 2014. ISSN 1475-2859. .
28. Dong Wei, Feng Chen, Gu Chen, XueWu Zhang, LongJun Liu, and Hao Zhang. Enhanced production of lutein in heterotrophic Chlorella protothecoides by oxidative stress. *Science in China Series C: Life Sciences*, 51(12):1088–1093, December 2008. ISSN 1862-2798. .
29. N. K. Choudhury and R. K. Behera. Photoinhibition of Photosynthesis: Role of Carotenoids in Photoprotection of Chloroplast Constituents. *Photosynthetica*, 39(4):481–488, December 2001. ISSN 03003604, 15739058. . Publisher: Photosynthetica.
30. Héctor Arvayo-Enríquez, Iram Mondaca-Fernández, Pablo Gortárez-Moroyoqui, Jaime López-Cervantes, and Roberto Rodríguez-Ramírez. Carotenoids extraction and quantification: a review. *Analytical Methods*, 5(12):2916–2924, 2013. . Publisher: Royal Society of Chemistry.
31. Alan R. Wellburn. The Spectral Determination of Chlorophylls a and b, as well as Total Carotenoids, Using Various Solvents with Spectrophotometers of Different Resolution. *Journal of Plant Physiology*, 144(3):307–313, September 1994. ISSN 0176-1617. .
32. Karen H. Wiltshire, Maarten Boersma, Anita Möller, and Heinke Buhtz. Extraction of pigments and fatty acids from the green alga Scenedesmus obliquus (Chlorophyceae). *Aquatic Ecology*, 34(2):119–126, June 2000. ISSN 1573-5125. .
33. Victor Pozzobon, Wendie Levasseur, Cédric Guerin, Nathalie Gaveau-Vaillant, Marion Pointcheval, and Patrick Perré. Desmodesmus sp. pigment and FAME profiles under different illuminations and nitrogen status. *Bioresource Technology Reports*, 10:100409, June 2020. ISSN 2589-014X. .
34. Lúcia Maia, Susana Casal, and M. Beatriz P. P. Oliveira. Validation of a Micromethod for Quantification of Lutein and β-Carotene in Olive Oil. *Journal of Liquid Chromatography & Related Technologies*, 31(5):733–742, February 2008. ISSN 1082-6076. . Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10826070701854139.
35. Nawel Achir, Adrien Servent, Marvin Soto, and Claudie Dhuique-Mayer. Feasibility of Individual Carotenoid Quantification in Mixtures Using UV-Vis Spectrophotometry with Multivariate Curve Resolution Alternating Least Squares (MCR-ALS). *Journal of Spectroscopy*, 2022:e4509523, February 2022. ISSN 2314-4920. . Publisher: Hindawi.
36. Erdal Dinç, Dumitru Baleanu, and Feyyaz Onur. Spectrophotometric multicomponent analysis of a mixture of metamizol, acetaminophen and caffeine in pharmaceutical formulations by two chemometric techniques. *Journal of Pharmaceutical and Biomedical Analysis*, 26 (5):949–957, December 2001. ISSN 0731-7085. .
37. Víctor Cerdà, Piyawan Phansi, and Sergio Ferreira. From mono- to multicomponent methods in UV-VIS spectrophotometric and fluorimetric quantitative analysis – A review. *TrAC Trends in Analytical Chemistry*, 157:116772, December 2022. ISSN 0165-9936. .
38. Yongnian Ni and Xiaofeng Gong. Simultaneous spectrophotometric determination of mixtures of food colorants. *Analytica Chimica Acta*, 354(1):163–171, November 1997. ISSN 0003-2670. .
39. Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2):109–130, October 2001. ISSN 0169-7439. .
40. Svein Jarle Horn, Einar Moen, and Kjetill Østgaard. Direct determination of alginate content in brown algae by near infra-red (NIR) spectroscopy. *Journal of Applied Phycology*, 11(1): 9–13, February 1999. ISSN 1573-5176. .
41. Anggara Mahardika, A. B. Susanto, Rini Pramesti, Hiroko Matsuyoshi, Bibin Bintang Andriana, Yusuke Matsuda, and Hidetoshi Sato. Application of imaging Raman spectroscopy to study the distribution of Kappa carrageenan in the seaweed Kappaphycus alvarezii. *Journal of Applied Phycology*, 31(2):1383–1390, April 2019. ISSN 1573-5176. .
42. Paul Geladi and Bruce R. Kowalski. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17, January 1986. ISSN 0003-2670. .
43. Robert A. Andersen. Algal Culturing Techniques Appendix A—Recipes for Freshwater and Seawater Media. In *Algal Culturing Techniques*. Academic Press, Burlington, Mass, 1 edition edition, February 2005. ISBN 978-0-12-088426-1.
44. Robert J. Porra. A simple method for extracting chlorophylls from the recalcitrant alga, Nannochloris atomus, without formation of spectroscopically-different magnesium-rhodochlorin derivatives. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1019(2):137–141, August 1990. ISSN 0005-2728. .
45. Nourhane Ahmad, Jihane Rahbani Mounsef, and Roger Lteif. A simple and fast experimental protocol for the extraction of xanthophylls from microalga Chlorella luteoviridis. *Preparative Biochemistry & Biotechnology*, 51(10):1071–1075, November 2021. ISSN 1082-6068. . Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10826068.2021.1901231.
46. Ming-Chang Chan, Shih-Hsin Ho, Duu-Jong Lee, Chun-Yen Chen, Chieh-Chen Huang, and

Jo-Shu Chang. Characterization, extraction and purification of lutein produced by an indigenous microalga Scenedesmus obliquus CNW-N. *Biochemical Engineering Journal*, 78: 24–31, September 2013. ISSN 1369-703X. .

47. Enrique F. Schisterman, Albert Vexler, Brian W. Whitcomb, and Aiyi Liu. The Limitations due to Exposure Detection Limits for Regression Models. *American journal of epidemiology*, 163 (4):374–383, February 2006. ISSN 0002-9262. .
48. Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, August 2012. ISSN 0169-7439. .
49. Federico Marini and Beata Walczak. Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 149:153–165, December 2015. ISSN 0169-7439. .
50. Silvia Bellato, Viviana Del Frate, Rita Redaelli, Daniela Sgrulletta, Remo Bucci, Andrea D. Magrì, and Federico Marini. Use of Near Infrared Reflectance and Transmittance Coupled to Robust Calibration for the Evaluation of Nutritional Value in Naked Oats. *Journal of Agricultural and Food Chemistry*, 59(9):4349–4360, May 2011. ISSN 0021-8561. . Publisher: American Chemical Society.
51. Abdolmajid Lababpour and Choul-Gyun Lee. Simultaneous measurement of chlorophyll and astaxanthin in Haematococcus pluvialis cells by first-order derivative ultraviolet-visible spectrophotometry. *Journal of Bioscience and Bioengineering*, 101(2):104–110, February 2006. ISSN 1389-1723. .
52. Raymond J. Ritchie and Suhailar Sma-Air. Using integrating sphere spectrophotometry in unicellular algal research. *Journal of Applied Phycology*, 32(5):2947–2958, October 2020. ISSN 1573-5176. .
53. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928.