Machine Learning with Applications 24 (2026) 100879

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

# A deep learning framework for spectrophotometric quantification of key microalgal pigments

Omar Bayomie [a,b,c], Victor Pozzobon [d,*]

[a] The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, London, UK
[b] SSPC, The Science Foundation Ireland Research Centre for Pharmaceuticals, University College Dublin, Belfield, Dublin, Ireland
[c] Institut des Systemes Intelligents et de Robotique, Sorbonne University, Paris 75005, France
[d] Université Paris-Saclay, CentraleSupélec, Laboratoire de Génie des Procédés et Matériaux, Centre Européen de Biotechnologie et de Bioéconomie (CEBB), 3 rue des Rouges Terres 51110 Pomacle, France

ARTICLE INFO

Dataset link: huggingface.co/PozzobonV/CNNspectroMicroalgae

*Keywords:*
Pigment quantification
Microalgae
Deep learning
CNN
Learning curve

ABSTRACT

This article presents a deep learning framework that links microalgae extract spectra to the quantification of individual pigments. To do so, it relies on Convolution Neural Network architectures. First, architectures from the literature were implemented and challenged. While showing good results for most pigments, they failed to predict adequately the zeaxanthin concentration (present in a very low amount, MAPE above 20%). Consequently, a specific network was designed. Upon finetuning, it reached 14.6% MAPE on the validation set. In addition to network architecture, data augmentation and preprocessing were explored. The results show that data augmentation by derivation alone (without extra preprocessing) yields the best results. Finally, the correlation between training dataset size and performance was investigated. Using the newly introduced learning curve tool, it was possible to evaluate the best achievable performance (3.10 to 8.57% MAPE) and convergence rate (approximately square root to quadratic, pigment-dependent) for the major pigments.

## 1. Introduction

When driving a microalgae process, pigments have always been of interest. First of all, because they are indicators of the well-being of the cells (for example, low nitrogen availability leads to chlorophyll breakdown, while repleting nitrogen drives cells regreening (Grimme & Porra, 1974; Pyliotis et al., 1975)). Second of all, because they are indicators of the market value of the biomass (the higher, the better) (Barros et al., 2019), which is especially true for carotenoids.

Indeed, for a long time, the health benefits of carotenoids have been alleged and investigated (O'Neill et al., 2001). While it is difficult to ascertain the effect of carotenoids in general, as this family of molecules holds more than 700 representatives (Arvayo-Enríquez et al., 2013), some specific molecules have emerged. Astaxanthin, which is primarily sourced from microalgae in its natural form, might be the flagship of this group, with its branding as the most potent natural antioxidant and historical applications as a feed supplement (Higuera-Ciapara et al., 2006). Still, this hype casts a shadow on other valuable carotenoid pigments, such as lutein. Indeed, over the past decade, scientists have demonstrated the health benefits of this molecule (against brain and eye aging, cardiovascular diseases, as pro-vitamin A, …). For an overall view, the kind reader is referred to the recent reviews on the topic (Buscemi et al., 2018; Stringham et al., 2019), or dedicated cohort studies (Min & Min, 2013), and *in vivo* investigations (Nazari et al., 2022).

Unfortunately, lutein is present in our diet in too small an amount (about 1.7 mg/day while between 6 and 14 mg/day would be needed to claim the full extent of its benefits) (Camarena-Bernard & Pozzobon, 2024). Therefore, dietary supplementation can be envisioned. Currently, lutein supplements are produced using Marigold flower (*Tagetes eracta linus*) (Bosma et al., 2003), whose production is deemed unsustainable (water-intensive, chemical-intensive, poor-quality labor), seasonal, and restricted to certain climates (China and Mexico mainly) (Ochoa Becerra et al., 2020).

In this context, microalgae have emerged as a potential avenue. Indeed, they feature several advantages, such as non-seasonal and non-location-specific production, medium to high added value bioeconomy jobs (Ronzon et al., 2022), and about 100 times higher lutein content (2 to 10 mg/g for microalgae versus about 1 to 100 μg/g for common vegetables (O'Neill et al., 2001)). Acknowledging this opportunity, scholars and engineers strive to unearth a viable microalgae-based lutein production process. Numerous strategies have been explored: photoautotrophy (Kona et al., 2021) and chemoheterotrophy (Shi et al.,

[*] Corresponding author.
*E-mail addresses:* omar.bayomie.23@ucl.ac.uk (O. Bayomie), victor.pozzobon@centralesupelec.fr (V. Pozzobon).

https://doi.org/10.1016/j.mlwa.2026.100879
Received 2 December 2025; Received in revised form 14 February 2026; Accepted 28 February 2026
Available online 1 March 2026

1999), 1-stage and 2-stage (Fan et al., 2012; Flórez-Miranda et al., 2017), as well as stresses (light, nitrogen, temperature, salinity, pH, or oxidative, as recently reviewed by Camarena-Bernard and Pozzobon (2024)).

For all these processes, a fast, reliable, and inexpensive lutein quantification method is required. Yet, the carotenoid quantification techniques are few, and those allowing for pinpointing a specific carotenoid are even fewer. They all start by extracting carotenoids in an organic solvent, as carotenoids are $C_{40}$, they are not water-soluble. Among the available methods, HPLC is considered the reference as it separates the compounds on a chromatographic column before quantifying them individually. Unfortunately, HPLC features several drawbacks. First of all, it implies sample destruction. Then, it is not immediate, as classical analysis lasts between 30 min (assuming the apparatus is free) and several hours if a sequence is in progress. Finally, this equipment is costly and requires expertise to operate and maintain.

Consequently, and for a long time now, the higher plants and microalgae scientific communities have used spectrophotometric methods as an alternative means to access total carotenoids information. Here, the works of Wellburn, Lichtenthaler, and Porra are to be acknowledged (Lichtenthaler & Buschmann, 2001; Lichtenthaler & Wellburn, 1983; Porra, 1990; Wellburn, 1994). Yet, while spectrophotometric methods are efficient, low-cost, and straightforward to implement, two limitations can be noticed.

Firstly, these methods assess only the total carotenoid concentration of the extract. Thus, when lutein (or any other specific carotenoid) is of interest, one has to assign a fraction of the total measurement to lutein (or else). In some case, this assumption can hold true for most of the unstressed conditions. This is, for example, the case for *Chlorella vulgaris* or *Scenedesmus obliquus* for whom lutein represents two-thirds of the total amount of carotenoids (Pozzobon et al., 2020; Wiltshire et al., 2000). But confidence in the results obtained with this approach is always questionable.

Secondly, to elaborate these methods, pigment-specific wavelengths are typically selected based on peaks in absorbance spectra for the considered species alone (Bennett & Bogorad, 1973). This approach raises concerns, especially in the case of carotenoid mixtures, as individual species exhibit similar spectra. To address these issues, different alternatives have been explored recently, though not always in the context of microalgal lutein production. For example, when only a few carotenoids are present, it is possible to separate non-polar carotenoids (*e.g.*, $\beta$-carotene) from polar ones (*e.g.*, lutein) using solvents like hexane and dimethylformamide, respectively (Maia et al., 2008). In another approach, when dealing with relatively simple matrices, pigment concentrations can be estimated by comparing the sample's spectrum with convoluted spectra of standards (Achir et al., 2022). However, these methods are generally unsuitable for microalgal samples, which typically contain complex mixtures of carotenoids and exhibit matrix effects that are difficult to untangle.

To alleviate these caveats, we previously developed a machine learning pipeline allowing to link HPLC readings and methanol extraction spectra (Pozzobon & Camarena-Bernard, 2022). Indeed, machine learning has been applied for a long time to this type of problem. For a review, the kind reader is referred to the excellent article (Cerdà et al., 2022), while specific examples will be detailed here to illustrate how the pipeline complexity grows with the data entanglement. First, when the peaks of the different chemical species are clearly isolated, Inverse Least Squares or Principal Component Regression are suited to recover the individual concentrations (Dinç et al., 2001). Then comes the first level of complexity, a few but overlapping species; in this case, MultiLinear Regression can still deliver good performance, but it has to be coupled with signal derivation (Lababpour & Lee, 2006). Finally, when species are numerous and individual spectra overlap, the Partial Least Squares (a.k.a. PLS) approach holds the upper hand, compared to Principal Component Regression and MultiLinear Regression. Furthermore, derivation (using the Savitzky-Golay algorithm) is of great help in further improving the results (PRESS score reduced from 0.32 to 0.11 using PLS, in the context of food colorant concentration determination) (Ni & Gong, 1997). In a nutshell, the previously introduced method is based on PLS (Wold et al., 2001) and Particle Swarm Optimizer (a.k.a. PSO) (Marini & Walczak, 2015) algorithms to identify the best subset of wavelengths to include in the process. PLS ensures good collinearity management (Geladi & Kowalski, 1986), while PSO allows overall optimization in a discontinuous search space (Bornatico et al., 2012). This pipeline was powered using a database containing numerous methanol extracts generated in our laboratory over the course of different projects (involving *Chlorella vulgaris* and *Scenedesmus almeriensis*, and both photoautotrophic and heterotrophic cultivations). All in all, it allowed the construction of seven-parameter equations for each of the quantified pigments (chlorophyll *a* and *b*, lutein, violaxanthin, and zeaxanthin). These equations yielded ± 10% accurate prediction, with zero-centered errors, except for zeaxanthin.

The present work takes up the torch of this first endeavor and explores how deep learning, through Convolutional Neural Networks (CNN), can improve the performance further.  The goals are to retain the individual pigment quantification feature, the major advantage of the PLS procedure previously introduced, while driving its accuracy closer to the HPLC one.

While such an attempt cannot be found in the literature, if one expands the scope beyond microalgae-related pigments, several contributions can be identified, especially in the IR spectra processing community. For example, Bjerrum et al. employed the technique to link spectra to tablet dosages in the context of pharmaceutical quality checks (Bjerrum et al., 2017). Their approach relied on minimal preprocessing (data scaling and Extended Multiplicative Scattering Correction, EMSC) and data augmentation. On top of achieving good raw performances (RMSE of 1.80 mg, for tablets containing up to 240 mg), the authors noted that the training of the CNN tends to select specific regions of the spectrum and perform derivatives on its own. They underlined how CNN can reconstruct human know-how to improve prediction (Ni & Gong, 1997). Cui et al. deployed a similar approach for wheat flour protein content determination and reached the same conclusion when analyzing the activation layers: the CNN computes spectra derivatives (Cui & Fearn, 2018). In addition, both teams agreed that PLS can be considered the gold standard to compare with (supporting the relevance of the comparison with our previous work). Finally, Wang et al. tried to improve the feature selection process and direct the CNN toward the most relevant features (Wang et al., 2022). Still, while relevant, this initiative is hampered by the lack of key implementation details. Overall, the comparison between machine learning and deep learning techniques is agglomerated in Table 1.

Unfortunately, these works cannot directly translate to our application. Indeed, key differences exist between the application cases introduced above and microalgae pigment extracts processing. The baseline deviation may be the main factor. Indeed, IR spectra used by the authors feature marked linear (or moderately quadratic) baseline shifts. Surprisingly, the authors did not preprocess their data by derivation (first- or second-order), which is known to be extremely efficient in this type of configuration (Ni & Gong, 1997). In addition, the datasets they used contained up to thousands of samples, which cannot be produced easily in the context of microalgae pigment quantification.

Given these differences, this paper will pay special attention to the need for baseline management techniques. In addition, it will emphasize how many samples are required to ensure high-quality predictions, as sample generation is a time-consuming process. Finally, all the data and scripts are available in an open-access repository. In terms of structure, this manuscript first introduces the dataset, its curation, augmentation, and preprocessing strategies. Then it presents the investigated CNN architectures and details the hyperparameter optimization workflow, including a pigment-specific adaptation. It continues by reporting the results (the influence of preprocessing, architecture, and data requirements). Finally, it discusses the results and compares them with conventional machine learning techniques, then provides insights into performance limits and dataset size requirements.

**Table 1**
Comparison between machine leaning and deep learning technique in view of their application to microalgal pigments quantification.

| | Machine learning | Deep learning |
|---|---|---|
| | (PLS, ILS, PCR, SVM, MLR) | (CNN) |
| Pros | No need (or minimal) for a case-specific adjustment<br>Fast training<br>Low RAM and GPU load | Feature selection<br>Non-linear processing<br>Modularity/Adaptability to the case at hand |
| Cons | No feature selection<br>No non-linear processing<br>SVM also does not scale well (specific techniques are required)<br>Early limitations of the simplest techniques (*e.g.*, ILS when spectra are entangled) | Difficult to design (architecture and hyperparameter range)<br>Complex (*e.g.*, learning rate management) and costly to train<br>High RAM and GPU load |

## 2. Dataset

### *2.1. Data acquisition & curation*

#### *2.1.1. Microalgal cells & cultivation*

Data were obtained from several experiments featuring pigment quantification within our laboratory. Different strains were used for this study. This choice was made for three reasons. First, it shortened the time required to agglomerate a large sample bank to carry out this study. Second, it allowed the capturing of diverse pigment profiles, increasing the robustness of the obtained equations. Third, it demonstrated the generalizability of the process (not restricted to one specific strain).

Part of the experiments involved *Chlorella vulgaris* (CV 211-11b) (SAG Culture Collection, Germany), which was cultivated in photoautotrophy on B3N medium (Andersen & Phycological Society of America, 2005) under different illuminations and culture conditions (salt stress, nitrogen starvation, cold stress, …). The other experiments were carried out with *Scenedesmus almeriensis* (kindly supplied by Pr. Gabriel Acien from the University of Almeria, Department of Engineering), which was cultivated using B3N medium under photoautotrophy, chemoheterotrophy, and mixotrophy.

### *2.2. Pigment spectra acquisition and HPLC quantification*

For each sample, cells were washed twice by centrifugation (4 °C, 11 000 rpm, 10 min). 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 beater. The suspension was cooked for 20 min at 60 °C (shaded from light) (Porra, 1990). The liquid 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 min. Compounds were identified by comparing their retention time and their UV spectra with standard solutions. UV spectra were recorded from 200 nm to 700 nm. Absorbance was recorded at 400, 450, 500, and 650 nm. Pigment quantifications were performed using the area of the peaks in external calibration for the most sensitive of the recorded wavelengths. 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.

### *2.3. Data structure & curation*

Once collected, the raw data were manually curated. 7 data points were excluded as they exhibited a substantial baseline deviation (*i.e.*, an absorbance value at 800 nm above 0.1, the first excluded sample had a value of 0.19). 5 data points were removed as they exhibited chaotic behavior upon derivation (*i.e.*, peak location deviation above 2 nm from the average or extreme noise). 1 data was excluded because of potential file mismanagement. 1 was excluded because of an error in the acquisition resolution. Overall, the generated dataset consisted of 75 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.

Afterward, the question of value imputation for undetected compounds (reported as 'N.A.' by HPLC) was addressed following the guidelines of other scholars (Schisterman et al., 2006). According to their conclusions, when the actual value is known, it can be used to replace the erroneous machine reading. Otherwise, replacing the value with 0 is a safe procedure as it does not induce a bias and limits variance. In our case, we do not have access to the known value because of its biological origin. Thus, we followed Schisterman et al. advice and replaced values below the detection limit with 0.

Finally, special care was taken in minimizing data similarity. Therefore, although most of these experiments were carried out as biological triplicates, either a single replicate or a duplicate (when sufficiently different) was included in the dataset. Furthermore, no analytical/technical replicate was included in the dataset. Overall, for duplicated runs (22 runs, *i.e.*, 44 points), when analyzing the pairwise absolute deviation, one obtains an average of 23.0% and a standard deviation of 40.4% (with a minimum of 0.2%, at the inoculation, and a maximum of 390.3%, 16-week aging stress).

### *2.4. Data augmentation*

Once the database has been acquired and curated, the question of potential data augmentation can be addressed. From the CNN literature

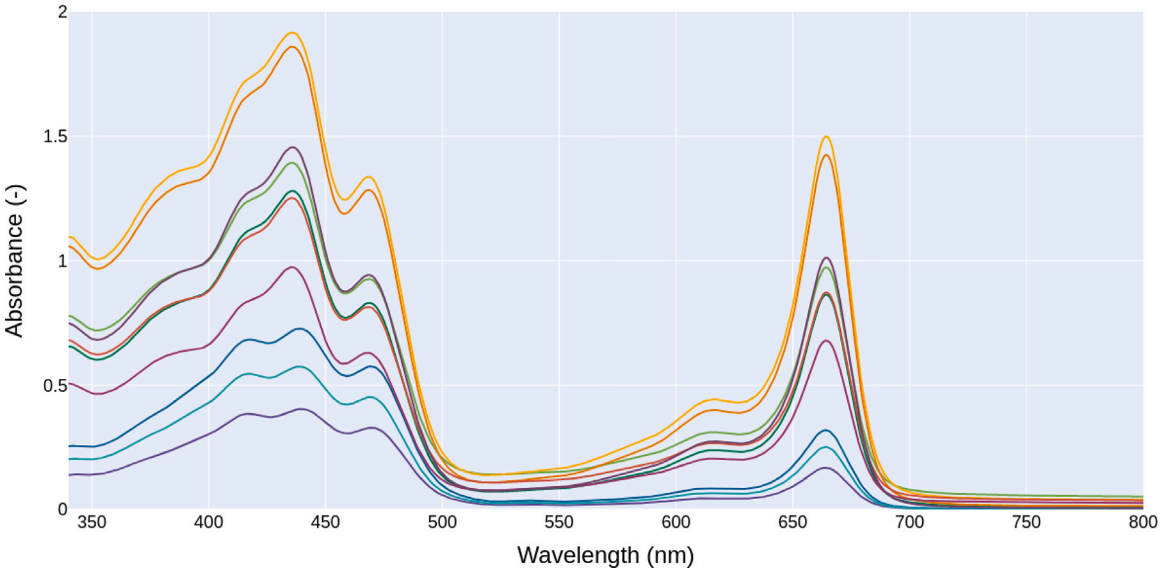


**Fig. 1.** Absorbance spectra of pigment extraction form 10 different biological samples (drawn randomly from the dataset)

survey exposed in the introduction, two avenues can be explored: derivation of the spectra (Bjerrum et al., 2017; Cui & Fearn, 2018) and increasing the number of spectra (Bjerrum et al., 2017). The first approach would aim at limiting the workload during the CNN training by allowing it to focus on feature selection (as opposed to feature derivation and then selection). Furthermore, the use of derivation has been documented as a good practice in the community for decades now (the kind reader is referred to the following review for an extended discussion on the topic (Cerdà et al., 2022)). The second approach (increasing the number of spectra in the dataset) is done by multiplying both spectra and associated compounds' concentrations by a random number between 0.5 and 1.5. The implicit assumption is the pure additivity of the compounds' contributions, with no non-linear effects.

These two options were investigated as preliminary actions for this work. The addition of derivatives (3 nm step, forward differentiating) to the feature set consistently improved the performance of the CNN (on average 2.51 ± $3.68\%_{Abs.}$ better Mean Absolute Percentage Error - MAPE -), and was therefore employed for the work presented below. On the contrary, artificially increasing the number of spectra by multiplication did not substantially improve the quality of the results, but substantially lengthened the training time. Given its elusive benefits and underlying assumption (even though weak), this data augmentation approach was not retained for this study.

### 2.5. Data preprocessing

When it comes to spectrum-to-concentration relations, it is common practice to use preprocessing methods to minimize noise while maximizing signal quality. Yet, the relevance of preprocessing techniques is case- and data-quality-dependent (*e.g.*, the presence of a baseline deviation). Our research investigated two preprocessing techniques (in addition to raw data usage) to establish their effects on model results.

#### 2.5.1. Raw data

First, this work used the unmodified spectral absorbance data (and associated first and second order derivatives) without any preprocessing technique (referred to as *Derivatives only*). This approach is intended to be considered as a baseline. It maintains all original spectral characteristics together with any noise or potential baseline shifts present in the data. Analyzing raw data gives us a starting point to determine whether advanced preprocessing steps actually improve the precision of the predictive model.

#### 2.5.2. Multiplicative scatter correction

Multiplicative Scatter Correction (MSC) represents a popular scattering correction spectroscopic technique which addresses additive and multiplicative effects within spectral data, firstly presented by Geladi et al. (1985), Sjöström et al. (1983). The MSC method adjusts each spectrum to an ideal reference spectrum (usually the average spectrum). Each spectrum $x_i$ (j as wavelength index) undergoes a series of transformations to remove unwanted physical variations while preserving chemical information. The process begins with mean centering, where

$$x_i^{\text{centered}} = x_i - \frac{1}{n}\sum_{j=1}^{n} x_{ij} \tag{1}$$

removes baseline offsets. A reference spectrum, with $m$ number of samples, is then calculated as

$$x_i^{\text{ref}} = \frac{1}{m}\sum_{i=1}^{m} x_i^{\text{centered}} \tag{2}$$

representing an ideal spectrum. For each centered spectrum, a linear regression against this reference follows the model

$$x_i^{\text{centered}} = b_i \cdot x_{\text{ref}} + a_i + e_i \tag{3}$$

where $b_i$ is the multiplicative coefficient related to scattering effects, $a_i$ is the additive coefficient related to baseline shifts, and $e_i$ contains the residual chemical information.

The final MSC-corrected spectrum is obtained by:

$$x_i^{\text{MSC}} = \frac{x_i^{\text{centered}} - a_i}{b_i} \tag{4}$$

which is a spectrum free of the estimated physical interferences while preserving the chemical signal contained in the residuals.

#### 2.5.3. Extended multiplicative scatter correction

Extended Multiplicative Scatter Correction (EMSC) builds upon the standard MSC framework by incorporating polynomial terms for chemical interferents and nonlinear baselines (Martens & Stark, 1991). From a mathematical perspective, a single $x_i$ raw spectrum is modeled as

$$x_i \;=\; b_{i,\text{ref}}\, x_{\text{ref}} \;+\; \sum_{j=0}^{P} b_{i,j}\, p_j \;+\; \sum_{k=1}^{K} b_{i,k}\, s_k \;+\; e_i.$$

where:

- $b_{i,\mathrm{ref}}$ is the multiplicative coefficient for the reference spectrum $x_{\mathrm{ref}}$.
- $p_j$ denotes a set of $P+1$ polynomial baseline basis vectors; *e.g.*, $p_0$ is a constant (all ones), $p_1$ is linear, $p_2$ is quadratic, and so on.
- $s_k$ represents the known spectrum of the $k$th chemical interferent.
- $b_{i,j}$ and $b_{i,k}$ are the coefficients for the baseline and interferent components in spectrum $i$.
- $e_i$ is the residual, representing the corrected chemical spectrum.

All $b$-parameters are estimated simultaneously via multiple linear regression (ordinary least squares). This leads to a vector representing the total modeled additive part to be removed:

$$A_i = \sum_{j=0}^{P} b_{i,j}\, p_j + \sum_{k=1}^{K} b_{i,k}\, s_k.$$

The final EMSC-corrected spectrum is:

$$x_i^{\mathrm{EMSC}} = \frac{x_i - A_i}{b_{i,\mathrm{ref}}}.$$

### 2.5.4. Centered-reducing scaling

Standardization or z-score normalization represents a data normalization preprocessing technique (Yan, 2025). It is more frequently encountered in the data processing community than in the spectroscopy one, but is advised by some authors in the context of CNN use (Bjerrum et al., 2017). In this work we used this technique that centers and reduces the spectral data. Each wavelength variable receives a transformation that brings both its mean to zero and its variance to one across all samples using this method:

$$x_{scaled} = \frac{x-\mu}{\sigma} \tag{5}$$

Where $\mu$ is the mean value for a given wavelength over the whole dataset and $\sigma$ is the associated standard deviation. Standardization enables all variables to operate in the same range by removing scale differences between wavelengths. Therefore, the high absorption and low absorption regions of the spectra show similar magnitudes after this preprocessing. The method demonstrates its best performance when absorbance signals at different wavelengths exhibit substantial variation, which is typical in pigment absorption spectra.

### 2.6. Data workflow

The dataset was shuffled and split into two subsets, one for training (80% of the total, randomly drawn) and one for validation (complementary 20%). The validation was used to assess the final performance of the optimized CNNs. On the contrary, the training set can be split between learning and testing subsets, randomly and repeatedly in the training and optimization process.

Furthermore, for the specific optimization workflow of the low-abundance pigment (zeaxanthin). A nested validation scheme was implemented, where the training pool of 80% of the dataset was used 5-fold cross-validation was used during optimization. An unseen hold-out validation dataset (20% of the original data) is used at the end to select the best model architecture.

## 3. Algorithms

Once the data, their curation, their augmentation, and their processing have been presented, the CNN algorithms using them are to be introduced.

### 3.1. Architectures derived from the literature

(Cui & Fearn, 2018) introduced a consolidated Convolutional Neural Network framework tailored for multivariate regression in the context of NIR calibration and redefined spectroscopic data exploration. Using automatic spectral preprocessing, CNN utilizes convolutional layers that autonomously identify effective spectral filters, eliminating the need for manual selection of preprocessing strategies. The configuration begins with spectral data entering through the initial input layer, proceeding through a convolutional preprocessing layer, followed by layers for extracting representative features and recognizing patterns, and concludes with an output layer for prediction.

The study included an extensive examination of activation mechanisms, revealing that Exponential Linear Units (ELU) outperformed Rectified Linear Units (ReLU), in terms of pure performance. Yet, they noted that a combination of ELU and ReLU attained the most balanced compromise between predictive precision and computational requirements in various testing conditions. In addition, the researchers illustrated how the application of L2 regularization, when paired with well-chosen parameters, aids in managing the structural complexity of the model and reducing noise in regression coefficients, resulting in more consistent and interpretable modeling outcomes.

The investigation led by Bjerrum et al. (2017) devised methods for the expansion of spectral data by embedding controlled offset and slope variations, as well as multiplicative transformations, to diversify the training set. This allowed the models to adapt to inherent fluctuations found in spectroscopic measurements. Their findings demonstrated that merging these augmentation methods with Extended Multiplicative Scatter Correction (EMSC) produced more favorable results than either method alone, although this synthesis initially appeared paradoxical. Furthermore, the incorporation of dropout for structural simplification was shown to be beneficial in minimizing overfitting, ensuring that the model performed reliably across different instruments and concentration settings. Also, in-depth visual assessments of CNN kernel responses were conducted, which indicated that the extracted filters emulate traditional preprocessing procedures such as smoothing, derivative calculation, and selection of specific regions. This highlighted a valuable bridge between long-established chemometric practices and newer data-driven methodologies.

### 3.2. Hyperparameters optimization

The study employs Optuna, a hyperparameter optimization framework, to explore neural network configurations across different preprocessing techniques. Optuna's design facilitates an efficient search across a high-dimensional hyperparameter space to identify model configurations that yield optimal performance, in this case, minimizing the MAPE for pigment prediction. Preliminary optimization runs for all pigments were conducted to narrow down the ranges of the hyperparameters shown below in and in Fig. 2 representing the highest range of attributes in Table 2 for visualization.

### 3.3. Pigment-specific architecture

As it will be shown in the Results section, the suggested architectures tend to produce high MAPE for pigments present in the lowest amounts. Indeed, for zeaxanthin, the high frequency of zero-concentration samples rendered the standard MAPE unstable which in some cases caused gradient explosions. To address this, we performed a dedicated hyperparameter search comparing robust loss functions, including Huber Loss, Mean Squared Error (MSE) and Log-Cosh Loss. In addition, in order to overcome the low prediction results, an alternative architecture is investigated.

Three neural network architectures were tested: (i) a fully connected dense feedforward network, (ii) a one-dimensional convolutional neural network with different blocks, and (iii) a hybrid architecture combining CNN and dense network architectures. The models included

**Table 2**
Overview of hyperparameters, their types, and value ranges, and selection strategies for the Optuna double convolutions study (applicable to Cui's and Bjerrum's architectures).

| Parameter | Type | Range | Sampling method |
|---|---|---|---|
| First kernel number | Integer | 15–30 | Uniform |
| First kernel width | Integer | 15–25 | Uniform |
| Second kernel number | Integer | 20–35 | Uniform |
| Second kernel width | Integer | 4–12 | Uniform |
| Fully connected layer size | Integer | 600–1000 | Discrete step |
| Dropout | Float | 0.08–0.2 | Uniform |
| $L_2$ regularization | Float | 0.05–0.15 | Uniform |
| Base learning rate | Float | 0.0008–0.003 | Uniform-Log scale |

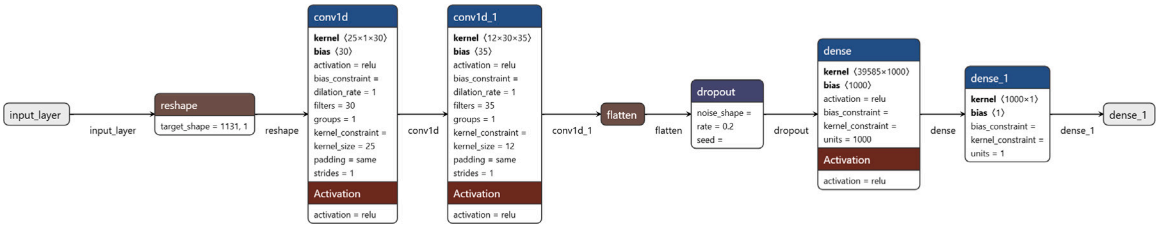


**Fig. 2.** Double convolutions CNN architecture and attributes used in the Optuna framework.

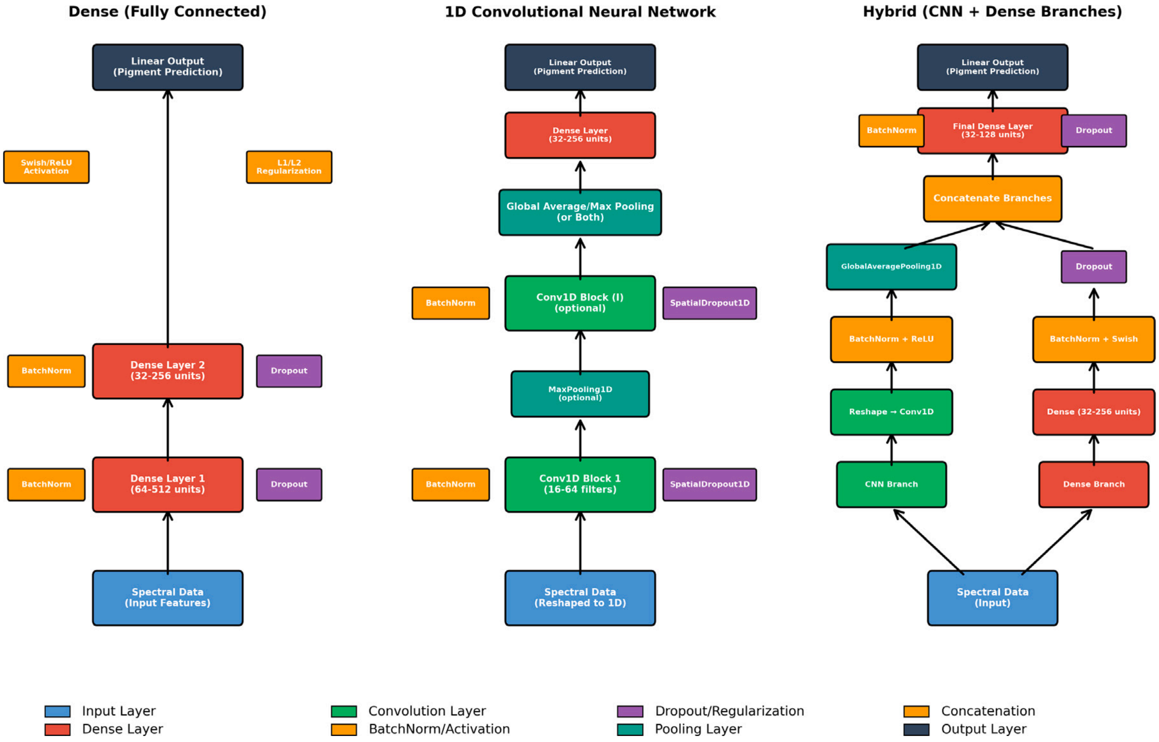


**Fig. 3.** Pigment-specific architectures workflow optimization.

multiple stabilization and regularization techniques through Batch Normalization, Dropout layers, and activation functions (swish, ReLU, ELU) that were determined through hyperparameter optimization. The CNN models used between one and three convolutional blocks, which allowed us to test different kernel sizes, filter numbers, and pooling methods (average, max, or concatenated). The hybrid model combined CNN and dense pathway outputs before adding more fully connected layers to the network (Fig. 3).

The data in the training set (80% of the original data) was further split into actual training (80%) and testing (20%). The unseen validation dataset (20% of the original data) is used at the end to find the best model architecture.

**Table 3**
Top models configurations for zeaxanthin configuration with the custom CNN architecture.

| Trial | Test MAPE | $R^2$ | RMSE | Model Type | Loss | Transform |
|---|---|---|---|---|---|---|
| 56 | 13.66 | 0.9300 | 0.01346 | Hybrid | Log-Cosh | Log |
| 96 | 13.93 | 0.9145 | 0.01489 | Hybrid | Log-Cosh | Log |
| 63 | 14.05 | 0.9392 | 0.01255 | Hybrid | Log-Cosh | Log |
| 74 | 14.41 | 0.9170 | 0.01467 | Hybrid | Log-Cosh | Log |
| 72 | 14.55 | 0.9261 | 0.01384 | Hybrid | Log-Cosh | Log |
| 32 (different run) | 14.58 | 0.9063 | – | Hybrid | Huber | Sqrt |

Finally, for a synthetic view of the process, with an algorithm pseudo-code presentation, the reader is referred to the Supplementary Materials.

## 4. Results

### 4.1. Training time

The study was performed on a GTX 1080Ti Nvidia GPU, with CUDA version 11.5 and TensorFlow version 2.19.0. 25 trials were set for each architecture and preprocessing technique parameterization. This resulted in a total computation time of 63.79 h, with an average trial duration of 6.12 min and around 12.76 h of optimization time per pigment. As one can see in Fig. 4 (left), the average training time was pigment dependent, with chlorophyll a being the fastest (around 3 min on average) and violaxanthin being the slowest (around 8 min on average). This behavior can be explained by the distribution of information on the spectra (Fig. 1). Indeed, the chlorophyll a spectrum is the most prominent one. Hence, identifying key features is a fast and efficient process. On the contrary, carotenoids are entangled. Hence, fine-tuning the hyperparameters requires more time.

Going one step further, it is possible to analyze how the performance of a given trial correlates with its training time (Fig. 4 (right)). Here, no clear trend emerges, as a longer training time does not lead to lower MAPE. Therefore, it can be concluded that the initial guess of Optuna is important in determining the final performance. This observation reinforces the need for a large number of trials during an Optuna optimization.

### 4.2. Influence of preprocessing

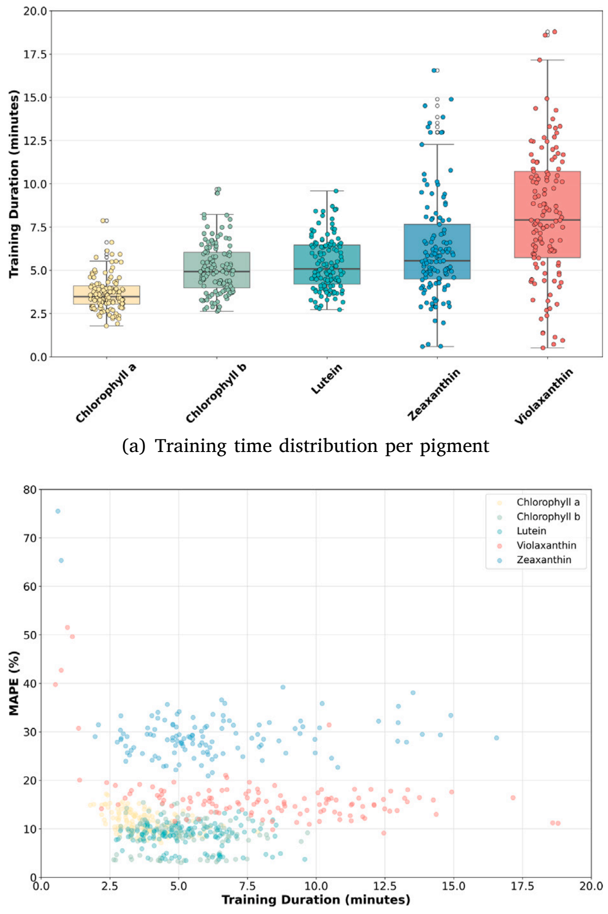
One of the primary goals was to study the effect of preprocessing on total performance. Therefore, Fig. 5 presents the best results obtained for each pigment and each preprocessing technique. As one can see, on average, the best results were achieved without any preprocessing. This finding propounds that the raw spectroscopic data (including first and second order derivatives) contained the most relevant information for pigment prediction. It is especially true for chlorophyll b and lutein, where no preprocessing resulted in twice better MAPE compared to other techniques.

### 4.3. PCA analysis of the hyperparameters

For each pigment, the best trial was identified (all coming from Bjerrum's architecture). Then, CNN parameters were assembled into a six-dimensional feature vector (first kernel number, first kernel width, second kernel number, second kernel width, fully connected layer size, dropout) and projected onto two principal components, $PC_1$ and $PC_2$, via PCA. The result is a two-dimensional representation graph with five points (one per pigment), where coordinates $(PC_1, PC_2)$ represent the original six-dimensional vector in a lower-dimensional space suitable for visualization.

The PCA clustering for the best hyperparameters (Fig. 6) shows three distinct groups where chlorophyll a and chlorophyll b appear together on the left side because they have similar optimal architectures. Yet carotenoids are separated into distinct groups, with zeaxanthin and lutein isolated from each other and from chlorophylls. Violaxanthin is positioned in the upper right corner. The model requirements for zeaxanthin appeared to differ by significant margin from the CNN architectures of other pigments except chlorophyll a and chlorophyll b (best performers with lutein). Since it lies in a distant position at the bottom of the plot. The separation between zeaxanthin and lutein remains significant even though they belong to the same carotenoid group. This observation aligns with the one on the training time and supports that transfer learning is not the right path for zeaxanthin. Therefore, another strategy, like a substantially different model and hyperparameters, would be more reasonable.

The analysis of architectural parameters reveals important patterns in feature activation. Different pigments benefit from different kernel



(a) Training time distribution per pigment

(b) Training time vs. performance

**Fig. 4.** Distribution of training time, as a function of the pigment, and versus performance (for all pretreatments, for both Cui's and Bjerrum's architectures)

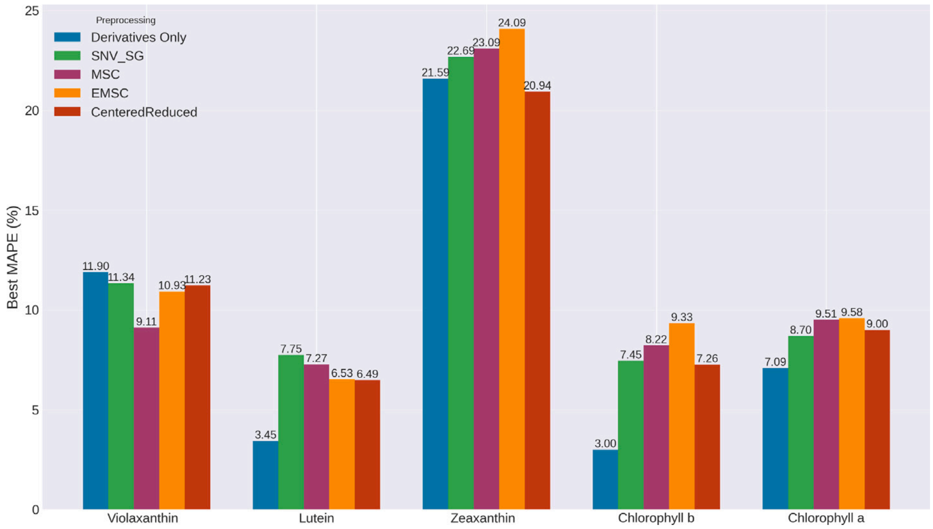


**Fig. 5.** Best performing trial per preprocessing technique for each pigment (Cui's and Bjerrum's architectures)

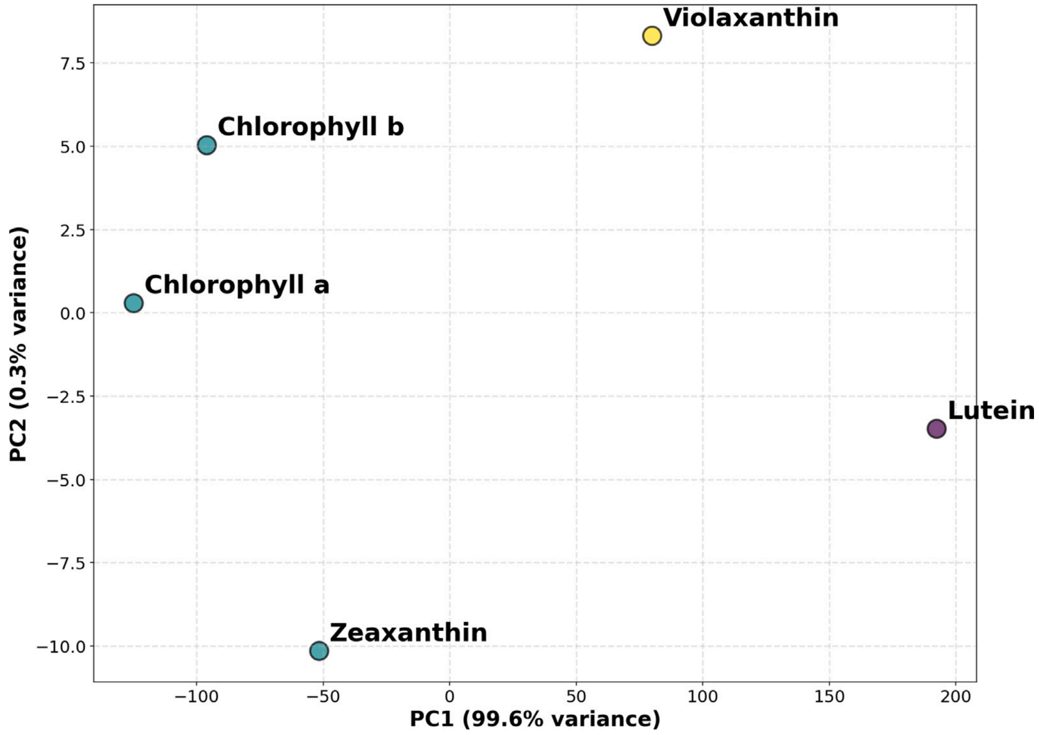


**Fig. 6.** PCA projection clustering on best performing parameters for each pigment.

sizes and numbers. Optimal fully connected layer sizes vary significantly by pigment. Base learning rates required fine-tuning for each pigment as well as for regularization, Dropout rates show pigment-specific optimal values.

### *4.4. Zeaxanthin architecture optimization*

Since zeaxanthin is the pigment present in the lowest amount, the associated MAPE tends to be the highest (almost 20%). It can easily be explained by the unfavorable signal-to-noise ratio compared to the other pigments. Consequently, reaching MAPE below 20% is a non-trivial challenge to overcome, which was tackled by a pigment-specific optimization workflow. The zeaxanthin optimization was performed over 100 trials using a 5-fold cross-validation scheme. A median pruner was applied to stop trials early when the intermediate cross-validation accuracy failed to improve relative to previous trials. By basing the pruning decisions solely on the internal cross-validation folds, the final holdout set remained completely isolated for the posterior performance assessment reported in Table 3. The maximum number of training epochs was 3000 before early stopping occurred. GPU memory growth was enabled to prevent allocation errors. Performance was assessed using multiple regression metrics: MAPE as the main criteria, Root Mean Squared Error (RMSE), and coefficient of determination ($R^2$). We treated feature scaling as a hyperparameter by comparing robust scaling (median and inter-quartile range normalization) and standard scaling (z-score normalization) to unscaled inputs. Target transformations through logarithmic and square-root transforms were also added as a hyperparameter. Finally, the three architectures presented in Fig. 3 were tested.

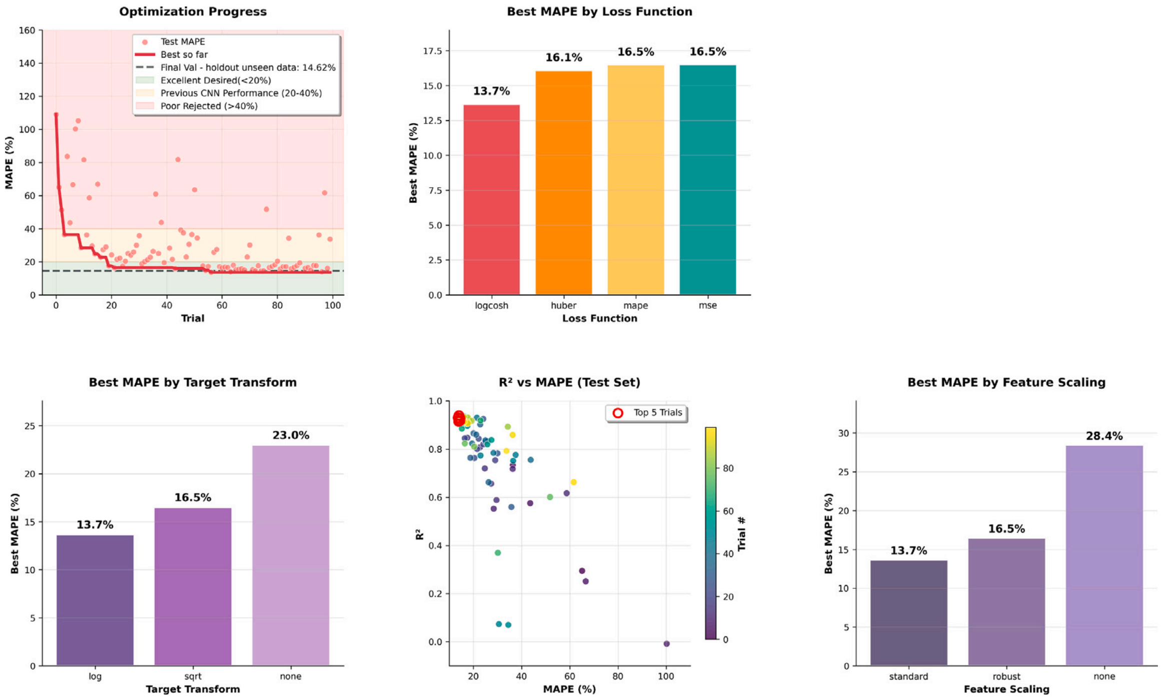


**Fig. 7.** Zeaxanthin optimization results.

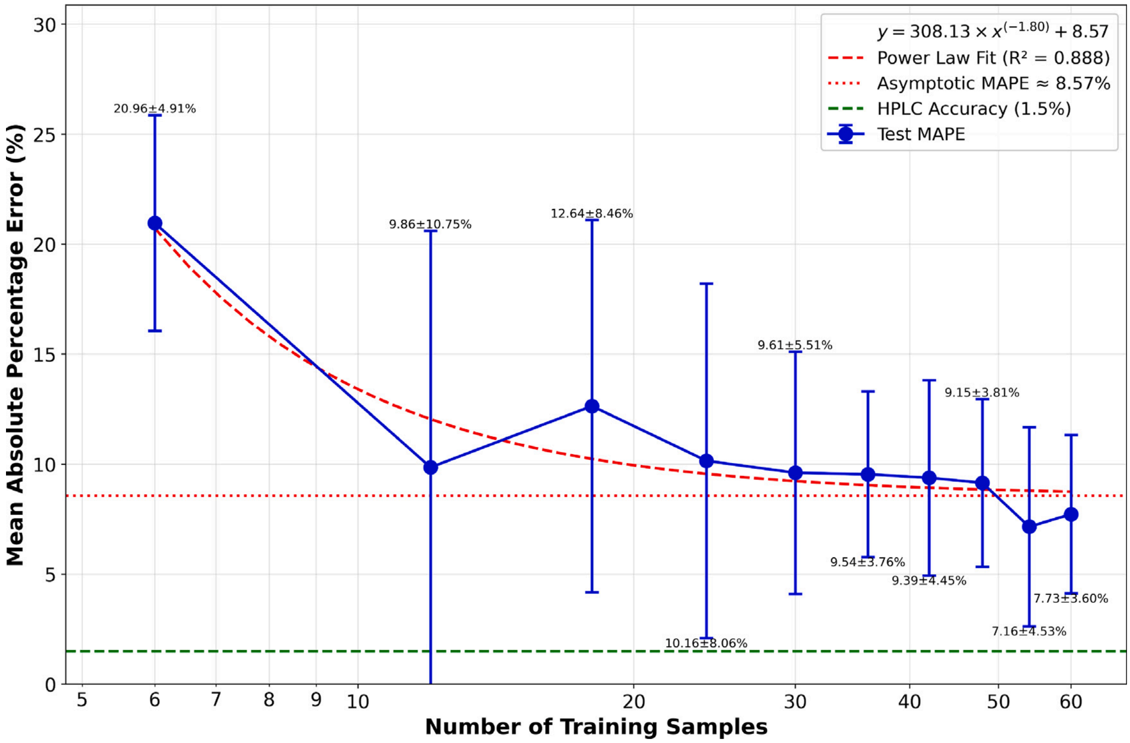


**Fig. 8.** Relationship between dataset size and predictive error for lutein. The blue line represents the observed error with variation, the red curve shows the fitted decay, and the dashed lines indicate limits and reference thresholds.

Results showed that among the three proposed architectures, the hybrid one produced the best results with the lowest MAPE (Table 3). The top model resulted in 14.62% MAPE on the unseen holdout validation and 13.66% on the test set. Overall screening results are illustrated in Fig. 7, which shows optimization progress.  These data shows that the Log-Cosh function was shown as the best performer for the final zeaxanthin model due to its smooth behavior near zero. Consequently, the optimal configuration for zeaxanthin differs from the others, utilizing a Log-Cosh loss and a logarithmic target transformation.

### *4.5. Size of the learning database*

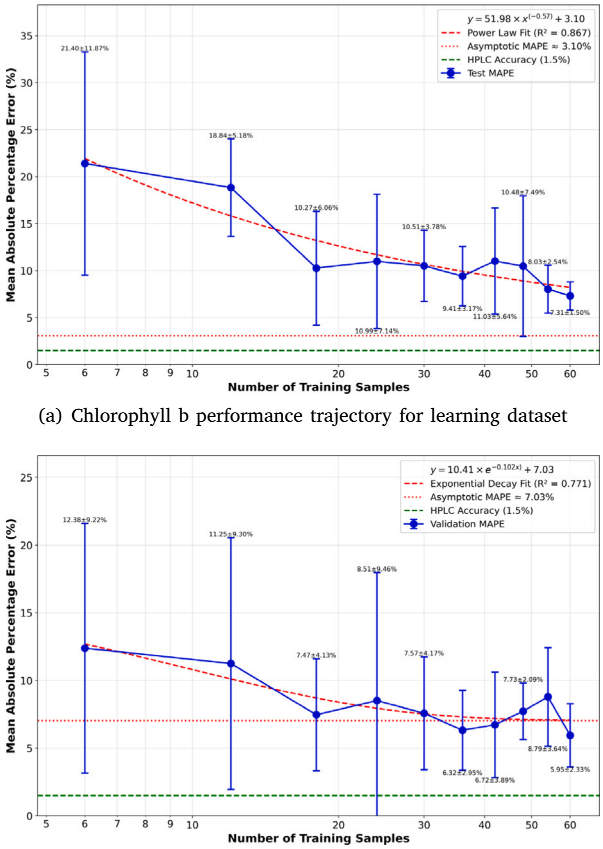
Collecting, preparing, and analyzing microalgae samples for pigment content analysis requires extensive labor and resources. The determination of minimum sample requirements enables researchers to optimize their workflow resources. Indeed, if one knows that 30 samples give 90% of the performance of 60 samples, one can argue on the cost efficiency of producing extra data. A systematic analysis was conducted for each pigment and each data size, which was repeated 5 times with different random seeds to show the variance in the results acts as cross-validation.

Fig. 8 presents the results for lutein (the carotenoid of primary interest in this case). Each point represents a trial with a different seed. The mean of these points represents the learning curve, as introduced by Viering and Loog (Viering & Loog, 2023). The horizontal green line at 1.5% MAPE represents the HPLC accuracy, which serves as the primary reference point.

Following the guidelines of Viering and Loog, the learning curve was fitted with a power-law. This method allows to derive the general sense of convergence of a particular algorithm as a function of the amount of data provided. Two parameters are important, the asymptotic value, which probes how close deep learning is from the accuracy of the expensive reference method (HPLC), and the decay rate, which indicates the expected gain from adding extra samples. In the case of lutein, the asymptotic value is 8.57% and the decay rate is a power −1.80, meaning that doubling the amount of sample would bring one 3.48 times closer to the asymptote.

Similarly, Fig. 9 (left) presents results for chlorophyll b. Here again, the decline in predictive error is again evident, with a sharper reduction suggesting that chlorophyll b becomes efficient in comparatively smaller data sets, until stabilizing in an asymptotic MAPE of 3.10%, based also on a power-law fit.

Fig. 9 (right) further illustrates this analysis for chlorophyll a, incorporating an exponential decay fit that shows consistent improvement in predictive accuracy as more samples are integrated. This trend establishes a robust asymptotic predictive error boundary at approximately 7.03%.



(a) Chlorophyll b performance trajectory for learning dataset

(b) Chlorophyll a performance trajectory for learning dataset

**Fig. 9.** Relationship between learning dataset size and predictive error. The blue line represents the observed error with variation, the red curve shows the fitted decay, and the dashed lines indicate limits and reference thresholds.

Since violaxanthin datasize analysis was behaving differently than lutein and chlorophyll (Fig. 10 (left)), in a comprehensive comparative approach integrates multiple model fits with evaluation of their respective effectiveness in predicting errors for violaxanthin. In particular, the piecewise linear model outperforms others, reflected in its highest value $R^2$ 0.699, which affirms its suitability for predictive modeling based on data sets, meanwhile, it could not depict a specific asymptotic MAPE with this piecewise linear equation:

$$y = \begin{cases} 1.9178x + 0.7123, & \text{for } x \leq 12.0 \\ -0.1180x + 19.6523, & \text{for } x > 12.0 \end{cases}$$

where 12 is the parametric breakpoint of the piecewise linear equation, where the behavior changes.

On the other hand, the zeaxanthin data size learning was implemented using the best performing hybrid architecture model and parameters described previously with 3 repetitions of each dataset size to ensure robustness of results. Finding that polynomial (degree 3) is best performing fitted learning curve.

$$y = -1.041 \times 10^{-3}x^3 + 1.59 \times 10^{-1}x^2 - 5.15x + 75.54$$

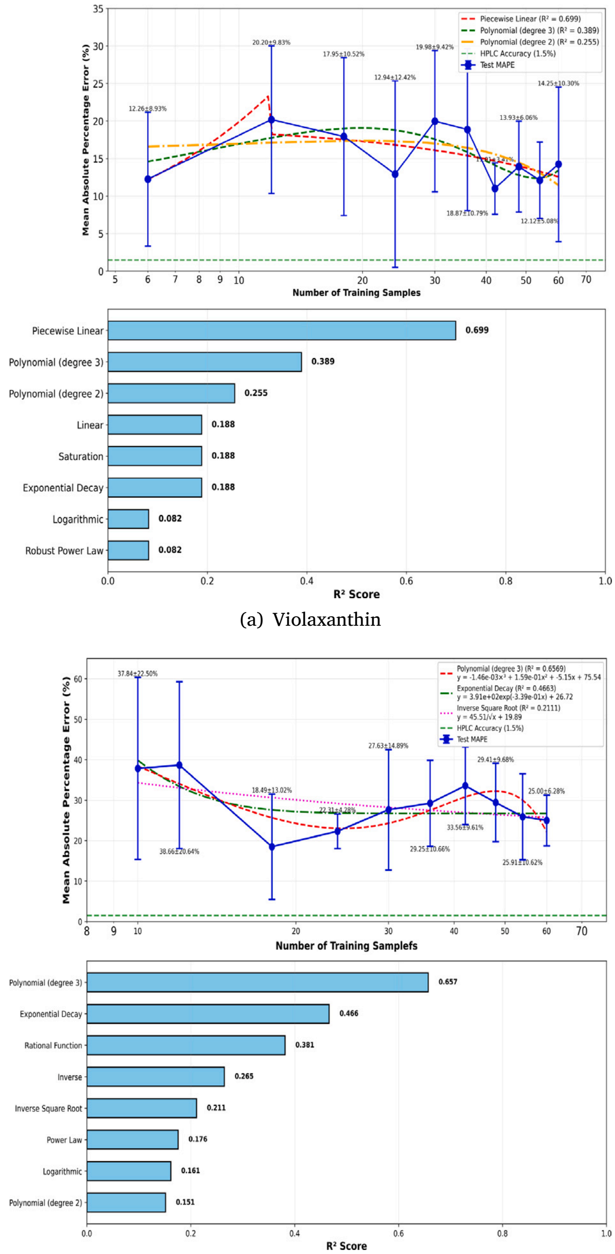
where $y$ represents the MAPE and $x$ denotes the number of training samples. This model achieved a coefficient of determination of $R^2 = 0.657$, indicating a moderately good explanatory power. The lower panel for Fig. 10 (right) shows this finding through a comparative analysis $R^2$, the polynomial (degree 3) fit outperformed all other tested models, followed by exponential decay ($R^2 = 0.466$) and rational function ($R^2 = 0.381$) fits. Together, the figure elucidates how zeaxanthin prediction accuracy does not benefit much from increased data volume, where we can see some sort of a saturation point at 30% of the data samples, where the model does not improve, unlike the case of chlorophylls and lutein, and how higher-order polynomial fitting captures the non-linear dynamics governing model performance.

Taking a step back, the disappointing learning performances on violaxanthin and zeaxanthin can be explained biologically. Indeed, individual carotenoid spectra exhibit significant overlap; therefore, distinguishing them is a challenge. Additionally, these two pigments are present in limited amounts in the cells, particularly compared to lutein. This explains why the CNN can identify lutein easily (the major carotenoid) and shows good learning performance, while it struggles with violaxanthin and zeaxanthin. Still, one should keep in mind that CNN outperforms our previously optimized Partial Least Squares model in every sense.

Finally, Table 4 consolidates these findings, summarizing the minimum and asymptotic MAPE values alongside critical dataset sizes required to achieve specified predictive accuracy thresholds for lutein and chlorophylls. This systematic investigation demonstrates that dataset

**Table 4**
Performance of data size Analysis by chlorophylls and lutein Pigments, results of asymptotic MAPE is based on fitted equations, and samples needed for different MAPE targets.

| Pigment | Fit Eq. | Min MAPE Mean | Asymptotic MAPE | 15.0% | 10.0% | 7.5% | 5.0% |
|---|---|---|---|---|---|---|---|
| Lutein | $y = 308.14 \cdot x^{-1.80} + 8.57$ | 7.16 | 8.57 | 9 | 20 | – | – |
| Chl b | $y = 51.98 \cdot x^{-0.57} + 3.10$ | 7.31 | 3.10 | 14 | 36 | 79 | 346 |
| Chl a | $y = 10.41 \cdot e^{-0.102 \cdot x} + 7.03$ | 5.95 | 7.03 | 6 | 13 | 31 | – |



(a) Violaxanthin

(b) Zeaxanthin

**Fig. 10.** Comparison of performance and model fit quality across carotenoids pigments.

size for training reduces predictive error across various pigments, and that one can get a relatively low prediction error with lower data samples and describes the relationship for that if one wants to build their own network or use the CNNs showed in this work.

## 5. Discussion

While insightful, the obtained results also have to be criticized with a retrospective look at the initial objectives. First, the aim was to retain the individual pigment prediction capability, which is the case. Second, the goal was to achieve predictions of higher quality than conventional machine learning algorithms. To evaluate this point, three machine learning algorithms (Partial Least Squares, Ridge, and Support Vector regressions) were optimized and tested on the data (details available in Supplementary Materials). Table 5 offers a comparison of the performance of each of the models with the proposed CNNs. As one can see, the architectures presented here represent a sizable improvement (from 17 to 63% reduction of MAPE, pigment-wise). Also, the question of the number of samples needed to achieve the desired level of performance was addressed, as it is an important indicator of the labour required to achieve such high-quality prediction. Based on our learning curve analysis, approximately 100–150 samples may be required to approach the asymptotic performance limit for all pigments.

Also, it is interesting to note that in all cases but one, applying no additional preprocessing is the approach yielding the best results. This is true for CNNs as well as machine learning models and calls for a comment. First of all, one has to remember that spectra were derived twice, which is a common technique to manage baseline deviation. Therefore, the addition aiming at managing baseline deviation (MSC, EMSC, and standard scaling) may not be needed *per se*. Second, for all samples, the baseline distortion is minimal. This is a major difference with NIR spectra that systematically feature a baseline slope because of scattering. This can be explained by the nature of the biological samples: extracted molecules, solubilized in the dedicated solvent, and filtered to avoid any particulate matter that could alter the reading. Finally, a comment is to be drawn on why preprocessings lead to underperformance, which is detrimental, as opposed to simply being unnecessary. When the pigments contribute to the overall spectrum, they have an additive contribution (as long as they are diluted enough, which is the case here). Therefore, when spectra are preprocessed, especially in a non-linear fashion, the sense of relative scaling originating from the physical addition can be altered, leading to less informative spectra.

While the proposed architectures overpass conventional machine learning (pure performance), the robustness of their prediction is to be challenged to ensure applicability. To do so, we focused on lutein, the flagship pigment for this study and implemented nested cross-validation scheme (5-fold × 5 seeds = 25 evaluations) with ensemble predictions/training for averaging the 9 best models to provide performance estimates with variability. The retained algorithm achieved 5.59 ± 3.08% MAPE (CV) and 4.60% MAPE (holdout), with $R^2$ = 0.8751 ± 0.2438 (CV) and 0.9844 (holdout) (detailed performance presented in Fig. 11). The relatively high variance in $R^2$ reflects the small test set size in each CV fold (12 samples), where a single outlier prediction can substantially impact the coefficient of determination. The MAPE metric, being more robust to individual large errors, shows more stable estimates. These results are consistent with our learning curve analysis (Fig. 8), which predicted an asymptotic MAPE of 8.57% for single models. The improved performance observed here (5.59%) can be credited to ensemble averaging, which is known to reduce prediction variance. The holdout performance (4.60%) falling within the cross-validation distribution further validates the robustness of our approach.

**Table 5**
Performance comparison on MAPE metric (%) between baseline machine learning algorithms and the CNN presented in this work. Between bracket - the optimal preprocessing method. D.o. - Derivatives only.

| Pigment | PLS best perf. | Ridge best perf. | SVR best perf. | This article's CNN |
|---|---|---|---|---|
| Chlorophyll a | 13.02 ± 2.14 (D.o.) | 11.91 ± 0.00 (D.o.) | 15.78 ± 0.00 (D.o.) | 7.09 |
| Chlorophyll b | 10.04 ± 4.31 (D.o.) | 8.10 ± 1.96 (D.o.) | 8.38 ± 0.00 (D.o.) | 3.00 |
| Lutein | 9.21 ± 0.15 (D.o.) | 9.76 ± 0.00 (D.o.) | 6.54 ± 1.02 (D.o.) | 3.45 |
| Violaxanthin | 15.52 ± 0.90 (D.o.) | 14.13 ± 1.29 (D.o.) | 16.59 ± 0.50 (D.o.) | 9.11 |
| Zeaxanthin | 27.43 ± 3.35 (D.o.) | 26.02 ± 1.94 (D.o.) | 16.56 ± 0.14 (EMSC) | 13.70 (hybrid) |

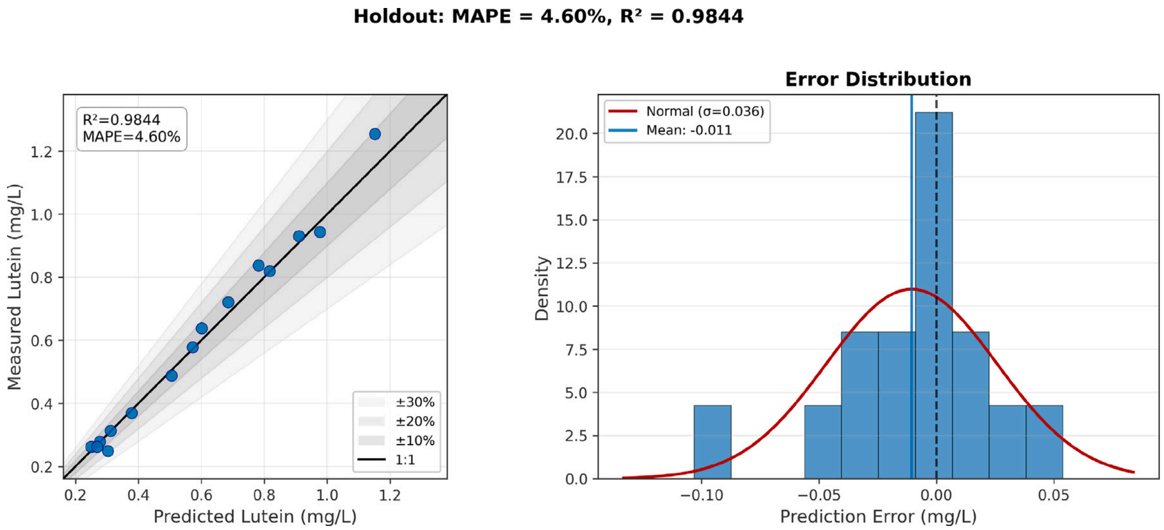


**Fig. 11.** Lutein parity plot and error analysis.

Limitations also have to be acknowledged. The first of them might be the type of microalgae pigments the proposed algorithms can process. Indeed, they were designed and calibrated using green microalgae pigment extracts. Applying them to extracts originating from other microalgae (brown, red, blue–green, . . . ) may require adjustments, or even new developments, as it was the case for zeaxanthin in this work. Another limitation is the extrapolation of the algorithm, which was not tested. For example, the most concentrated samples could have been kept out and used to assess for extrapolation capabilities. Still, one should note that extrapolation would be limited by two aspects: the limit of detection of the spectrophotometer acquiring the spectra, and the limit of the additivity assumption when dealing with a mixture of pigments.

Finally, this work calls for perspectives. Three can be underlined. Firstly, the activation of the different layers of the CNN could be dissected. Indeed, other scholars previously showed that the CNN they used performed both derivation and selection. In the present case, as derivatives are supplied to the CNN, investigating activation could reveal another mathematical operation improving the predictions. To do so, Grad-CAM could be applied to the trained CNNs to visualize precisely which spectral wavelengths the model focuses on. Indeed, this system has the potential to link deep learning models, which operate as black boxes, to biological interpretation. Secondly, compressing these optimized CNN models through quantization to run on portable spectrometers, enabling real-time, in-field pigment quantification for precision farming, would be an interesting line of implementation. Finally, multi-task learning (Zhang & Yang, 2022) by predicting all five pigments concurrently at once through a shared CNN backbone, which produces pigment-specific output, may help detect inter-pigment correlations and enhance predictions of minor carotenoids that share spectral characteristics with dominant pigments.

## 6. Conclusion

In the present work, a deep learning framework was implemented for pigments quantification. Using a double-convolution neural network architecture, as well as a single CNN, successfully delivered improved predictions compared to the previously optimized Partial Least Squares. Extensive workflow and hyperparameter studies were performed to ensure high-accuracy quantification results. The results show that data augmentation by derivation alone (without extra preprocessing) yields the best results. CNN techniques allowed to break a barrier and bring zeaxanthin MAPE below 20%. Finally, the correlation between training dataset size and performance was investigated. Using the newly introduced learning curve tool, it was possible to evaluate the best achievable performance and convergence rate for the major pigments.

The code and the database are available on GitHub, and the trained CNNs are available on HuggingFace. (to be adjusted upon article acceptance)

## Nomenclature

| Latin Symbols | Meaning |
|---|---|
| a | Additive spectrum in MSC |
| A | Additive spectrum in EMSC |
| b | Multiplicative coefficient in MSC and EMSC |
| e | Error (vector or individual, based on indices) |
| p | Polynomial baseline vectors in EMSC |
| $R^2$ | Coefficient of determination |
| s | Spectrum in EMSC |
| x | Input feature (vector or individual, based on indices) |
| y | Output feature (vector or individual, based on indices) |

| Greek Symbols | Meaning |
|---|---|
| $\mu$ | Mean |
| $\sigma$ | Standard deviation |
| ref | Reference in MSC and EMSC |
| scaled | Standard scaling |

| Indices | Meaning |
|---|---|
| Lower case i | Feature index in a vector |
| Lower case j | Run index in the sample batch |

| Acronym | Meaning |
|---|---|
| CNN | Convolutional Neural Network |
| CV | Cross Validation |
| ELU | Exponential Linear Unit |
| EMSC | Extended Multiplicative Scattering Correction |
| GPU | Graphics Processing Unit |
| HPLC | High Pressure Liquid Chromatography |
| ILS | Inverse Least Squares |
| IR | Infra Red |
| MAPE | Mean Absolute Percent Error |
| MLR | Multi Linear Regression |
| MSC | Multiplicative Scattering Correction |
| MSE | Mean Squared Error |
| NA | Not Available |
| NIR | Near Infra Red |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| PLS | Partial Least Squares |
| PRESS | Predicted Residual Error Sum of Squares |
| PSO | Particle Swarm Optimization |
| ReLU | Rectified Linear Unit |
| RMSE | Root Mean Square Error |
| SVM | Support Vector Machine |
| UV | Ultra Violet |

## CRediT authorship contribution statement

**Omar Bayomie:** Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Victor Pozzobon:** Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

This project was funded by the French government as part of the France 2030 initiative. Communauté urbaine du Grand Reims, Département de la Marne, Région Grand Est and European Union (FEDER Grand Est 2021–2027) are acknowledged for their financial support to the Chair of Biotechnology of CentraleSupélec and the Centre Européen de Biotechnologie et de Bioéconomie (CEBB).

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mlwa.2026.100879.

## Data availability

Algorithms can be found on HuggingFace huggingface.co/PozzobonV/CNNspectroMicroalgaehttps://huggingface.co/Pozzob

## References

Achir, N., Servent, A., Soto, M., & Dhuique-Mayer, C. (2022). Feasibility of individual carotenoid quantification in mixtures using UV-vis spectrophotometry with multivariate curve resolution alternating least squares (MCR-ALS). *Journal of Spectroscopy*, *2022*, Article e4509523, Publisher: Hindawi.

Andersen, R. A., & Phycological Society of America (2005). *Algal culturing techniques*. Academic Press, Google-Books-ID: 9NADUHyFZaEC.

Arvayo-Enríquez, H., Mondaca-Fernández, I., Gortárez-Moroyoqui, P., López-Cervantes, J., & Rodríguez-Ramírez, R. (2013). Carotenoids extraction and quantification: a review. *Analytical Methods*, *5*(12), 2916–2924, Publisher: Royal Society of Chemistry.

Barros, A., Pereira, H., Campos, J., Marques, A., Varela, J., & Silva, J. (2019). Heterotrophy as a tool to overcome the long and costly autotrophic scale-up process for large scale production of microalgae. *Scientific Reports*, *9*(1), 13935, Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Applied microbiology;Industrial microbiology Subject_term_id: applied-microbiology;industrial-microbiology.

Bennett, A., & Bogorad, L. (1973). Complementary chromatic adaptation in a filamentous blue-green alga. *Journal of Cell Biology*, *58*(2), 419–435.

Bjerrum, E. J., Glahder, M., & Skov, T. (2017). Data augmentation of spectral data for convolutional neural network (CNN) based deep chemometrics. URL http://arxiv.org/abs/1710.01927, arXiv:1710.01927 [cs].

Bornatico, R., Pfeiffer, M., Witzig, A., & Guzzella, L. (2012). Optimal sizing of a solar thermal building installation using particle swarm optimization. *Energy*, *41*(1), 31–37.

Bosma, T. L., Dole, J. M., & Maness, N. O. (2003). Optimizing marigold (tagetes erecta l.) petal and pigment yield. *Crop Science*, *43*(6), 2118–2124, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2003.2118.

Buscemi, S., Corleo, D., Di Pace, F., Petroni, M. L., Satriano, A., & Marchesini, G. (2018). The effect of lutein on eye and extra-eye health. *Nutrients*, *10*(9), 1321, Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

Camarena-Bernard, C., & Pozzobon, V. (2024). Evolving perspectives on lutein production from microalgae - a focus on productivity and heterotrophic culture. *Biotechnology Advances*, Article 108375.

Cerdà, V., Phansi, P., & Ferreira, S. (2022). From mono- to multicomponent methods in UV-VIS spectrophotometric and fluorimetric quantitative analysis – A review. *TRAC Trends in Analytical Chemistry*, *157*, Article 116772.

Cui, C., & Fearn, T. (2018). Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems*, *182*, 9–20.

Dinç, E., Baleanu, D., & Onur, F. (2001). 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.

Fan, J., Huang, J., Li, Y., Han, F., Wang, J., Li, X., Wang, W., & Li, S. (2012). Sequential heterotrophy–dilution–photoinduction cultivation for efficient microalgal biomass and lipid production. *Bioresource Technology*, *112*, 206–211.

Flórez-Miranda, L., Cañizares-Villanueva, R. O., Melchy-Antonio, O., Martínez-Jerónimo, F., & Flores-Ortíz, C. M. (2017). Two stage heterotrophy/photoinduction culture of scenedesmus incrassatulus: potential for lutein production. *Journal of Biotechnology*, *262*, 67–74.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1–17.

Geladi, P., MacDougall, D., & Martens, H. (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, *39*(3), 491–500.

Grimme, L. H., & Porra, R. J. (1974). The regreening of nitrogen-deficient Chlorella fusca. *Archives of Microbiology*, *99*(1), 173–179.

Higuera-Ciapara, I., Félix-Valenzuela, L., & Goycoolea, F. M. (2006). Astaxanthin: A review of its chemistry and applications. *Critical Reviews in Food Science and Nutrition*, *46*(2), 185–196. http://dx.doi.org/10.1080/10408690590957188, Publisher: Taylor & Francis.

Kona, R., Pallerla, P., Addipilli, R., Sripadi, P., & Venkata Mohan, S. (2021). Lutein and *beta*-carotene biosynthesis in scenedesmus sp. SVMIICT1 through differential light intensities. *Bioresource Technology*, *341*, Article 125814.

Lababpour, A., & Lee, C.-G. (2006). 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.

Lichtenthaler, H. K., & Buschmann, C. (2001). Chlorophylls and carotenoids: Measurement and characterization by UV-VIS spectroscopy. *Current Protocols in Food Analytical Chemistry*, *1*(1), F4.3.1–F4.3.8, _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142913.faf0403s01.

Lichtenthaler, H. K., & Wellburn, A. R. (1983). Determinations of total carotenoids and chlorophylls a and b of leaf extracts in different solvents. *Biochemical Society Transactions*, *11*(5), 591–592.

Maia, L., Casal, S., & Oliveira, M. B. P. P. (2008). Validation of a micromethod for quantification of lutein and *beta*-carotene in olive oil. *Journal of Liquid Chromatography & Related Technologies*, *31*(5), 733–742. http://dx.doi.org/10.1080/10826070701854139, Publisher: Taylor & Francis.

Marini, F., & Walczak, B. (2015). Particle swarm optimization (PSO). A tutorial. *Chemometrics and Intelligent Laboratory Systems*, *149*, 153–165.

Martens, H., & Stark, E. (1991). Extended multiplicative signal correction and spectral interference subtraction: New preprocessing methods for near infrared spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, *9*(8), 625–635, Invited Papers from the International Symposium organized by the Swedish Academy of Pharmaceutical Sciences.

Min, J.-y., & Min, K.-b. (2013). Serum lycopene, lutein and zeaxanthin, and the risk of alzheimer's disease mortality in older adults. *Dementia and Geriatric Cognitive Disorders*, *37*(3–4), 246–256.

Nazari, L., Komaki, S., Salehi, I., Raoufi, S., Golipoor, Z., Kourosh-Arami, M., & Komaki, A. (2022). Investigation of the protective effects of lutein on memory and learning using behavioral methods in a male rat model of alzheimer's disease. *Journal of Functional Foods*, *99*, Article 105319.

Ni, Y., & Gong, X. (1997). Simultaneous spectrophotometric determination of mixtures of food colorants. *Analytica Chimica Acta*, *354*(1), 163–171.

Ochoa Becerra, M., Mojica Contreras, L., Hsieh Lo, M., Mateos Díaz, J., & Castillo Herrera, G. (2020). Lutein as a functional food ingredient: Stability and bioavailability. *Journal of Functional Foods*, *66*, Article 103771.

O'Neill, M. E., Carroll, Y., Corridan, B., Olmedilla, B., Granado, F., Blanco, I., Berg, H. V. d., Hininger, I., Rousell, A.-M., Chopra, M., Southon, S., & Thurnham, D. I. (2001). A European carotenoid database to assess carotenoid intakes and its use in a five-country comparative study. *British Journal of Nutrition*, *85*(4), 499–507.

Porra, R. J. (1990). 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.

Pozzobon, V., & Camarena-Bernard, C. (2022). Lutein, violaxanthin, and zeaxanthin spectrophotometric quantification: A machine learning approach. *Journal of Applied Phycology*.

Pozzobon, V., Levasseur, W., Guerin, C., Gaveau-Vaillant, N., Pointcheval, M., & Perré, P. (2020). Desmodesmus sp. pigment and FAME profiles under different illuminations and nitrogen status. *Bioresource Technology Reports*, *10*, Article 100409.

Pyliotis, N. A., Goodchild, D. J., & Grimme, L. H. (1975). The regreening of nitrogen-deficient Chlorella fusca. *Archives of Microbiology*, *103*(1), 259–270.

Ronzon, T., Tamosiunas, S., & M'Barek, R. (2022). *Jobs and growth in the bioeconomy: Technical report JRC128361*, European Commission - Joint Research Centre, ISBN: 9276471308.

Schisterman, E. F., Vexler, A., Whitcomb, B. W., & Liu, A. (2006). The limitations due to exposure detection limits for regression models. *American Journal of Epidemiology*, *163*(4), 374–383.

Shi, X.-M., Liu, H.-J., Zhang, X.-W., & Chen, F. (1999). Production of biomass and lutein by Chlorella protothecoides at various glucose concentrations in heterotrophic cultures. *Process Biochemistry*, *34*(4), 341–347.

Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å., & Martens, H. (1983). A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Analytica Chimica Acta*, *150*, 61–70.

Stringham, J. M., Johnson, E. J., & Hammond, B. R. (2019). Lutein across the lifespan: From childhood cognitive performance to the aging eye and brain. *Current Developments in Nutrition*, *3*(7), Publisher: Oxford Academic.

Viering, T., & Loog, M. (2023). The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7799–7819.

Wang, Y., Li, M., Ji, R., Wang, M., Zhang, Y., & Zheng, L. (2022). Mark-spectra: A convolutional neural network for quantitative spectral analysis overcoming spatial relationships. *Computers and Electronics in Agriculture*, *192*, Article 106624.

Wellburn, A. R. (1994). 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.

Wiltshire, K. H., Boersma, M., Möller, A., & Buhtz, H. (2000). Extraction of pigments and fatty acids from the green alga Scenedesmus obliquus (Chlorophyceae). *Aquatic Ecology*, *34*(2), 119–126.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130.

Yan, C. (2025). A review on spectral data preprocessing techniques for machine learning and quantitative analysis. *IScience*, *28*(7), Article 112759.

Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609.