Bioresource Technology 418 (2025) 131988

Contents lists available at ScienceDirect

Bioresource Technology

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

Integration of spectroscopic techniques and machine learning for optimizing *Phaeodactylum tricornutum* cell and fucoxanthin productivity

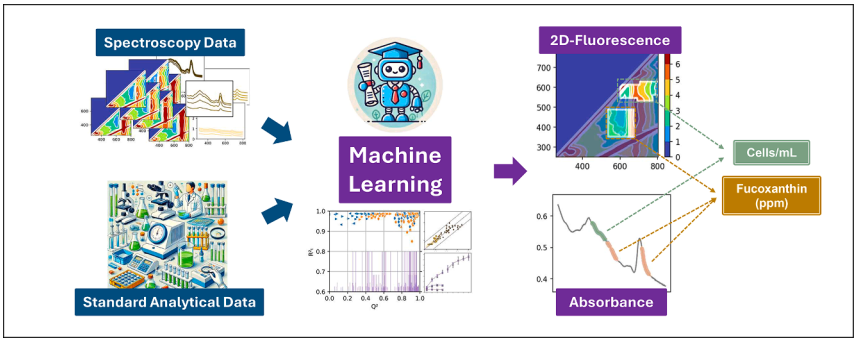
Pedro Reynolds-Brandão [a,b,c], Francisco Quintas-Nunes [b,c], Constança D.F. Bertrand [b,c], Rodrigo M. Martins [b,c], Maria T.B. Crespo [b,c], Cláudia F. Galinha [a,*], Francisco X. Nascimento [b,c]

[a] LAQV-REQUIMTE, Chemistry Dept., NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
[b] iBET, Instituto de Biologia Experimental e Tecnológica, 2781-901 Oeiras, Portugal
[c] Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, 2780-157, Oeiras, Portugal

HIGHLIGHTS

- Media formulations impacted *Phaeodactylum tricornutum* productivity.
- Absorbance and 2D-fluorescence captured biological variance.
- A comprehensive dataset was generated and machine learning models applied.
- Spectroscopy-based models for monitoring *P. tricornutum* cultures were developed.
- Robust and sensitive monitoring of *P. tricornutum* fucoxanthin and cell productivity.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Microalgae
Culture
Monitoring
Carotenoids
2D-fluorescence
Absorbance

ABSTRACT

The development of sustainable and controlled microalgae bioprocesses relies on robust and rapid monitoring tools that facilitate continuous process optimization, ensuring high productivity and minimizing response times.

In this work, we analyse the influence of medium formulation on the growth and productivity of axenic *Phaeodactylum tricornutum* cultures and use the resulting data to develop machine learning (ML) models based on spectroscopy. Our culture assays produced a comprehensive dataset of 255 observations, enabling us to train 55 (24+31) robust models that predict cells or fucoxanthin directly from either absorbance or 2D-fluorescence spectroscopy.

We demonstrate that medium formulation significantly affects cell and fucoxanthin concentrations, and that these effects can be effectively monitored using the developed models, free of overfitting. On a separate data subset, the models demonstrated high accuracy (cell: $R^2 = 0.98$, RMSEP = 2.41×10^6 cells/mL; fucoxanthin: $R^2 = 0.91$ and RMSEP = 0.65 ppm), providing a practical, cost-effective, and environmentally friendly alternative to standard analytical methods.

* Corresponding author.
E-mail address: cf.galinha@fct.unl.pt (C.F. Galinha).

https://doi.org/10.1016/j.biortech.2024.131988
Received 1 August 2024; Received in revised form 13 December 2024; Accepted 14 December 2024
Available online 16 December 2024

1. Introduction

Algae carotenoid pigments have garnered significant attention due to their therapeutic properties and health benefits (Pradhan et al., 2023; Rowles and Erdman, 2020; Vílchez et al., 2011). Particularly, fucoxanthin, a xantophyll carotenoid pigment, has demonstrated powerful antioxidant and anti-inflammatory properties, which can be beneficial for cancer therapy, cardiovascular health, and obesity management (Butler et al., 2020; Leong et al., 2022; Neumann et al., 2019). One of the most promising sources of fucoxanthin is the marine diatom, *Phaeodactylum tricornutum* (Zhuang et al., 2023). With efficient photosynthetic activity and fast growth rates, *P. tricornutum* cells can produce up to 100 times the dry weight fucoxanthin content of brown seaweeds, which have traditionally been the commercial source for this pigment (Butler et al., 2020; Leong et al., 2022).

When producing fucoxanthin from *P. tricornutum* cultivations, various factors ranging from temperature and light intensity to nutrient levels must be optimized so that the desired productivity is achieved (McClure et al., 2018; Wang et al., 2022b; Yang and Wei, 2020; Yi et al., 2019; Ashokkumar et al., 2023) For this purpose, it is essential to attentively monitor the culture's productivity in response to these factors, with a particular focus cell concentration and fucoxanthin content. High-Performance Liquid Chromatography (HPLC) and Flow Cytometry are the standard analytical methods used to accurately quantify pigments and cells in microalgae bioprocesses. However, these methods are time-consuming, do not provide real-time information, and thus limit the effectiveness of the monitoring task. Additionally, these methods are expensive and unsustainable, requiring highly trained personnel, costly equipment, and polluting consumables. This not only undermines the intended sustainability of microalgae-based bioprocesses but also limits the global potential of microalgae cultivation, as process optimization becomes restricted to high-income countries.

Spectroscopy and optical probes are tools widely proposed to facilitate monitoring of bioprocesses (Busse et al., 2017; Faassen & Hitzmann, 2015; Liu et al., 2021). These tools can be applied directly inline, providing real-time information. Moreover, when compared to standard analytical methods, spectroscopy is operationally inexpensive and environmentally green. Specifically, absorbance spectroscopy has been already used for monitoring microalgae bioprocesses (Griffiths et al., 2011; Lichtenthaler and Buschmann, 2001), including fucoxanthin production by *P. tricornutum* (Li et al., 2018). However, these methods are not completely reliable due to their susceptibility to noise and non-linear interferences caused by overlapping spectra (Chen et al., 2017; Lakowicz, 2006). Furthermore, some methods still require labour-intensive sample preprocessing techniques, such as solvent-mediated pigment extraction, which impairs real-time monitoring.

To address these limitations, machine learning (ML) models have shown promise for monitoring bioprocesses using spectroscopy and optical probes. In the context of spectroscopy, ML models can overcome signal noise and non-linearity, surpassing the capabilities of linear regression models for leveraging optical probes and spectroscopy (Biancolillo and Marini, 2018; Faassen and Hitzmann, 2015). Several bioprocess applications, such as wastewater treatment and the cultivation of yeast, bacteria, and animal cells, already demonstrate the potential of ML models for spectroscopic monitoring (Bayer et al., 2020; Galinha et al., 2012; Grote et al., 2014; Teixeira et al., 2011). For monitoring microalgae cultivations, a few studies showed that specific modelling techniques make fluorescence spectroscopy a robust monitoring tool (Brandão et al., 2023; Havlik et al., 2022; Sá et al., 2022; Porras Reyes et al., 2024).

In the present work, we build a comprehensive dataset from the thorough monitoring of *P. tricornutum* cultivation assays for studying medium formulation effects. For that purpose, standard analytical methods were employed in parallel with absorbance or 2D-fluorescence spectroscopy measurements. The dataset was then used to train sensitive and robust machine learning models that effectively predict cell growth and productivity directly from either of the spectroscopies. It provides numerous observations on diverse profiles of cell growth, fucoxanthin productivity, and corresponding spectroscopies, all reflecting the microalga's response to varying nutrient concentrations, specifically Nitrogen (N) and Phosphorus (P) sources. This facilitated the application of a sophisticated machine learning methodology, designed to develop models capable of predicting cell growth and productivity while remaining robust to the variability observed in assay profiles.

2. Material and methods

2.1. Phaeodactylum tricornutum cultivation experiments for the development and validation of Machine learning models

2.1.1. Preparation of axenic inoculum

Phaeodactylum tricornutum CCAP 1055/1 was acquired from the Culture Collection of Algae and Protozoa (CCAP, Scotland, United Kingdom). The axenic inoculum was obtained via serial dilution and antibiotic treatment in F/2 media and incubated at 22 °C under 250 μmol/sec/m^2 white LED light in a 16:8-hour light–dark cycle for 12 days. To confirm the absence of associated bacteria, single axenic *P. tricornutum* colonies were selected and grown in Marine Agar 2216 (MA), as well as resuspended in liquid to be observed by microscopy. Afterwards, the inoculum of *P. tricornutum* was scaled up in sterile 2 L Schott flasks containing 1.2 L of F/2 medium prepared using natural seawater collected in the Oeiras region, Portugal. The inoculum cultures were performed at 22 °C with aeration at 0.2 L/min with filtered compressed air (Millex-FG 50 mm 0.2 μm PTFE) and dispersed through a 33cD DURAN® Gas distribution tube and conducted under sterile conditions within a 16:8-hour photoperiod. Samples were taken and checked for contamination daily by spreading 100 μL of 10 mL suspensions on Marine Agar, with no bacterial or fungal colonies observed after 12 days at 26 °C. The inoculum cultures reached the stationary phase (~1 x 10^7 cells/mL) after 12 days and were immediately used for the following experiments.

2.1.2. Cultivation assays varying abundance of N and P sources

To obtain data for developing robust models, three independent axenic *P. tricornutum* cultivation assays were conducted using medium prepared with seawater with different nutrient concentrations (F/2, F/2 + N, F/2 + N + P): a simple F/2 medium, a F/2 medium supplemented with 1 g/L of $NaNO_3$, and a F/2 medium supplemented with 0.5 g/L of $NaNO_3$ and 0.05 g/L of Na_2HPO_4. The assays were conducted under the same vessels, volume, temperature, light, and aeration conditions as previously described in 2.1.1. A total of fifteen culture replicates per assay were performed. Each replicate started with an initial cell concentration of 1x10^6 cells/mL. Observations from the F/2 and F/2 + N assays were collected at four time points (3, 5, 7, and 9 days after inoculation; until no cell growth was observed). F/2 + N + P assay was sampled at nine time points (days 3, 5, 7,9, 12, 14, 16, 18 and 20), as significant growth was observed up to 20 days after inoculation in F/2 + N + P experiment (see results section 3.1). As the F/2 + N + P assay was longer, it was divided into two parts: F/2 + N + P part 1 (days 3 to 9) and part 2 (days 12 to 20). All samples were analysed by both 2D-fluorescence and absorbance spectroscopy, as well as by standard analytical methods (as described in section 2.2.4).

2.2. Analysis of Phaeodactylum tricornutum samples

2.2.1. Cytometry analysis for quantification of cell count and single-cell fluorescence

Samples from the *P. tricornutum* cultivation experiments were retrieved and directly analysed by flow cytometry using a Guava® Muse® Cell Analyzer (Luminex, USA). Parameters such as cell count (cells/mL), cell size (forward scatter, FSC) and red auto-fluorescence per cell (RED) were obtained. The latter parameter is directly obtained from

the Guava® Muse® System as its operating laser, a Class IIIb 532-nm laser wavelength in CW mode, induces fluorescence of the intracellular chlorophyll molecules.

2.2.2. High-Performance liquid chromatography (HPLC) analysis of methanol extracts for quantification of fucoxanthin (ppm)

The fucoxanthin concentration in culture samples was quantified by preparing clear methanol extracts and analysing them using high-performance liquid chromatography (HPLC). Culture samples were centrifuged (5000 G, 5 min) and the resulting pellets were resuspended in 5 mL of pure methanol and left in the dark for 24 h. The volume of sample centrifuged was adjusted so no more than $5x10^7$ cells were pelleted. Afterwards, the methanol biomass suspensions were again centrifuged and the extracts (i.e., the supernatants) were collected and filtered (Millex-FG 20 mm 1 μm PTFE); the effectiveness of the extraction was confirmed by checking that the pellets were completely white. HPLC analysis was then conducted using a Waters Alliance Separations Module e2695 (Waters, Dublin, Ireland) coupled with Photodiode Array Detector Module e2998 (HPLC-PDA). Separation of fucoxanthin was achieved using a C18 reverse phase column (Phenomenex Luna 3u C18 (2) 100A 75*4.60 mm) at a constant flow rate of 1 mL/min and a gradient elution with the following profile: 65 % acetonitrile (ACN) and 35 % milli-Q water (MQ) from 0 to 8 min, increasing until 90 % ACN and 10 % MQ from 8 to 11 min and maintained until 14 min and then decreasing to 65 % ACN and 35 % MQ from 14 to 20 min. The temperature of the column oven was 40 °C and the sample injection volume was 20 μL. The chromatogram was recorded using the PDA module at the absorbance maximum wavelength for fucoxanthin. To quantify, fucoxanthin standards (Sigma-Aldrich, fucoxanthin analytical standard, MFCD01745140) were prepared in pure methanol at concentrations ranging from 0.05–0.6 ppm and analysed by HPLC with the same methodology. This allowed the creation of a calibration curve relating chromatogram peak integration to fucoxanthin concentration could be obtained. The fucoxanthin peak within the chromatogram was detected at a residence time of 7 min.

2.2.3. Quantification of nitrate and phosphate in culture supernatants

Nitrate (NO_3) and phosphate (PO_4) contents in culture samples were quantified from their aqueous supernatants, obtained from the centrifugation process described in the previous section. NO3 quantification was performed using the spectrophotometric method described by (Wang et al., 2022a) , while PO_4 quantification followed the methodology outlined by (Ducklow and Dickson, 1994). Both methods are specifically designed for seawater analysis.

2.2.4. Absorbance and 2D-fluorescence spectroscopy analysis

Culture samples were analysed directly by absorbance spectroscopy and 2D-fluorescence spectroscopy without any pre-treatment. Absorbance measurements were performed using an Ultrospec 2100 Pro UV/Vis spectrophotometer and 1 mL plastic cuvettes (10 mm optical path), producing an absorbance spectrum between 300 and 800 nm per sample, with a scan speed of 20 nm per second. Additionally, methanol extracts prepared in Section 2.2.2 were analysed by absorbance spectroscopy at a single wavelength of 445 nm. 2D-fluorescence measurements were performed using a Varian Cary Eclipse spectrofluorometer and 400 μL quartz cuvettes, generating an excitation-emission matrix (EEM) for each sample. Excitation wavelengths ranged from 250 to 790 nm, and emission was detected between 260 and 800 nm, both in 5 nm steps. The monochromator slit widths were set to 20 nm for both excitation and emission. The scan speed was 200 nm of emission reads per second per excitation wavelength, resulting in an analysis duration of approximately 10 min per sample.

2.3. Statistical analysis of Phaeodactylum tricornutum cultivation assays

2.3.1. Significance and normality tests

The significance of differences between assays and/or time points regarding culture fucoxanthin and cell contents was confirmed using a statistical test pipeline: 1) Normality test – Shapiro-Wilk test was applied to each data point in each assay; 2) Homoscedasticity test – Performed only if the normality test indicated $p > 0.05$ for all data points; 3) Significance and post hoc tests – Applied based on the outcomes of the previous steps.

If data was found unlikely to be normally distributed ($p < 0.05$), Kruskal-Wallis test to identify significance, followed by Dunn's post-hoc test for pairwise comparisons. If all showed Shapiro-Wilk p-value $>$ 0.05, a Levene test was used to verify homoscedasticity. In cases where homoscedasticity was unlikely (Levene's test p-value < 0.05), Welch's ANOVA test was applied, followed by Tukey's post-hoc test. Otherwise, a standard ANOVA was used, also followed by Tukey's post-hoc. The analysis was conducted using a custom script developed in-house, utilizing functions from scipy.stats library. The script is available on the developer platform GitHub within PT-FucoFromSpec repository (https://github.com/ibetbio/PT-FucoFromSpec).

2.3.2. Principal component analysis

Principal Component Analysis was performed on the dataset, for data overview and outlier detection. Custom Python scripts were developed in house using Spyder IDE within a Conda environment, utilizing Scikit-Learn and Matplotlib libraries. These scripts are available in PT-FucoFromSpec repository on GitHub (https://github.com/ibetbio/PT-FucoFromSpec). Three PCA were performed: a) standard analytical data analysis, including fucoxanthin (ppm), cell count (Million cells/mL), fucoxanthin per cell (pg/cell), and CytoRed (a.u.); b) 2D-fluorescence spectroscopy data analysis, including the 6103 fluorescence variables (excitation-emission wavelength pairs), and c) absorbance spectroscopy data analysis, including the 500 absorbance variables (wavelengths). Two principal components were selected for each analysis. The resulting scores and loadings were visualized as follows: for standard analytical data, biplots were generated; for the spectroscopic data, loadings were represented within spectral plots as heatmaps.

2.4. Development of machine learning models to predict P. Tricornutum cultivation parameters based on spectroscopy data

2.4.1. Python scripts

To implement the methodology described in the following sections, custom Python scripts were developed in-house using the Spyder IDE within a Conda environment. Libraries were installed for various purposes, including data manipulation (Pandas) and serialization (Pickle), scientific computing (NumPy), machine learning (Scikit-learn, TensorFlow, Keras), and data visualization (Matplotlib, Seaborn). The scripts are available on the GitHub developer platform (https://github.com/ibetbio/PT-FucoFromSpec).

2.4.2. Preparation of the data for machine learning processes

The methodology described in 2.2. produced a data set consisting of 255 observations, all derived from the analysis of *P. tricornutum* culture samples. This dataset is available in the PT-FucoFromSpec repository on GitHub (https://github.com/ibetbio/PT-FucoFromSpec). Each observation includes data from standard analytical methods, specifically, fucoxanthin volumetric concentrations (Fx ppm), cell count (Million cells/mL), cell autofluorescence, CytoRed (a.u.). Additionally, it contains data from spectroscopic analysis: absorbance spectra with 500 absorbance variables, and 2D-fluorescence spectra with 6103 fluorescence variables.

The observations were divided into two data subsets: a training subset, used to develop models; and a testing subset, used to validate

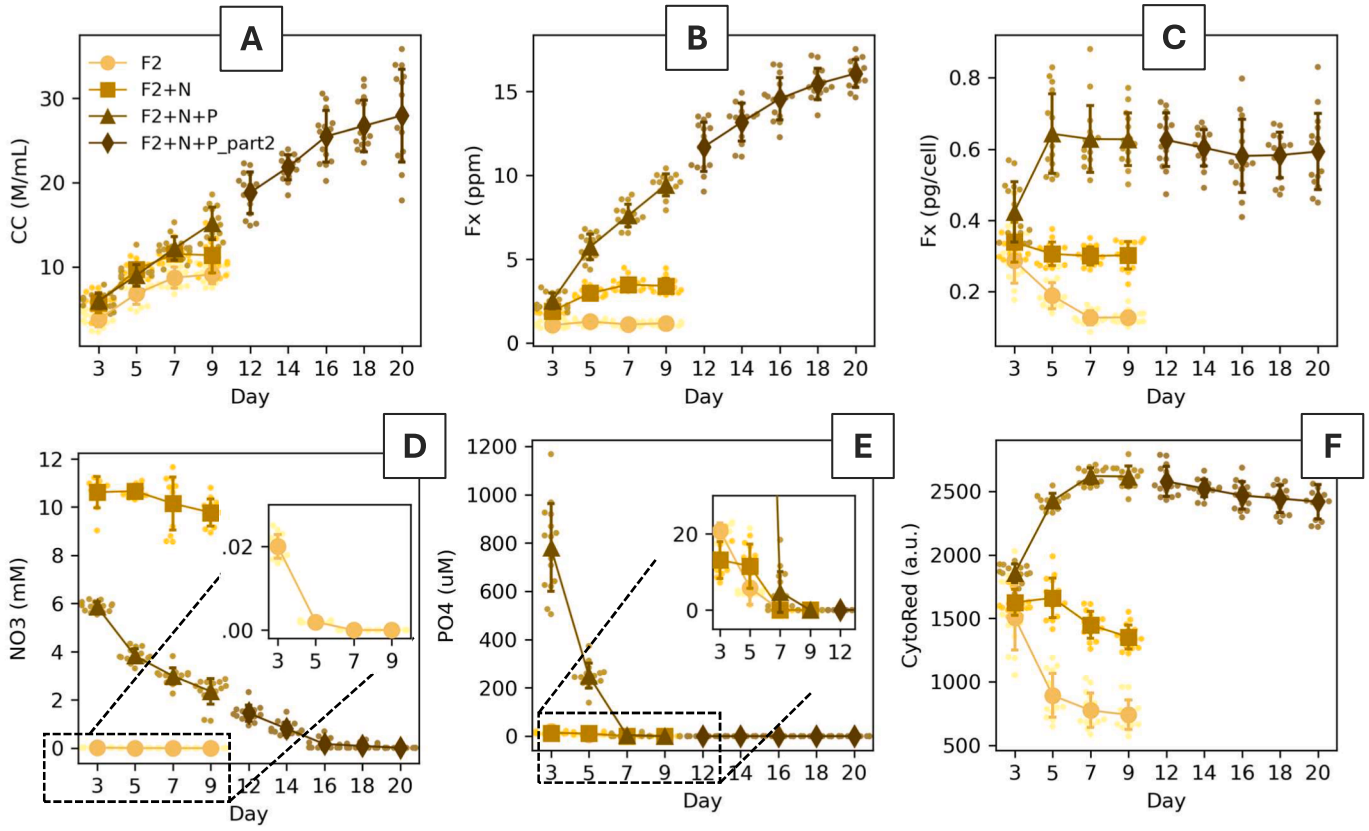


Fig. 1. Growth, productivity, and consumption profiles of *Phaeodactylum tricornutum* axenic cultivation obtained for each medium formulation assay (i.e., F/2, F2/+N, F/2 + N + P). The profiles include cell (CC), fucoxanthin (Fx), NO_3, and PO_4 contents, as well as single-cell auto-fluorescence (CytoRed).

them. The inclusion of data in each subset was deliberately non-random to prevent data leakage and, thus, overfitting (Cawley and Talbot, 2010). Instead, entire culture replicates (batches) were selected: replicates 1–10 were used for the training set, while observations from replicates 11–15 were used for the testing set. To enhance the training data, data augmentation was applied (Zhang et al., 2019; Esben et al., 2017). This accounted for the expected human experimental error inherent in standard analytical methods. Additional details about the augmentation method are provided as comments within the scripts available in the PT-FucoFromSpec repository on GitHub (https://github.com/ibetbio/PT-FucoFromSpec).

2.4.3. Machine learning methodology

A three-step methodology was implemented to develop the machine learning models (see supplementary materials). Step 1 is the definition of machine learning pipelines (described in Section 2.4.3.1), involving the generation of combinations of machine learning tools (i.e., algorithm + feature selection method + pre-process transformation) and defining search spaces for hyperparameter tuning; it does not require data. Step 2 works exclusively with the training subset and involves tuning, training, and selection of models (described in Section 2.4.3.2); and Step 3 works exclusively with the testing subset and involves testing and benchmarking the models developed in step 2 (described in Section 2.4.3.3).

2.4.3.1. Step 1: Definition of machine learning pipelines. To generated machine learning (ML) pipelines, five algorithms were combined with four feature selection methods and three pre-processing transformations. The algorithms included Projection to Latent Structures (PLS), Convolutional Neural Networks (CNN), Random Forest (RF), Support Vector Machines (SVM) and Extreme Gradient Boost (XGBoost). To implement these algorithms, the Scikit-learn and Tensorflow libraries were used. The feature selection methods consisted of variable importance to projection (VIP) and moving window (MW) (Forina et al., 2004). Scikit-learn was used also for VIP, while MW was implemented through a custom script developed in house (https://github.com/ibetbio/PT-FucoFromSpec). The pre-process transformations were mean centering and standard normalization (MCSN), log transformation (Log) applied to all data, and log transformation to the standard analytical data exclusively. Step 1 resulted in 190 distinct ML pipelines, each consisting of a ML algorithm, a feature selection method, and pre-process transformation, and a hyperparameter search space for tuning. The intervals for hyperparameter search were defined based on the degrees of freedom offered by each component (further details available at https://github.com/ibetbio/PT-FucoFromSpec).

2.4.3.2. Step 2: Tuning, training, and selection of models. The model tuning process involved identifying the optimal hyperparameter values for each pipeline within the defined search spaces. The criterion used for this optimization was robustness to leave-one-assay-out cross-validation (LOAOCV). In LOAOCV, models are trained using data from all assays except one, which is held out for testing. This process is repeated until each assay has been excluded and tested once. Robustness was evaluated using the coefficient of determination (Q^2) and root-mean-squared-error (RMSECV). The tuning process concluded once the hyperparameter search space was fully explored. The optimal hyperparameter values were those that resulted in the lowest RMSECV. The training phase involved using the tuned hyperparameters to fit the models to the entire training subset. The resulting training accuracy was evaluated using the coefficient of determination (R^2_{Train}) and root-mean-squared error (RMSET). Animated visual representations of the tuning and training processes for all pipelines in Step 2 are provided as supplementary

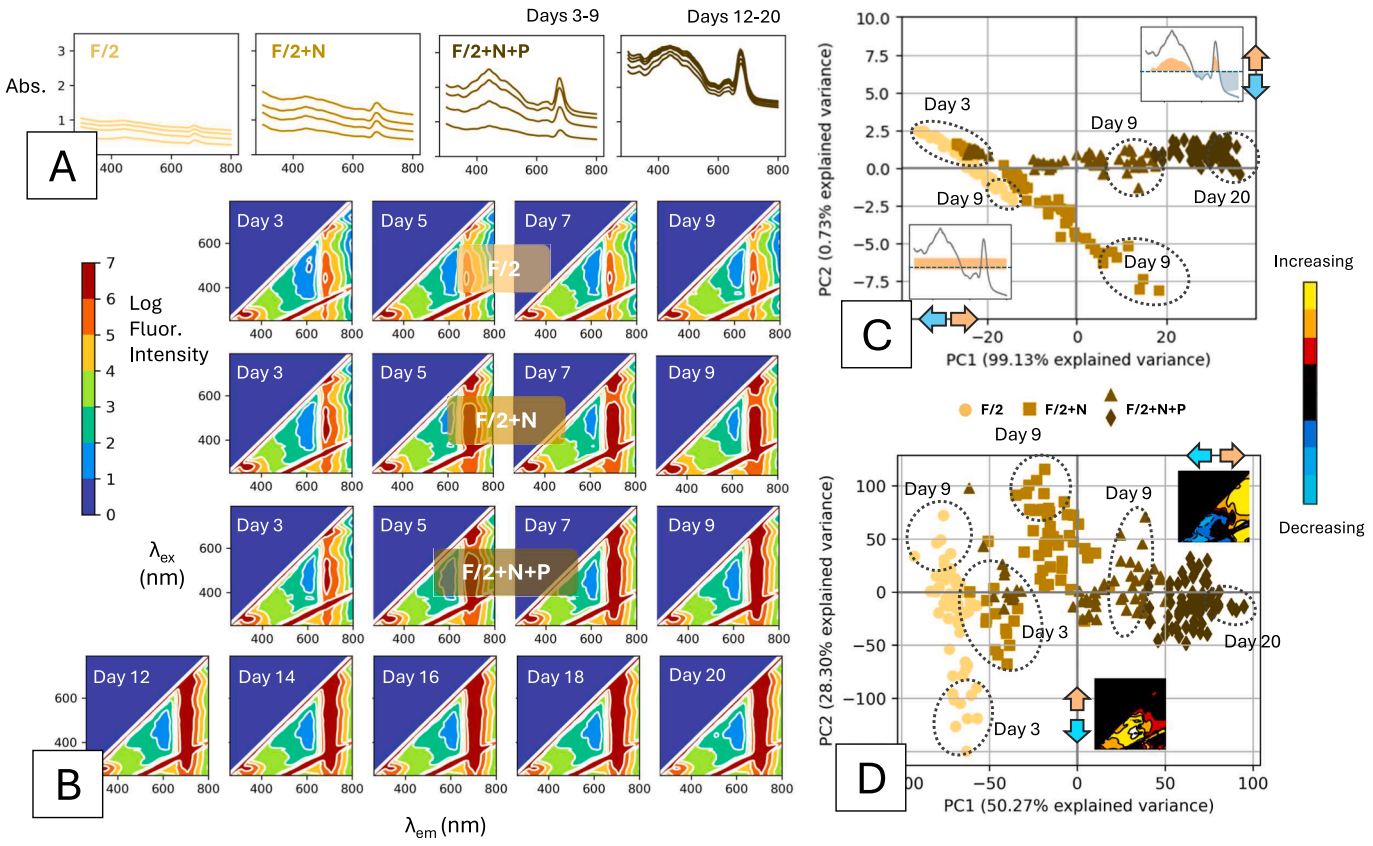


Fig. 2. *Phaeodactylum tricornutum* cultivation profiles for Absorbance and Fluorescence Spectroscopy in F/2 media supplemented with different NO_3 and PO_4 concentrations. On the right, the average spectroscopy profiles, on the left the PCA score plots including all data on each spectroscopy.

videos (video links: GridSearchCVMEx_2DF.gif, GridSearchCVMW_2DF.gif, GridSearchCVMW_abs.gif, GridSearchCVVIP_2DF.gif, GridSearchCVVIP_abs.gif, RandomsearchCV_2DF.gif, RandomSearchCV_abs.gif). The supplementary materials include.gif files illustrating the spectral selection profiles, training accuracy, and cross-validation accuracy for all pipelines. Model selection was then performed based on training and cross-validation performance to eliminate overfitted models. For this purpose, both R^2_{Train} and Q^2 were considered: higher R^2_{Train} indicated better model fit, while higher Q^2 reflected greater architecture robustness. Discrepancies between R^2_{Train} and Q^2 were interpreted as indicating model overfitting. To aid in model selection, linear regression was performed using individual spectral elements within each spectroscopy type. A total of 500 linear regression models were computed from absorbance and 6103 from 2D-fluorescence. All these models were evaluated using LOAOCV, with the highest Q^2 serving as a reference for ML model selection. A ML model with lower Q^2 than that of linear regression was considered not worth the additional computational effort. Animated representations of this process are included as supplementary videos (links: 'GridSearchCV_LinReg_2DF.gif', 'GridSearchCV_LinReg_abs.gif').

2.4.3.3. Step 3: Model testing and benchmarking. The models selected in Step 2 were applied to the testing data subset to make predictions based on the spectroscopy data. As in the previous step, the performance was evaluated using the coefficient of determination (R^2_{Test}) and the root mean squared error of prediction (RMSEP). This step includes also the utilization of the traditional alternatives to standard analytical methods as benchmarks for the ML models. The traditional alternatives used for predicting cell count (CC, Million cells per mL) and fucoxanthin concentrations (Fx, ppm) were optical density of the culture samples at 750 nm (OD_{750}) and absorbance at 445 nm of pigment extracts from culture samples ($extOD_{445}$), respectively. Linear regression models were developed using the training data subset to relate CC and Fx with OD_{750} and $extOD_{445}$, respectively. These simple models were then applied to the testing set, following the same procedure as the ML models. Details on these models are available in the Supplementary Material. The benchmarking process involved not only comparing comparing R^2_{Test} and RMSEP across the different models but also performing statistical analysis. For this purpose, the statistical pipeline described in section 2.3.1 was used to employed to test the following null hypothesis: "*The spectroscopy-based machine learning model provides data that does not differ significantly from the standard analytical method*".

3. Results and discussion

3.1. Phaeodactylum tricornutum cell and fucoxanthin production profiles in F/2 media supplemented with different NO_3 and PO_4 concentrations

The growth and productivity profiles resultant of the *P. tricornutum* cultivation assays are shown in Fig. 1. Cell concentration (CC, million cells per mL), fucoxanthin concentration (Fx, ppm), fucoxanthin per cell (Fx, pg/cell), and photosynthetic activity (CytoRed, a.u.) are plotted through operating days for all cultivations of each assay (Fig. 1 — **A, B, C and F**), where it is possible to observe that each assay resulted in significantly different profiles (all statistical analysis results can be found in Supplementary Materials). Maximal growth and productivity were observed when *P. tricornutum* was cultivated in F/2 medium supplemented with both nitrate (NO_3) and phosphate (PO_4), referred to as F/2 + N + P assay. Under these conditions, cell concentration and fucoxanthin concentration consistently improved over time, with significant increases observed until 20 days after inoculation. The values reached as high as 17 ppm of fucoxanthin and 30 million cells/mL. During this period, the concentrations of NO_3 and PO_4 in the F/2 + N + P medium decreased significantly, as it can be clearly observed in Fig. 1,

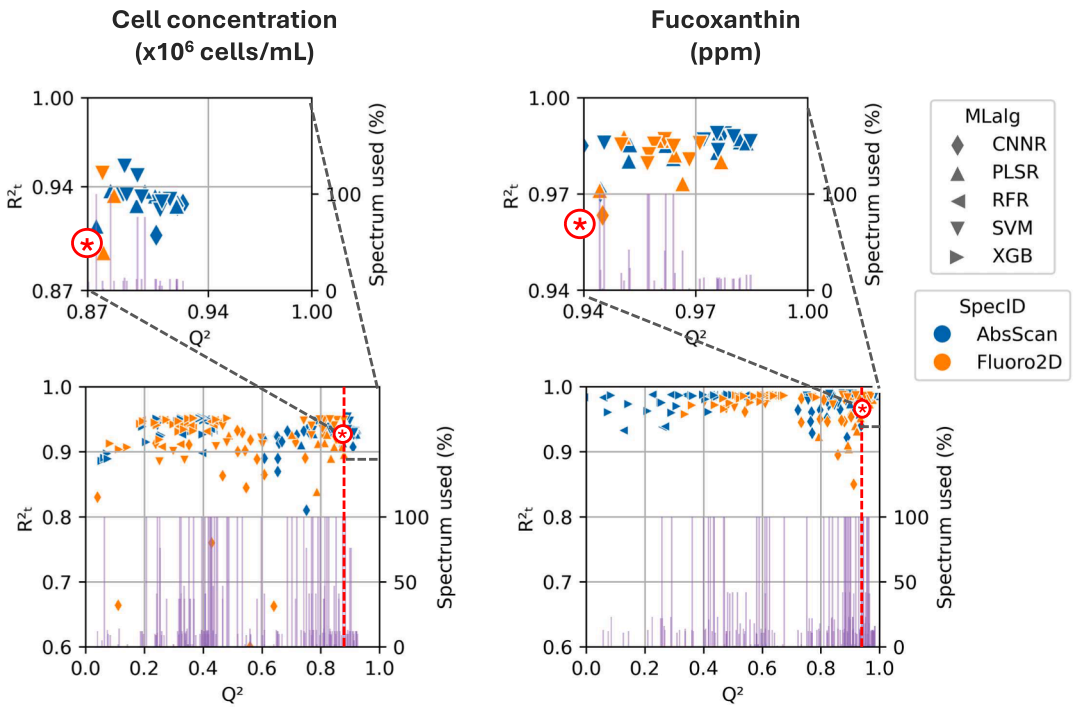


Fig. 3. Machine learning accuracy results for the tuning (Q^2) and training (R^2_{Train}) of 200 models computed for predicting cell count (million cells/mL) or fucoxanthin (ppm) of *P. tricornutum* cultivations; the architectures are based on a % (violet bar) of either absorbance (blue) or 2D-fluorescence (orange), and each uses 1 out of 5 ML algorithms (data point markers); (*) – threshold for acceptable robustness, marked by the most robust linear regression model possibly obtained. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

D and E, with PO_4 fully depleted by day 9. In comparison to other assays, i.e., cultivation in F/2 medium with extra NO_3 (F/2 + N) and in F/2 medium alone (F/2), F/2 + N + P showed significantly higher fucoxanthin levels as early as day 3. This was also observed for photosynthetic activity (measured by cell autofluorescence, CytoRed, a.u.) and fucoxanthin accumulation per cell (Fx, pg/cell), which significantly increased until day 9 and remained stable until day 20 (Fig. 1**, C and F**). Growth limitations and decreased fucoxanthin production and photosynthetic activity were observed when *P. tricornutum* was cultivated in F/2 or F/2 + N medium. These assays showed no significant increments in cell growth and Fx productivity from day 7 forward. In the case of the F/2 + N assay, fucoxanthin content became stationary at ~ 3 ppm and CC at 15 million cells/mL. In the case of the F/2 assay, no significant variation of fucoxanthin (ppm) was verified during the entire cultivation period. F/2 assay remained stagnant at ~ 1.2 ppm of fucoxanthin, and CC increased to just 10 million cells/mL. These limitations were linked to the lower levels of NO_3 or PO_4 in these assays. Indeed, the residual NO_3 and PO_4 concentration in these assays depleted by day 7, coinciding with the cessation of cell growth observed in both assays. Notably, F/2 + N cultures remained at very high NO_3 concentrations until day 9, varying between 11 and 9 mM. Small decreases of NO_3 were only detected with significance between 4-day long time intervals (i.e., between days 3 and 7, and between days 5 and 9). This contrasts with assay F/2 + N + P, where, at day 9, 80 % of the NO_3 was already consumed/uptaken by *P. tricornutum* cells, having decreased significantly between every time point. This suggests that the consumption of NO_3 is dependent on the abundance of PO_4, explaining the limitations in growth and fucoxanthin production observed in the F/2 + N. treatment. Still, *P. tricornutum* cultivations in F/2 + N showed significantly better growth, productivity, and photosynthetic activity than in F/2, and this was verified for all time points. This result suggests that NO_3 supplementation alone to F/2 is beneficial for *P. tricornutum* cultivation, although to a lesser extent than when PO_4 is abundant.

Overall, the results on cell and fucoxanthin production profiles suggest that adequate quantities for N and P sources are vital for optimal *P. tricornutum* autotrophic growth, photosynthetic activities, and fucoxanthin production, which is consistent with results obtained in previous studies (Afonso et al., 2022; Huang et al., 2019; Levitan et al., 2015; McClure et al., 2018).

3.2. Absorbance and 2D-fluorescence spectroscopy: Variation within P. Tricornutum cultivation assays

The absorbance and 2D-fluorescence spectroscopy profiles resultant of the *P. tricornutum* cultivation assays are shown in Fig. 2**, A and B**. As expected, *P. tricornutum* cultivations conducted in F/2 + N + P medium showed higher variance and higher final values for absorbance and fluorescence intensity values when compared to the F/2 + N and F/2 treatments. This difference is particularly evident in the F/2 treatment, which showed an apparently stagnant fluorescence profile, and very low increments of absorbance. Being spectral data highly multivariate (i.e., 500 wavelengths of absorbance, 6103 excitation-emission pairs of fluorescence), the spectral differences between assays are not straightforward. Because of this, a 2-component Principal Component Analysis (PCA) was used, being applied separately to absorbance and 2D-fluorescence (Fig. 2**, C and D)**. Within the plots, the contributions of the spectral elements for each principal component are represented by a heatmap: hotter regions imply positive correlation with the major directions of variance, while cold imply a negative correlation.

In the case of absorbance (Fig. 2**, C**), principal component one (PC 1) explained 99.1 % of the observed variance, for which all absorbance wavelengths equally contributed. PC 1 distinguished data points throughout time and allowed the distinction between F/2 and the other treatment assays. PC 2 explained only 0.73 % of the variance observed, but it is what allowed for a clear distinction between assays F/2 + N and F/2 + N + P. The major contributions for PC 2 were increments in wavelengths related with chlorophyll and carotenoids, namely violet-blue absorbances (400–500 nm), and red absorbance (640–680 nm), and decreases in debris- related regions, namely yellow-orange absorbances (600–640 nm) and red to infrared absorbances (690–800 nm)

Table 1
Training/testing accuracy of machine learning (ML) models for predicting CC (M cells /mL) using Absorbance or 2D-fluorescence spectroscopy, trained using the architectures tuned with significant robustness.

		ML Pipeline Tuned				*Testing*	
Model ID	**Spectroscopy**	**Algorithm**	**Pre-process**	**Feature Selection / Spectrum used (%)**		**R^2**	**RMSEP (million cells/mL)**
CC/A-15	Abs	SVM	LogBoth	VIP	76	0.91	2.41
CC/A-13	Abs	PLSR	LogBoth	VIP	76	0.90	2.42
CC/A-16	Abs	PLSR	LogBoth	MW-1	12	0.90	2.42
CC/A-20	Abs	PLSR	LogBoth	None	100	0.90	2.44
CC/A-12	Abs	PLSR	MCSN	MW-1	12	0.90	2.49
CC/F-3	2DF	SVM	MCSN	MW-2	10	0.89	2.56
CC/F-LR	2DF	LR	None	EEP 725/765	<1	0.89	2.57
CC/A-8	Abs	SVM	LogOutput	VIP	5	0.89	2.62
CC/A-18	Abs	SVM	LogBoth	MW-3	13	0.89	2.64
CC/F-1	2DF	PLSR	MCSN	MW-2	10	0.89	2.65
CC/A-5	Abs	PLSR	LogOutput	MW-3	12	0.88	2.66
CC/A-6	Abs	PLSR	LogOutput	MW-1	12	0.88	2.66
CC/A-3	Abs	PLSR	LogOutput	VIP	5	0.88	2.66
CC/A-4	Abs	PLSR	LogOutput	MW-2	10	0.88	2.66
CC/A-19	Abs	SVM	MCSN	MW-2	10	0.88	2.66
CC/A-10	Abs	CNNR	LogOutput	VIP	5	0.88	2.67
CC/A-17	Abs	SVM	None	MW-2	10	0.88	2.68
CC/A-9	Abs	SVM	LogOutput	MW-2	10	0.88	2.71
CC/F-2	2DF	PLSR	MCSN	MW-3	2	0.88	2.72
CC/A-1	Abs	CNNR	LogOutput	MW-2	10	0.88	2.73
CC/A-2	Abs	CNNR	LogOutput	MW-3	12	0.88	2.73
CC/A-7	Abs	SVM	LogOutput	MW-3	12	0.88	2.73
CC/A-21	Abs	PLSR	LogOutput	None	100	0.88	2.74
CC/A-14	Abs	SVM	LogOutput	MW-1	12	0.88	2.76
CC/A-11	Abs	CNNR	LogOutput	MW-1	12	0.87	2.85
CC/A-LR	Abs	LR	None	OD663	<1	0.87	2.86
CC/BM	N/A					0.79	3.59

Table 2
Training/testing accuracy of machine learning (ML) models for predicting Fucoxanthin (Fx, ppm) using Absorbance or 2D-fluorescence spectroscopy, trained using the architectures tuned with significant robustness.

		ML Pipeline Tuned				*Testing*	
Model ID	**Spectroscopy**	**Algorithm**	**Pre-process**	**Feature Selection / Spectrum used (%)**		**R^2**	**RMSEP (M/mL)**
Fx/BM	N/A					0.99	0.42
Fx/A-12	Abs	SVM	LogBoth	MW-3	13	0.98	0.65
Fx/A-1	Abs	SVM	LogBoth	MW-2	16	0.98	0.65
Fx/A-7	Abs	SVM	None	MW-3	13	0.98	0.66
Fx/A-11	Abs	SVM	MCSN	MW-3	13	0.98	0.66
Fx/A-3	Abs	PLSR	LogBoth	MW-3	13	0.98	0.67
Fx/A-8	Abs	PLSR	MCSN	MW-3	13	0.98	0.67
Fx/A-9	Abs	SVM	LogBoth	MW-1	12	0.98	0.68
Fx/A-2	Abs	PLSR	LogBoth	MW-2	16	0.98	0.68
Fx/A-5	Abs	SVM	None	MW-1	12	0.98	0.68
Fx/A-10	Abs	PLSR	LogBoth	MW-1	12	0.98	0.68
Fx/A-6	Abs	SVM	MCSN	MW-1	12	0.98	0.69
Fx/A-4	Abs	PLSR	MCSN	MW-1	12	0.98	0.70
Fx/A-17	Abs	SVM	LogBoth	None	100	0.98	0.75
Fx/A-15	Abs	PLSR	MCSN	MW-2	23	0.98	0.75
Fx/F-2	2DF	SVM	LogBoth	MW-1	12	0.98	0.76
Fx/A-14	Abs	PLSR	MCSN	None	100	0.98	0.81
Fx/F-6	2DF	SVM	MCSN	MW-3	2	0.98	0.82
Fx/F-1	2DF	PLSR	LogBoth	MW-1	12	0.98	0.82
Fx/F-12	2DF	PLSR	MCSN	MW-2	10	0.98	0.83
Fx/F-8	2DF	PLSR	MCSN	MW-3	2	0.98	0.84
Fx/F-11	2DF	SVM	LogOutput	None	100	0.98	0.84
Fx/F-13	2DF	SVM	LogBoth	MW-2	18	0.97	0.84
Fx/F-10	2DF	SVM	LogBoth	None	100	0.97	0.84
Fx/F-7	2DF	SVM	LogBoth	VIP	41	0.97	0.85
Fx/F-9	2DF	SVM	MCSN	MW-2	10	0.97	0.86
Fx/A-13	Abs	PLSR	LogBoth	None	100	0.97	0.87
Fx/F-3	2DF	SVM	LogBoth	MW-3	2	0.97	0.89
Fx/F-4	2DF	PLSR	LogBoth	VIP	41	0.96	1.00
Fx/A-16	Abs	PLSR	MCSN	VIP	42	0.97	0.93
Fx/A-LR	Abs	LR	LogOutput	OD442	<1	0.96	1.01
Fx/F-14	2DF	CNNR_2D	None	MW-3	2	0.96	1.07
Fx/F-5	2DF	PLSR	LogBoth	MW-2	18	0.94	1.28
Fx/F-LR	2DF	LR	LogOutput	EEP 645/690	<1	0.92	1.45

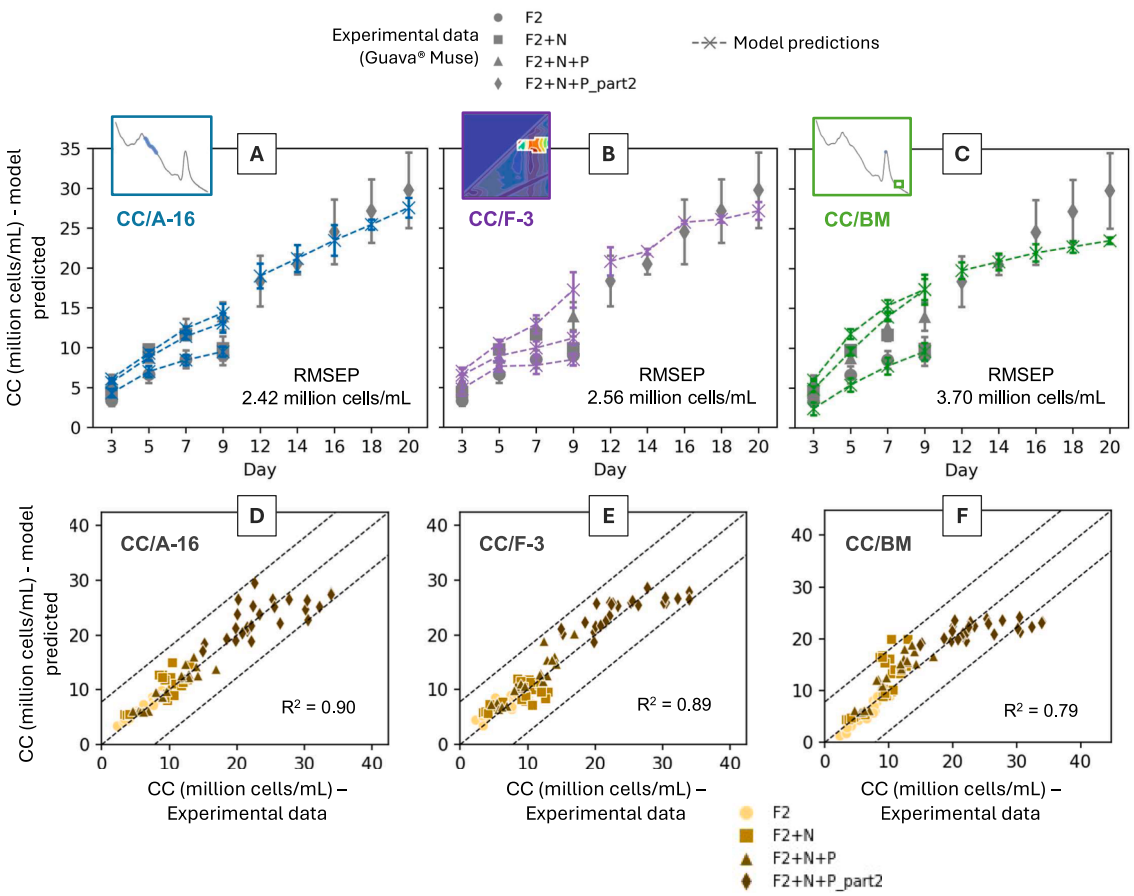


Fig. 4. Testing of machine learning models CC/A-16, CC/F-3, as well of benchmark model CC/BM, showing their accuracy with the experimental data obtained with standard analytical method (Guava® Muse). While A, B and C compare the temporal profiles of each model (colored dotted lines with crosses) with the one provided by the experimental data (grey markers), D, E and F show if models' predictions are leading to deviations bigger than the data's expected dispersion (dotted diagonal lines). The experimental data is plotted with markers according to the assay of origin (circles – F/2 assay, squares – F/2 + N assay, triangles – F/2 + N + P assay, diamonds – F/2 + N + P assay part 2).

(Fig. 2**, C**).

For 2D-fluorescence (Fig. 2**, D**), PC 1 explained 50.3 % of the variance, being its major contributors increments in the chlorophyll fluorescence region (emission from 650 nm) and decreases in amino acid/protein-like (260–300 nm excitation for 280–400 nm emission) and humic compound-like (400–450 nm excitation for 420–500 nm emission) fluorescence regions. The observed variance allowed to distinguish between the assays, also explaining the temporal variation of F/2 + N + P and F/2 + N treatments. PC 2 explained 28.3 % of the variance observed, for which the major contributors were increments in the amino acid/protein and humic compound-like fluorescence regions. Notably, the temporal variation of F/2 was mostly explained by PC2, while the one of F/2 + N + P was explained by PC1.

In summary, these results showed that both spectroscopies provided data to discriminate (in an unsupervised way) the observations of *P. tricornutum* cultures, with 2D-fluorescence providing higher resolution. The results supported the utility for absorbance and 2D-fluorescence for detecting and monitoring the *P. tricornutum* cultivations. Additionally, the 2D-fluorescence spectroscopy PCA model suggests that a considerable portion of the variance detected in the cultivation assays may be linked to the production of other metabolites than pigments; this was observed for assays F/2 and F/2 + N, both showing increased fluorescence intensity in regions related to presence of amino acid/protein-like and humic compound-like substances.

3.3. Machine learning models for predicting fucoxanthin and cell count of P. Tricornutum cultures using absorbance or 2D-fluorescence

3.3.1. Tuning, training, and selection of robust models

For each productivity parameter, fucoxanthin (Fx, ppm) and cell count (CC, million cells/mL), a total of 190 different machine learning (ML) pipelines were generated and tuned for training models: 95 based on absorbance and 95 based on 2D-fluorescence. The architectures employed different modelling techniques (combination of pre-processing, feature selection, ML algorithm and ML hyperparameters) and their performance was evaluated based on the training performance of their models, R^2_{Train} and RMSET, and their robustness to leave-one-assay-out cross-validation, Q^2 and RMSECV.

Fig. 3 represents the 190 models according to their tuning and training performances: each is represented by a data point plotted according to its R^2_{Train} (xx axis) and Q^2 (yy axis). The models are distinguished by the spectroscopy used (colours) and algorithm used (symbols). Moreover, below each model performance point, a bar quantifying the % of spectrum used by the corresponding model is shown. The obtained results showed that most of the models presented very high training performance ($R^2_{Train} \sim 0.90$), with the ones for predicting fucoxanthin showing overall better accuracy ($R^2_{Train} \sim 0.95$) than the ones for predicting CC ($R^2_{Train} \sim 0.90$). This result suggests that fitting models for predicting CC from spectroscopy was a more difficult challenge. This was expected since pigments engage in more interactions with light than any other cellular component in microalgal cells (Gillbro and Cogdell, 1989; Katoh et al., 1991; Wolf and Stevens, 1967) and thus

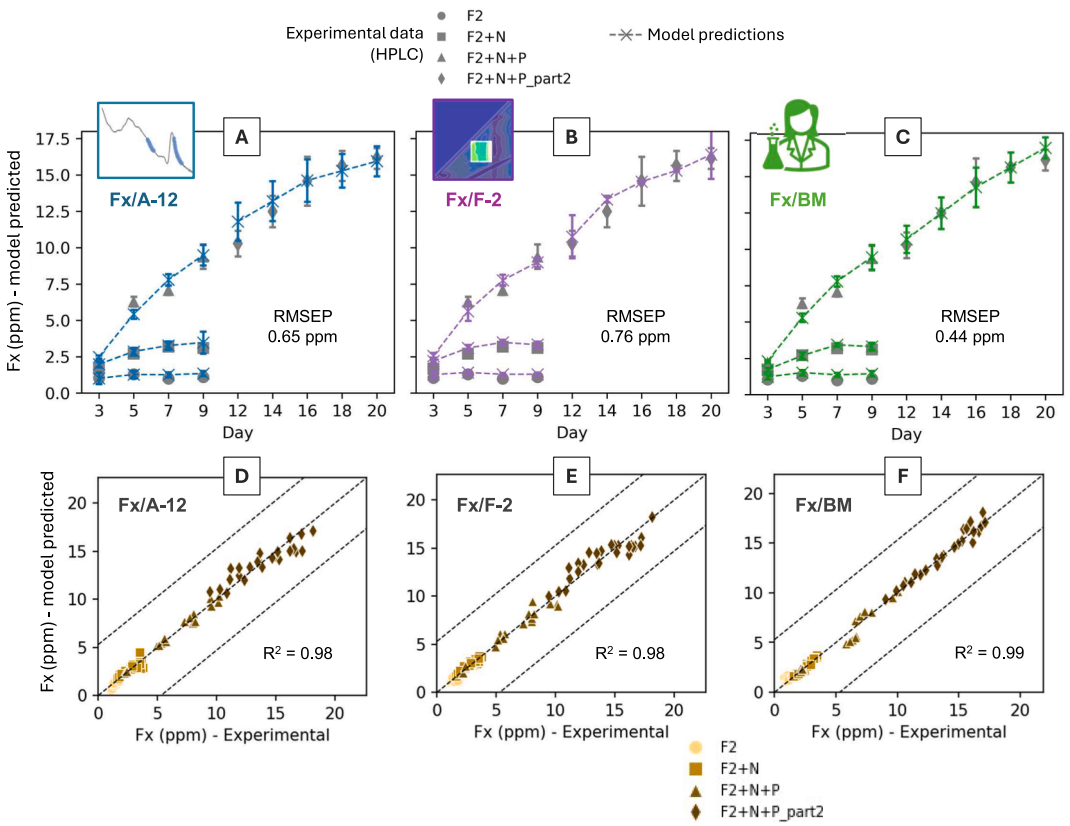


Fig. 5. Testing of machine learning models Fx/A-12, Fx/F-2, as well of benchmark model Fx/BM, showing their accuracy with the experimental data obtained with standard analytical method (HPLC). While A, B and C compare the temporal profiles of each model (colored dotted lines with crosses) with the one provided by the experimental data (grey markers), D, E and F show if models' predictions are leading to deviations bigger than the data's expected dispersion (dotted diagonal lines). The experimental data is plotted with markers according to the assay of origin (circles – F/2 assay, squares – F/2 + N assay, triangles – F/2 + N + P assay, diamonds – F/2 + N + P assay part 2).

their derivation from spectroscopy should be more straightforward.

In contrast to training performance, most models showed poor robustness to leave-one-assay-out cross-validation (LOAOCV), with Q^2 going as low as 0 for models with $R^2_{Train} > 0.90$. This discrepancy between Q^2 and R^2_{Train}, where models showed similar training performance but very different robustness to LOAOCV, is a consequence of overfitting, an expected phenomenon in machine learning (Cawley and Talbot, 2010; Golbraikh and Tropsha, 2002). To select reliable models and avoid overfitting, a model cutoff was arbitrated based on a minimal value of Q^2 (Fig. 3**, marked with dotted line**), below which models were excluded from consideration. Linear regression was chosen as criterion for minimal acceptable Q^2. Being the simplest form of statistical learning, linear regression provides a universal reference for minimal performance of supervised ML models. If a complex ML model has equal or lower robustness than linear regression, it does not justify the additional computational effort. The best Q^2 obtained when applying linear regression to predict fucoxanthin from any single spectral element of absorbance or 2D-fluorescence spectroscopy was 0.94, and for cell concentration was 0.87 (Fig. 3**, marked with *;** see also Supplementary Materials). Thus, models with Q^2 lower than linear regression (i.e., models right to the dotted line in Fig. 3) were considered overfitting and thus excluded from consideration. This resulted in 24 model options for predicting CC (M cells/mL), i.e., 21 absorbance-based and 3 fluorescence-based, and 31 for predicting Fx (ppm), i.e., 17 absorbance-based and 13 fluorescence-based. These models were tested and benchmarked (see results in section 3.3.2, Tables 1 and 2). The model options required only a fraction (from 76 % down to 2 %) of the original spectrum, highlighting the importance of feature selection when developing machine learning models using spectroscopy, as shown by other works (Brá et al., 2008; Brandão et al., 2023; des Touches et al., 2023; Ferreira et al., 2005; Forina et al., 2004; Lu et al., 2014; Xu et al., 2023). Additionally, this provided a decrease in the time required for collecting the spectroscopy data, making the models more reliable for real-time monitoring. Interestingly, among these options, ML algorithms PLSR and SVM were predominant, although for CC the top 2 most robust architectures used the CNN algorithm. Moreover, log-transformation was also predominant. These results suggest that the relationship between productivity parameters (CC and fucoxanthin) and spectroscopy is of non-linear nature, since non-linear relationship algorithms and/or non-linear transformations were used. On the other hand, algorithms RF and XGB were not among the selected options. This suggests either inadequacy of these algorithms for addressing the current challenge, or insufficiency in exploration of the hyperparameter search space. In any case, this result showed the importance of employing more than one ML algorithm when addressing new modelling challenges.

3.3.2. *Testing and benchmarking the selected models*

In the previous section, 55 robust ML models were selected, 24 for predicting CC and 31 for predicting Fx, all reliable options for a user to choose from. To confirm their validity, all 55 options were tested for predictions using spectroscopy data from the testing subset. Their testing performance was evaluated based on R^2 and RMSEP, i.e., on the accuracy and the error in their predictions using the testing data subset (Table 1 **and** Table 2). Additionally, benchmark models were also tested, serving as performance reference for the ML models. The benchmark models chosen are traditional alternatives to standard analytical methods for predicting CC (million cells/mL) and fucoxanthin (ppm), and consist of linear regression models (see Supplementary Materials). They are based on a) optical density of the culture samples at 750 nm (OD750), b) absorbance at 445 nm of pigment extracts from

Table 3
Operational requirements of each technique for monitoring Cell Count (CC, cells/mL) and Fucoxanthin (Fx, ppm) in *Phaeodactylum tricornutum* cultures. (*) – sampling may not be mandatory, an inline probe may be constructed.

Monitoring Parameter	Method	Labour Requirements *Laboratory*	*Computer*
Cell count (cells/mL)	Cytometry (Guava® Muse)	1) Sampling 2) Serial dilution (10^{-1} up to 10^{-4}) 3) Equipment operation, cleaning, and quality check 4) Signal gain and gate geometry adjustments	1) Data collection
	Culture Optical Density at 750 nm (CC/BM)	1) Sampling (*) 2) Absorbance spectroscopy measurement	1) Data collection 2) Apply linear equation
	Absorbance ML Model (CC/A-n)	1) Sampling (*) 2) Absorbance spectroscopy measurement	1) Data collection 2) Run Python executable
	Fluorescence ML Model (CC/F-n)	1) Sampling (*) 2) Fluorescence spectroscopy measurement	1) Data collection 2) Run Python executable
Fucoxanthin (ppm)	HPLC	1) Sampling 2) Production of extracts (serial steps of centrifugation, extraction, filtration, and incubation) 3) Equipment operation, cleaning, and quality check 4) Standard preparation for calibration	1) Data collection 2) Peak selection and integration 3) Calibration curve calculation and peak integration quantification
	Methanol Extract Absorbance at 445 nm (Fx/BM)	1) Sampling 2) Production of extracts (serial steps of centrifugation, extraction, filtration, and incubation) 3) Absorbance spectroscopy measurement	1) Data collection 2) Apply linear equation
	Absorbance ML Model (Fx/A-n)	1) Sampling (*) 2) Absorbance spectroscopy measurement	1) Data collection 2) Run Python executable
	Fluorescence ML Model (Fx/F-n)	1) Sampling (*) 2) Fluorescence spectroscopy measurement	1) Data collection 2) Run Python executable

culture samples (extOD445).

In the case of CC, all 24 ML models outperformed the benchmark, with $R^2 > 0.87$ and RMSEP < 2.4 million cells/mL. These results suggest that the obtained ML models were a superior alternative to the traditional optical density method for cell concentration estimation. The evident choice would be the model presenting the highest R^2 or lowest RMSEP (e.g., CC/A-15 and CC/F-3). However, some ML models were more economic in spectrum usage (e.g. CC/F-2, using only 2 % of 2D-fluorescence spectra), or less complex (e.g. CC/A-12, using the simplest algorithm – PLSR – without log transformations).

Regarding the predictions of fucoxanthin concentrations, no ML model outperformed the benchmark (R^2 = 0.99 and RMSEP = 0.42 ppm). Nevertheless, many of the ML models developed were still very attractive options regarding performance (of the 31 ML models, 21 presented R^2 = 0.98, with RMSEP reaching as low as 0.65 ppm), not to mention the almost zero labour requirements (i.e., no methanol extraction required), as further discussed in the next section.

Based on these results, we suggest that the most suitable models for predicting cell and fucoxanthin concentrations are CC/A-16 and Fx/A-12 using absorbance, and models CC/F-3 and Fx and Fx/F-2 using 2D-fluorescence. The accuracy of these models, as well as the benchmarks, is presented graphically in Figs. 4 and 5 in two different ways: A, B and C show the models predictions over time alongside the standard analytical data, D, E and F display the model predictions versus the standard analytical data, irrespective of time dependence. Both representations serve distinct purposes: A, B and C demonstrate whether the models replicate the growth and productivity profiles provided by the standard analytical methods (Guava® Muse and HPLC) do, while D, E and F indicate whether the models' predictions results in deviations exceeding the expected data dispersion (with dotted diagonal lines marking +/- 1 standard deviation).

All six models, along with the remaining 51 reliable models, provide data that does not significantly differ from the information using standard methods (see statistical analysis results in the Supplementary Material). These models for further testing by any user on the GitHub platform (https://github.com/ibetbio/PT-FucoFromSpec), along with Python scripts for their application. They require only the spectroscopy data in ".xlsx" format, and they return CC (million cells/mL) or Fucoxanthin (ppm) values.

These models were tuned for robustness, so they should be applicable across various studies (e.g., variations in temperature, light, etc.), as long as the key responses to be monitored remain growth and productivity. Nonetheless, the methodology developed in this study is flexible and can easily be extended to accommodate new data into the existing dataset. Recalibration, or the development of new models using the current methodology, is only necessary if new cultivations produce growth and productivity responses that differ significantly from the wide variance already covered by our dataset, requiring the models to extrapolate rather than interpolate.

3.3.3. *Operational labour savings*

The results obtained in our work revealed that machine learning models based on spectroscopy are reliable alternatives to laborious and expensive standard analytical tools. Data presented in Table 3 shows the labour requirements of using ML models when compared to standard methods (HPLC and cytometry) and their respective traditional alternatives (extract absorbance and culture optical density). The ML models and the traditional alternatives to standard methods provide significant labour savings, although in the case of fucoxanthin the absorbance of extract (benchmark model Fx/BM) still requires procedures in the laboratory (i.e., production of extracts). This makes ML models very attractive options, as they provide comparable accuracy while saving much effort. In the case of cell concentration, optical density of the culture (CC/BM) requires the same laboratory effort as ML models but becomes the less attractive option because of its lower accuracy (see previous section). Indeed, the ML models based on spectroscopy, as well as the optical density of the culture, do not require any laboratorial procedures, except for sampling. Still, the sampling step can be removed by using inline optical probes, an advantage exclusive of spectroscopy-based models (Havlik et al., 2022; Shin et al., 2018). Specific probes tailored to each model configuration may be designed and applied, and this work is currently being done by our group.

4. Conclusion

Our findings demonstrated that spectroscopy-based machine learning models effectively monitor the significant impact of media formulations on the growth and fucoxanthin productivity of *P. tricornutum* cultivation. The developed models and the accompanying dataset are publicly available, enabling further research by the scientific community. Additionally, the straightforward deployment of these ML

models has the potential to replace expensive and environmentally harmful standard analytical methods, offering substantial savings in both cost and time.

Data availibility

The dataset generated in this study, comprising 255 observations of spectroscopy and standard analytical data from axenic *Phaeodactylum tricornutum* cultivation under various medium formulations, along with the Python scripts developed, is available for download on the developer platform GitHub (https://github.com/ibetbio/PT-FucoFromSpec).

CRediT authorship contribution statement

Pedro Reynolds-Brandão: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Francisco Quintas-Nunes:** Writing – review & editing, Methodology, Investigation. **Constança D.F. Bertrand:** Writing – review & editing, Methodology, Investigation. **Rodrigo M. Martins:** Writing – review & editing, Methodology, Investigation. **Maria T. B. Crespo:** Supervision. **Cláudia F. Galinha:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Francisco X. Nascimento:** Writing – review & editing, Supervision, Project administration, Conceptualization.

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.

Acknowledgements

For supporting this work, we acknowledge the project "PhycoμBiome: Understanding and harnessing the power of the microalgae microbiome aiming the maximization of marine microalgae productivity" funded by Fundação para a Ciência Tecnologia/ Ministério da Ciência, Tecnologia e Ensino Superior (FCT/MCTES, Portugal), grant PTDC/BAA-BIO/1262/2020 (http://doi.org/10.54499/PTDC/BAA-BIO/1262/2020). We also acknowledge the support of the European Project MULTI-STR3AM, funded by Bio Based Industries Joint Undertaking (JU) under grant agreement No. 512 887227. The JU receives support from the European Union's Horizon 2020 research and innovation program and the Bio Based Industries Consortium. The research was performed with the support of the R&D Units 'GREEN-IT-Bioresources for Sustainability' (UIDB/04551/2020, DOI: 10.54499/UIDB/04551/2020 and UIDP/04551/2020, DOI: 10.54499/UIDP/04551/2020), iNOVA4Health (UIDB/04462/2020, and UIDP/04462/2020), LS4FUTURE Associated Laboratory (LA/P/0087/2020, DOI: 10.54499/LA/P/0087/2020) and LAQV (LA/P/0008/2020 DOI 10.54499/LA/P/0008/2020, UIDP/50006/2020 DOI 10.54499/UIDP/50006/2020 and UIDB/50006/2020 DOI 10.54499/UIDB/50006/2020) funded by FCT/MCTES. FCT/MCTES is also acknowledge by PRB and CFG for funding the PhD grant 2021.07927.BD and funding through the Scientific Employment Stimulus - Individual Call (2022.04601.CEECIND).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2024.131988.

References

Ashokkumar, V., Flora, G., Sevanan, M., Sripriya, R., Chen, W. H., Park, J. H., Rajesh banu, J., & Kumar, G. (2023). Technological advances in the production of carotenoids and their applications– A critical review. In *Bioresource Technology* (Vol. 367). Elsevier Ltd. https://doi.org/10.1016/j.biortech.2022.128215.

Afonso, C., Bragança, A.R., Rebelo, B.A., Serra, T.S., Abranches, R., 2022. Optimal Nitrate Supplementation in Phaeodactylum tricornutum Culture Medium Increases Biomass and Fucoxanthin Production. Foods 11 (4). https://doi.org/10.3390/FOODS11040568.

Bayer, B., von Stosch, M., Melcher, M., Duerkop, M., Striedner, G., 2020. Soft sensor based on 2D-fluorescence and process data enabling real-time estimation of biomass in Escherichia coli cultivations. *Eng. Life Sci.* 20 (1–2), 26–35. https://doi.org/10.1002/elsc.201900076.

Biancolillo, A., Marini, F., 2018. Chemometric methods for spectroscopy-based pharmaceutical analysis. *Front. Chem.* 6 (NOV), 1–14. https://doi.org/10.3389/fchem.2018.00576.

Brá, L.P., Lopes, M., Ferreira, A.P., Menezes, J.C., 2008. A bootstrap-based strategy for spectral interval selection in PLS regression. *Journal of Chemometrics* 22 (11–12), 695–700. https://doi.org/10.1002/cem.1153.

Brandão, P.R., Sá, M., Galinha, C.F., 2023. Learning from fluorescence: A tool for online multiparameter monitoring of a microalgae culture. *Comput. Chem. Eng.* 179. https://doi.org/10.1016/j.compchemeng.2023.108452.

Busse, C., Biechele, P., de Vries, I., Reardon, K.F., Solle, D., Scheper, T., 2017. Sensors for disposable bioreactors. *Eng. Life Sci.* 17 (8), 940–952. https://doi.org/10.1002/elsc.201700049.

Butler, T., Kapoore, R.V., Vaidyanathan, S., 2020. Phaeodactylum tricornutum: A Diatom Cell Factory. *Trends Biotechnol.* 38 (6), 606–622. https://doi.org/10.1016/j.tibtech.2019.12.023.

Cawley, G.C., Talbot, N.L.C., 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. In. *J. Mach. Learn. Res.* 11.

Chen, C., Gong, N., Li, Z., Sun, C., Men, Z., 2017. Concentration effect on quenching of chlorophyll a fluorescence by all-trans-β-carotene in photosynthesis. *Molecules* 22 (10). https://doi.org/10.3390/molecules22101585.

des Touches, T., Munda, M., Cornet, T., Gerkens, P., Hellepute, T., 2023. Feature selection with prior knowledge improves interpretability of chemometrics models. *Chemometrics and Intelligent Laboratory Systems* 240. https://doi.org/10.1016/j.chemolab.2023.104905.

Ducklow, H., & Dickson, A. (1994, January). *Chapter 11. The Determination of Phosphorus in Sea Water* . JGOFS Protocols.

Esben, J., Bjerrum, M., & Glahder, T. S. (2017). *Data Augmentation of Spectral Data for Convolutional Neural Network (CNN) Based Deep Chemometrics*. https://doi.org/https://doi.org/10.48550/arXiv.1710.01927.

Faassen, S. M., & Hitzmann, B. (2015). Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. In *Sensors (Switzerland)* (Vol. 15, Issue 5, pp. 10271–10291). MDPI AG. https://doi.org/10.3390/s150510271.

Ferreira, A.P., Alves, T.P., Menezes, J.C., 2005. Monitoring complex media fermentations with near-infrared spectroscopy: Comparison of different variable selection methods. *Biotechnol. Bioeng.* 91 (4), 474–481. https://doi.org/10.1002/bit.20526.

Forina, M., Lanteri, S., Oliveros, M.C.C., Millan, C.P., 2004. Selection of useful predictors in multivariate calibration. *Anal. Bioanal. Chem.* 380 (3 SPEC.ISS.), 397–418. https://doi.org/10.1007/s00216-004-2768-x.

Galinha, C.F., Carvalho, G., Portugal, C.A.M., Guglielmi, G., Reis, M.A.M., Crespo, J.G., 2012. Multivariate statistically-based modelling of a membrane bioreactor for wastewater treatment using 2D fluorescence monitoring data. *Water Res.* 46 (11), 3623–3636. https://doi.org/10.1016/j.watres.2012.04.010.

Gillbro, T., Cogdell, R.J., 1989. Carotenoid fluorescence. *Chem. Phys. Lett.* 158 (3–4), 312–316. https://doi.org/10.1016/0009-2614(89)87342-7.

Golbraikh, A., Tropsha, A., 2002. Beware of q2! *J. Mol. Graph. Model.* 20 (4), 269–276. https://doi.org/10.1016/S1093-3263(01)00123-1.

Griffiths, M.J., Garcin, C., van Hille, R.P., Harrison, S.T.L., 2011. Interference by pigment in the estimation of microalgal biomass concentration by optical density. *J. Microbiol. Methods* 85 (2), 119–123. https://doi.org/10.1016/j.mimet.2011.02.005.

Grote, B., Zense, T., Hitzmann, B., 2014. 2D-fluorescence and multivariate data analysis for monitoring of sourdough fermentation process. *Food Control* 38 (1), 8–18. https://doi.org/10.1016/j.foodcont.2013.09.039.

Havlik, I., Beutel, S., Scheper, T., Reardon, K.F., 2022. On-Line Monitoring of Biological Parameters in Microalgal Bioprocesses Using Optical Methods. MDPI In *Energies* 15 (3). https://doi.org/10.3390/en15030875.

Huang, B., Marchand, J., Blanckaert, V., Lukomska, E., Ulmann, L., Wielgosz-Collin, G., Rabesaotra, V., Moreau, B., Bougaran, G., Mimouni, V., Morant-Manceau, A., 2019. Nitrogen and phosphorus limitations induce carbon partitioning and membrane lipid remodelling in the marine diatom Phaeodactylum tricornutum. Eur. J. Phycol. 54 (3), 342–358. https://doi.org/10.1080/09670262.2019.1567823.

Katoh, T., Nagashima, U., Mimuro, M., 1991. Fluorescence properties of the allenic carotenoid fucoxanthin : Implication for energy transfer in photosynthetic pigment systems. *Photosynth. Res.* 27, 221–226.

Lakowicz, J.R., 2006. *Principles of fluorescence spectroscopy*. Springer.

Leong, Y. K., Chen, C. Y., Varjani, S., & Chang, J. S. (2022). Producing fucoxanthin from algae – Recent advances in cultivation strategies and downstream processing. In *Bioresource Technology* (Vol. 344). Elsevier Ltd. https://doi.org/10.1016/j.biortech.2021.126170.

Levitan, O., Dinamarca, J., Zelzion, E., Lun, D.S., Guerra, L.T., Kim, M.K., Kim, J., Vam Mooy, Bhattacharya, D., Falkowski, P.G., 2015. Remodeling of intermediate metabolism in the diatom Phaeodactylum tricornutum under nitrogen stress. Proc. Natl. Acad. Sci. USA 112 (2), 412–417. https://doi.org/10.1073/PNAS.1419818112/SUPPL_FILE/PNAS.1419818112.SD01.XLSX.

Li, F.L., Wang, L.J., Fan, Y., Parsons, R.L., Hu, G.R., Zhang, P.Y., 2018. A rapid method for the determination of fucoxanthin in diatom. *Mar. Drugs* 16 (1), 1–13. https://doi.org/10.3390/md16010033.

Lichtenthaler, H.K., Buschmann, C., 2001. Chlorophylls and Carotenoids: Measurement and Characterization by UV-VIS Spectroscopy. F4.3.1–F4.3.8 *Curr. Protocol Food Anal. Chem.*. 1 (1). https://doi.org/10.1002/0471142913.faf0403s01.

Liu, J. Y., Zeng, L. H., & Ren, Z. H. (2021). The application of spectroscopy technology in the monitoring of microalgae cells concentration. In *Applied Spectroscopy Reviews* (Vol. 56, Issue 3, pp. 171–192). Bellwether Publishing, Ltd. https://doi.org/10.1080/05704928.2020.1763380.

Lu, B., Castillo, I., Chiang, L., Edgar, T.F., 2014. Industrial PLS model variable selection using moving window variable importance in projection. *Chemometrics and Intelligent Laboratory Systems* 135, 90–109. https://doi.org/10.1016/j.chemolab.2014.03.020.

McClure, D.D., Luiz, A., Gerber, B., Barton, G.W., Kavanagh, J.M., 2018. An investigation into the effect of culture conditions on fucoxanthin production using the marine microalgae Phaeodactylum tricornutum. Algal Res. 29, 41–48. https://doi.org/10.1016/J.ALGAL.2017.11.015.

Neumann, U., Derwenskus, F., Flister, V.F., Schmid-Staiger, U., Hirth, T., Bischoff, S.C., 2019. Fucoxanthin, a carotenoid derived from Phaeodactylum tricornutum exerts antiproliferative and antioxidant activities in vitro. *Antioxidants* 8 (6), 2019. https://doi.org/10.3390/antiox8060183.

Porras Reyes, L., Havlik, I., & Beutel, S. (2024). Software sensors in the monitoring of microalgae cultivations. In *Reviews in Environmental Science and Biotechnology* (Vol. 23, Issue 1, pp. 67–92). Springer Science and Business Media B.V. https://doi.org/10.1007/s11157-023-09679-8.

Pradhan, N., Kumar, S., Selvasembian, R., Rawat, S., Gangwar, A., Senthamizh, R., Yuen, Y.K., Luo, L., Ayothiraman, S., Saratale, G.D., Mal, J., 2023. Emerging trends in the pretreatment of microalgal biomass and recovery of value-added products: A review. *Bioresource Technology* 369. https://doi.org/10.1016/j.biortech.2022.128395.

Rowles, J.L., Erdman, J.W., 2020. Carotenoids and their role in cancer prevention. *BBA - Molecular and Cell Biology of Lipids* 1865 (11), 158613. https://doi.org/10.1016/j.bbalip.2020.158613.

Sá, M., Ferrer-Ledo, N., Gao, F., Bertinetto, C.G., Jansen, J., Crespo, J.G., Wijffels, R.H., Barbosa, M., Galinha, C.F., 2022. Perspectives of fluorescence spectroscopy for online monitoring in microalgae industry. *J. Microbial. Biotechnol.* 15 (6), 1824–1838. https://doi.org/10.1111/1751-7915.14013.

Shin, Y.H., Gutierrez-Wing, M.T., Choi, J.W., 2018. A field-deployable and handheld fluorometer for environmental water quality monitoring. Micro and Nano Syst. Lett. 6 (1). https://doi.org/10.1186/s40486-018-0078-x. Society of Micro and Nano Systems.

Teixeira, A.P., Duarte, T.M., Oliveira, R., Carrondo, M.J.T., Alves, P.M., 2011. High-throughput analysis of animal cell cultures using two-dimensional fluorometry. *J. Biotechnol.* 151 (3), 255–260. https://doi.org/10.1016/j.jbiotec.2010.11.015.

Vílchez, C., Forján, E., Cuaresma, M., Bédmar, F., Garbayo, I., Vega, J.M., 2011. Marine carotenoids: Biological functions and commercial applications. *Mar. Drugs* 9 (3), 319–333. https://doi.org/10.3390/md9030319.

Wang, L., Lin, K., Guo, H., Zhang, Y., 2022a. Spectrophotometric determination of nitrate in small volume of seawater samples using a simple resorcinol method. *Anal. Bioanal. Chem.* 414 (19), 5869–5876. https://doi.org/10.1007/s00216-022-04152-x.

Wang, Z.P., Wang, P.K., Ma, Y., Lin, J.X., Wang, C.L., Zhao, Y.X., Zhang, X.Y., Huang, B. C., Zhao, S.G., Gao, L., Jiang, J., Wang, H.Y., Chen, W., 2022b. Laminaria japonica hydrolysate promotes fucoxanthin accumulation in Phaeodactylum tricornutum. *Bioresource Technology* 344. https://doi.org/10.1016/j.biortech.2021.126117.

Wolf, F.T., Stevens, M.V., 1967. the Fluorescence of Carotenoids. *Photochem. Photobiol.* 6 (8), 597–599. https://doi.org/10.1111/j.1751-1097.1967.tb08761.x.

Xu, X., Teng, G., Wang, Q., Zhao, Z., Wei, K., Bao, M., Zheng, Y., Luo, T., 2023. Spectral preprocessing combined with feature selection improve model robustness for plastics samples classification by LIBS. *Front. Environ. Sci.* 11. https://doi.org/10.3389/fenvs.2023.1175392.

Yang, R., Wei, D., 2020. Improving Fucoxanthin Production in Mixotrophic Culture of Marine Diatom Phaeodactylum tricornutum by LED Light Shift and Nitrogen Supplementation. *Front. Bioeng. Biotechnol.* 8 (July). https://doi.org/10.3389/fbioe.2020.00820.

Yi, Z., Su, Y., Cherek, P., Nelson, D.R., Lin, J., Rolfsson, O., Wu, H., Salehi-Ashtiani, K., Brynjolfsson, S., Fu, W., 2019. Combined artificial high-silicate medium and LED illumination promote carotenoid accumulation in the marine diatom Phaeodactylum tricornutum. *Microb. Cell Fact.* 18 (1). https://doi.org/10.1186/s12934-019-1263-1.

Zhang, D., Del Rio-Chanona, Petsagkourakis, P., Wagner, J., 2019. Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. Biotechnol. Bioeng. 116 (11), 2919–2930. https://doi.org/10.1002/bit.27120.

Zhuang, G.J., Ye, Y., Zhao, J., Zhou, C., Zhu, J., Li, Y., Zhang, J., Yan, X., 2023. Valorization of Phaeodactylum tricornutum for integrated preparation of diadinoxanthin and fucoxanthin. *Bioresource Technology* 385. https://doi.org/10.1016/j.biortech.2023.129412.