Computers and Chemical Engineering 179 (2023) 108452

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

# Learning from fluorescence: A tool for online multiparameter monitoring of a microalgae culture

Pedro R. Brandão, Marta Sá [1], Claudia F. Galinha [*]

*LAQV-REQUIMTE, Department of Chemistry, NOVA School of Science and Technology, FCT NOVA, Universidade NOVA de Lisboa, Caparica 2829-516, Portugal*

ARTICLE INFO

*Keywords:*
2D fluorescence
Machine learning
Microalgae cultivation
Excitation-emission matrices (EEMs)
Bioprocess monitoring
Projection to latent structures regression (PLSR)

ABSTRACT

We propose a systematic approach for monitoring important productivity parameters in a *Dunaliella salina* culture using 2D fluorescence data. For this purpose, a methodology based on Machine Learning algorithm Projection to Latent Structures Regression (PLSR) coupled with variable selection strategies was used. Additionally, a robustness analysis is proposed to support the validation of the yielded models and provide a measure of their reliability. Robust (i.e., $Q^2 \geq 0.5$) and parsimonious (i.e., selecting down to 3 % of the fluorescence variables present in a 250–700 nm wavelength excitation-emission matrix) models were obtained for monitoring cell count, chlorophyll *b*, total carotenoids and β-carotene culture concentration, and the ratio between total carotenoids and total chlorophylls, all of which were validated with a left-out batch performing with $R^2$ higher than 0.7 except for β-carotene ($R^2 = 0.54$).

## 1. Introduction

Microalgae are a diverse group of photosynthetic microorganisms that became considerably popular, having captured increasing interest and investment in the last 40 years (Hamed, 2016; Patel et al., 2017). Although formerly viewed as sustainable feedstocks for inexpensive applications such as fuel and feed (Vanthoor-Koopmans et al., 2013), nowadays microalgae are recognized as efficient systems to produce a variety of high-value nutraceuticals and ingredients with important applications in human health and nutrition. Particularly, microalgae carotenoids such as lutein, zeaxanthin and beta-carotenes have been shown to have potent antioxidant capabilities, as well as capacity to modulate gene expression, enhance immune function, and exert anti-inflammatory effects (Barkia et al., 2019; Chew et al., 2017; Khan et al., 2018). Additionally, microalgae production is remarkably sustainable, being independent of arable land systems and allowing for water and carbon recycling (Khan et al., 2018).

Just like with any living cell culture, the process of cultivating microalgae and their products relies on the control of their metabolism, which is highly dependant on the culture's biochemical environment (Roth, 1978). Therefore, to maximize productivity from microalgae-based systems while maintaining the quality of the intended products, key culture parameters need to be accurately and constantly monitored. While physical parameters can be accurately monitored in real time, chemical and biological parameters require expensive and time-consuming chromatographic equipment, making them challenging to assess (Cuellar-Bermudez et al., 2015; Glindkamp et al., 2009).

Optical probes based on spectroscopy data (e.g., absorbence, fluorescence, Raman, NMR, etc.) are operationally inexpensive and are fit for online monitoring of biological systems (Li and Humphrey, 1991; Lindemann et al., 1998; Marose et al., 1998). Two dimensional (2D) has been shown to provide information not only on the activity of various fluorophores simultaneously but also on physicochemical properties such as pH, polarizability, ionic strength, solubility, etc., being already considered a status fingerprint for biological systems (Amigo and Marini, 2013; Forina et al., 1987; Galinha et al., 2011b; Lakowicz, 2006), making it a potentially great tool for monitoring microalgae cultures. In fact, several studies have already shown the successful application of 2D fluorescence spectroscopy for monitoring a variety of biological processes, such as wastewater treatment (Galinha et al., 2011b, 2011a; 2012), microbial fermentation (Bayer et al., 2020; Tartakovsky et al., 1996) and animal cell cultivation (Graf et al., 2019; Podrazký et al., 2003; Teixeira et al., 2011).

Regarding microalgae cultivation, previous work demonstrated the applicability of 2D fluorescence spectroscopy for monitoring important process-related biological parameters (e.g. cell number and viability,

* Corresponding author.
*E-mail address:* cf.galinha@fct.unl.pt (C.F. Galinha).
[1] Present address: Stichting imec Nederland - OnePlanet Research center, 6708WH Wageningen, The Netherlands.

https://doi.org/10.1016/j.compchemeng.2023.108452
Received 19 April 2023; Received in revised form 3 October 2023; Accepted 6 October 2023
Available online 7 October 2023

chlorophyll and carotenoid content) (Sá et al., 2020a, 2020b, 2019, 2017). In these works, machine learning models, based on algorithm Projection to Latent Structures Regression, were trained to derive biological parameters from a selection of principal components of compressed 2D fluorescence data and climatic data variables. For this purpose, a compression step was performed by Principal Component Analysis, where the multivariate EEMs were simplified to not more than a dozen principal components of variance, and then used alongside climatic variables for PLSR modelling. To our knowledge, these works are the first reports on application of 2D fluorescence spectroscopy in microalgae related bioprocesses.

The present work aims to continue the effort of optimizing the machine learning modelling methodology for using 2D fluorescence spectroscopy data for monitoring microalgae production. Using the previous strategy, the user not only has to collect climatic data but also is required to perform a vast 2D fluorescence scan, i.e., within 250 and 700 nm, affecting the real-time monitoring application since the acquisition of EEMs of this magnitude requires about 10 min. In the present work, a new modelling strategy demonstrates that portions that go as low as 3 % of the complete EEM are sufficient for validating 2D fluorescence as a standalone tool for biological parameter monitoring. For that purpose, a dataset on *Dunaliella salina* cultures from the previous work (Sá et al., 2020b) was further explored with this new strategy, where algorithm Projection to Latent Structures Regression (PLSR) is used directly and combined with variable selection strategies for identification of relevant wavelength areas within a 2D fluorescence excitation emission matrix (EEM) for each biological parameter.

## 2. Methods

The data used in this work was the same explored by Sá et al. (2020b), and it is a result of the monitoring of batch induction cultures of *Dunaliella salina* (from green to orange) performed outdoors at pilot-scale. A total of 6 batches, hereby noted A, B, C, D, E and F, were monitored through sampling each 2 to 4 days, resulting in 41 samples.

Each sample provides data on 11 biological parameters and an excitation-emission matrix (EEM) containing 4093 fluorescence variables. The projection to latent structures regression (PLSR) algorithm will take the fluorescence variables as model inputs and each of the biological parameters as the expected outputs.

### 2.1. Biological parameters

The biological parameters (Table 1) were measured using reference methods (Sá et al., 2020b), namely flow cytometry analysis using Guava MUSE Cell analyzer, and pigment analysis of methanol extracts made from the samples using either absorbence spectrophotometry, based on direct application of the modified Arnon's equations (Lichtenthaler and Buschmann, 1987), or using HPLC.

The resulting data was stored in a 41 × 11 matrix of outputs, $\boldsymbol{O}$

**Table 1**
Biological parameters and their identification number p.

| Biological parameter | p |
|---|---|
| Cell Count ($10^6$ cells/L) | 1 |
| Chlorophyll *b* (mg/L) | 2 |
| Chlorophyll a (mg/L) | 3 |
| Carotenoids (mg/L) | 4 |
| Carotenoids/Chlorophylls | 5 |
| Proteins (mg/L) | 6 |
| Lutein (mg/L) | 7 |
| Zeaxanthin (mg/L) | 8 |
| α-carotene (mg/L) | 9 |
| β-carotene (mg/L) | 10 |
| 9-cis-β-carotene (mg/L) | 11 |

$$\boldsymbol{O} = \begin{bmatrix} o_1^1 & o_2^1 & \cdots & o_{11}^1 \\ o_1^2 & o_2^2 & \cdots & o_{11}^2 \\ \vdots & \vdots & \ddots & \vdots \\ o_1^{41} & o_2^{41} & \cdots & o_{11}^{41} \end{bmatrix} \quad (1)$$

Where $o_p^i$ represents the value of biological parameter $p$ of sample $i$.

### 2.2. 2D fluorescence: excitation-emission matrices

The EEM scans were performed in the excitation range of 250 to 690 nm and in the emission range of 260 to 700 nm, with a 5 nm step. This results in 41 matrices containing 89 × 89 = 7921 fluorescence variables.

Each of the EEMs was first processed for removal of the fluorescence variables whose wavelengths of emission are below wavelength of excitation. This resulted in 41 EEMs with 4093 elements of interest, which were then unfolded into 4093-dimensional vectors with the format of equation 2.

$$\boldsymbol{\Lambda\Lambda}^i = \left(\lambda\lambda_1^i, \lambda\lambda_2^i, \ldots, \lambda\lambda_j^i, \ldots, \lambda\lambda_{4093}^i\right) \quad (2)$$

Where $\boldsymbol{\Lambda\Lambda}^i$ represents the unfolded form of the EEM that was scanned from sample $i$ and $\lambda\lambda_j^i$ represents fluorescence variable number $j$ of sample $i$ (e.g., $\lambda\lambda_1^1$ represents the fluorescence intensity emitted at 260 nm and excited at 250 nm for sample 1 and $\lambda\lambda_{4093}^{41}$ represents fluorescence intensity emitted at 700 nm and excited at 690 nm for sample 41).

All the 41 unfolded EEMs were then stored in a 41 × 4093 matrix of inputs, $\boldsymbol{I}$

$$\boldsymbol{I} = \begin{bmatrix} \left(\boldsymbol{\Lambda\Lambda}^1\right)^T \\ \left(\boldsymbol{\Lambda\Lambda}^2\right)^T \\ \vdots \\ \left(\boldsymbol{\Lambda\Lambda}^{41}\right)^T \end{bmatrix} = \begin{bmatrix} \lambda\lambda_1^1 & \lambda\lambda_2^1 & \cdots & \lambda\lambda_{4093}^1 \\ \lambda\lambda_1^2 & \lambda\lambda_2^2 & \cdots & \lambda\lambda_{4093}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda\lambda_1^{41} & \lambda\lambda_2^{41} & \cdots & \lambda\lambda_{4093}^{41} \end{bmatrix} \quad (3)$$

### 2.3. Principal component analysis for identification of outlier observations

Principal Component Analysis (PCA) was applied to both the data on the biological parameters and the data on 2D fluorescence, but singularly used for detecting outlier observations. The number of principal components was optimized by leave-one-out-cross-validation (LOOCV) (Efron and Gong, 1983). For this purpose, Hotelling's $T^2$ test was performed, and observations which exceeded 99 % of the $T^2$ range were deemed outliers.

### 2.4. Projection to latent structures regression (PLSR)

The algorithm used for Projection to Latent Structures Regression (PLSR) used in this work is SIMPLS (De Jong, 1993) for estimating a univariate output $\boldsymbol{o_p}$ from a multivariate input $\boldsymbol{I}$. More specifically, PLSR returns a model for making an estimate, $\widehat{\boldsymbol{o}}_{\boldsymbol{p}}$, for each biological parameter, $\boldsymbol{o_p}$, from the fluorescence variables, $\boldsymbol{I}$. The model is a multilinear equation where the fluorescence variables are multiplied by a vector of regression coefficients, $\boldsymbol{b}$, resulting in a prediction:

$$\widehat{\boldsymbol{o}}_{\boldsymbol{p}} = \boldsymbol{I} \,.\, \boldsymbol{b} \; + \boldsymbol{e} = \boldsymbol{I} \,.\, \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} + \boldsymbol{e} \quad (4)$$

where $\boldsymbol{b_j}$ represents the coefficient of regression attributed to $\boldsymbol{\lambda\lambda_j}$ and $\boldsymbol{e}$ the residuals vector.

### 2.5. Datasets

PLSR may not find acceptable solutions for cases where there is a non-linear relationship between inputs and outputs (Berglund and Wold, 2007). To overcome this possibility, the application of Box-transformations is commonly used in data analysis (Box and Cox, 1964). In the present work, we applied to the dataset a λ=0 Box-Cox transformation, which is an element-wise logarithm transformation. Thus, 2 datasets were used in this work (see Table 2): Dataset "Ori", which consists of writing both $\boldsymbol{I}$ and $\boldsymbol{O}$ matrices in a Microsoft Excel worksheet and Dataset "Log", which consists of an element-wise logarithm transformation of $\boldsymbol{I}$ and $\boldsymbol{O}$. In the case of $\boldsymbol{I}$ log transformation, the elements whose values were lower than 1 were set to 1. This condition is important because log $x$ varies extremely when $x$ tends to 0 and, thus it can amplify noise, which in this data ranges from 0 to 5.

### 2.6. Model-wise and data-wise selection of fluorescence variables within the EEM

The selection of fluorescence variables was performed by three methods: a model-wise method, a data-wise method, and a hybrid one (see Table 3).

The model-wise method consists of selecting fluorescence variables according to their importance to the modelling using the full EEM; the selection criteria is thus called variable importance to projection (VIP) coefficient (Lazraq et al., 2003). The computation of VIP for each fluorescence variable can be found in the supplementary material. This method consists of using a criterion $c$ for selecting variables with $VIP_j \geq c$ (see Table 3). This criterion takes values within an interval between a minimum and maximum with a variable step (it depends on the steepness of the decrease in number of fluorescence variables by varying criterion $c$).

The data-wise method consists of applying Moving-Window-PLSR (MWPLSR) (Balabin and Smirnov, 2011; Jiao et al., 2016). It consists of restricting quadrangular areas of variable side and position in the EEM, hence called moving window, and then training PLSR models with only the fluorescence variables within that area.

Finally, the hybrid method is a PLSR model applied to the top 3 best training-performing windows obtained by MWPLSR.

### 2.7. Modelling procedure

This work analyses PLSR without variable selection, and PLSR with variable selection based on VIP or MWPLSR. For this purpose, the following 2 step modelling procedure was followed: a Machine Learning step and a Robustness Analysis step. Fig. 1 schematizes the modelling procedure.

#### 2.7.1. Machine learning step: tuning a learning architecture

In this work we define the learning architecture as the set of parameters that need to be set for applying a machine learning algorithm. This learning architecture needs to be tuned for the data by trial and error until a certain criterion is met (e.g., root mean squared error minimization) (Bernardo and Smith, 1994), a procedure also known as training. In the present case, the parameters to be optimized are the type of dataset transformation, the fluorescence variables to use, and the number of modelling latent variables. These will hereby be named hyperparameters, to distinguish them from the culture biological parameters.

For this purpose, a data split (Fig. 1, box 1) was performed where batches A, B, D, E and F were used for training and batch C was left out for later testing (see Section 2.7.2).

To obtain the optimal learning architecture, i.e., the optimal combination of type of dataset transformation, $DT_{opt}$, fluorescence variables to be used, $FV_{opt}$, and number of latent variables to be used, $LV_{opt}$, a strategy of leave-one-out-cross-validation (LOOCV) was used (Fig. 1, box 2). For each learning architecture, PLSR models are computed for all the combinations of the dataset that can be obtained by leaving one observation out and then all those models are tested using the respective left-out observation. Each of the models' predictions obtained and respective true value are then used for calculating a percentage error of cross-validation, $ECV$ (%) (formula is found supplementary material). The optimal learning architecture is the one for which the cross-validation PLSR models yielded the lowest $ECV$.

Also, an early stop was implemented to the number of latent variables to avoid too much complexity in the models; this is because even though LOOCV is a good tool for optimization of PLSR models it is also susceptible to overfitting (Cawley and Talbot, 2010; Golbraikh and Tropsha, 2002), especially in a case where the number of observations (i.e., degrees of freedom) is low. Thus, the early stop criterion used in this work is that the latent variables should not be higher than 8 latent variables (i.e., 20 % of the number of observations).

#### 2.7.2. Second step: robustness analysis of the architecture

The optimal architecture is used to compute a PLSR model using the training stage data (i.e., batches A, B, D, E and F) yielding a model ready for work, hereby noted as model W (Fig. 1, connection from box 2 to box 3). The validity of this model relies strictly on how well it performs with data never seen before that represents the intended application for the model. This is why an entire batch is left out instead of choosing a random set of observations (Cawley and Talbot, 2010; Politis and Romano, 2003).

Thus, model W is tested using the left-out data (i.e., batch C, see Fig. 1 box 3) and computing two performance criteria: the Root Mean Square Error of Prediction (RMSEP) and a coefficient of determination of prediction ($R^2$), for which the formulas can be found in supplementary material. In this work it was considered that models with values of $R^2$ superior to 0.75 have minimal satisfactory performance; in other words, a minimal performing model should have the capacity to explain at least 75 % of the variance in the data.

#### 2.7.3. Third step: robustness analysis of the architecture

Finally, a robustness analysis of the learning architecture is performed, which was inspired in the concept of nested cross-validation (Cawley and Talbot, 2010; Varma and Simon, 2006). In the absence of further observations / batches for validation of model W, the learning

**Table 2**
Datasets used for applying PLSR modelling; the table presents their description and the transformation required for any element within the matrices of inputs (I) outputs (O); $\lambda\lambda_j^i$ and $o_j^i$ represent respectively the element of matrices I and O in line i and column j.

| Dataset | Description | Elements of $\boldsymbol{I}$ | Elements of $\boldsymbol{O}$ |
|---|---|---|---|
| Ori | Original | $\lambda\lambda_j^i$ | $o_p^i$ |
| Log | Element-wise logarithm application | $\log(\lambda\lambda_j^i)$, 1 if $\lambda\lambda_j^i \leq 1$ | $\log(o_p^i)$ |

**Table 3**
Strategies used for variable selection, their specific methods and correspondent basis, selection condition and threshold range; $b_j$ represents the regression coefficient attributed to fluorescent variable j obtained by PLSR, $VIP_j$ represents the variable importance to projection attributed to fluorescent variable j, and n represents the initial number of fluorescence variables.

| Method | Basis | Selection condition | Threshold range |
|---|---|---|---|
| VIP | Model-wise | $VIP_j > c$ | $c \in [\min(VIP_j), \max(VIP_j)]$ |
| MWPLSR | EEM theory | EEM with side $= c$ | $c \in \begin{bmatrix} 8, & 12, & 16 \\ 20, & 24, & 28 \end{bmatrix}$ |
| Top MWPLSR | Hybrid | Top 3 windows from MWPLSR | |

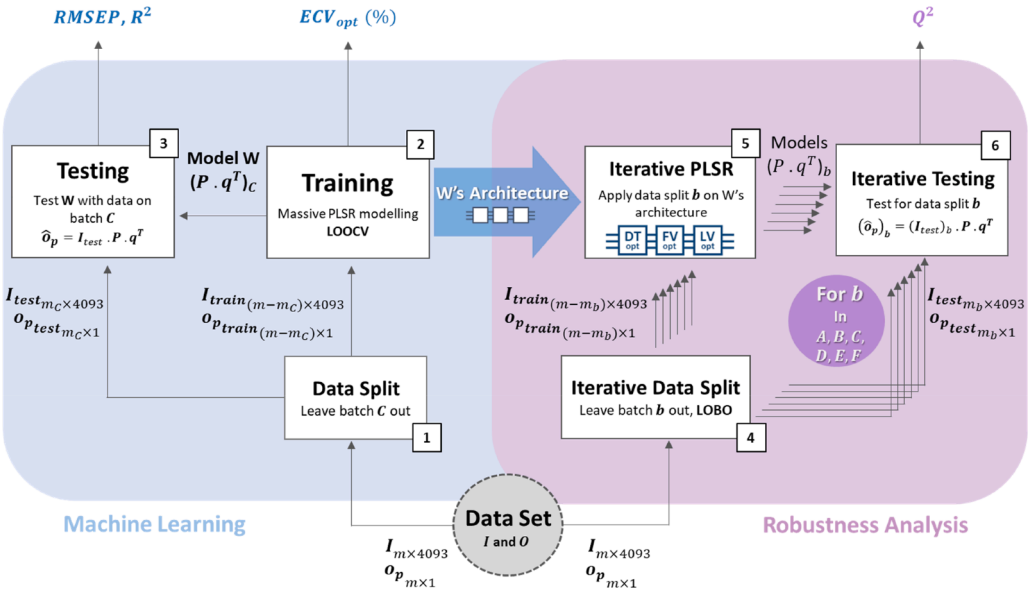


**Fig. 1.** Scheme of the modelling methodology developed in this work.

architecture is used to compute and validate models under different training/validation divisions i.e., for all 6 combinations of dataset division by leaving one batch out (LOBO, see Fig. 1 boxes 4, 5 and 6). In other words, a LOBO cross-validation is performed, and it yields the coefficient of determination for LOBO cross-validation ($Q^2$). Just as for RMSECV, $Q^2$ does not quantify the performance of a model (since it averages the performance of different models, although with the same learning architecture), but rather it provides an estimation of reliability of the modelling procedure. Formula and details are available in the supplementary material.

In this work it is assumed that a value of $Q^2$ greater than 0.5 implies satisfactory robustness, as done also by other works (Peng and Lai, 2012; Sartorius Stedim Data Analytics, 2017; Triba et al., 2015).

This analysis provides information to either support or refute the performance estimated by $R^2$, since it tests how robust is the model building by the learning architecture. It gives an idea on the likelihood of dataset division overfitting (Cawley and Talbot, 2010), i.e. the likelihood of the learning architecture yielding good performing models only if certain observations are left in/out from the validation set. Therefore, the robustness analysis result is prioritized over $R^2$ because, when a very low $Q^2$ is verified, no matter how good the value of $R^2$ is, the model should not be reliable as it probably resulted from a learning architecture that overfitted the dataset.

### 2.8. Random data modelling

The same methodology described in the sections above was applied to a dataset consisting of the fluorescent data matrix $\boldsymbol{I}$, and, instead of the biological parameters, a set of 22 random normally distributed variables. In the same logic stated in Section 2.7 this resulted in 616 random data models. This random parameter modelling is useful to assess whether the modelled information is related to the biological parameters or to noise (Ferreira et al., 2005), serving as a support for the significance of the robustness of the models obtained using the original data. It is expected that the values of $Q^2$ of the random data modelling should be significantly lower than the ones obtained with modelling the original data.

### 2.9. In silico implementation

The implementation of the outlier detection and modelling procedure, along with EEM restriction and the PLSR algorithm, were all performed by scripts developed in house using GNU Octave software. The algorithm SIMPLS (De Jong, 1993) was imported from the package 'statistics' of GNU Octave; the data was retrieved and written from and to Microsoft Excel spreadsheets using package 'io' of GNU Octave.

## 3. Results and discussion

### 3.1. PLSR modelling without variable selection

The 1st set of models was generated by PLSR modelling without fluorescent variable selection, thus the learning architecture does not include the optimal fluorescence variables selection, $FV_{opt}$. The results for the learning architecture and model W of each parameter are presented in Table 4.

According to these results, only Cell Count and the ratio Chlorophylls/Carotenoids can be modelled with a robust learning architecture (i.e., $Q^2 \geq 0.5$). In Fig. 2 is possible to observe that most predictions during robustness analysis do not exceed one standard deviation from equality, confirming the mentioned robustness. The model W of Cell Count predicts batch C with $R^2 = 0.94$, equivalent to a RMSEP of 1.51 M cells/mL, while the model W of Carotenoids/Chlorophylls achieved only $R^2 = 0.69$ with a RMSEP of 1.26. In Fig. 2 is possible to observe that the data points representing the predictions of model W follow the experimental data closely, confirming satisfactory predicting ability of the models for both biological parameters. In the case of Carotenoids/Chlorophylls, the low $R^2$ is mainly due to the underestimation of two data points as Fig. 2 shows.

Parameters Chlorophyll *b* and Carotenoids stand out from the others having non-negligible values of $Q^2$ ($\geq 0.37$). The accuracy plots of the robustness analysis for both these parameters (Fig. 2) show consistent overestimations in predicting batch E (green circles) and underestimations of the predictions for batch D (blue crosses) for both, being these apparently the main reasons for the low $Q^2$. These results suggest that what is being learned/modelled from the data is not random, and so the gathering of more data would probably result in acceptable learning architectures and model W.

Overall, the great majority of the parameters presented low values of $Q^2$, questioning the reliability of their corresponding models. These results may be due to one of these hypotheses:

**Table 4**
PLSR models obtained without variable selection; each model provides results on the Learning Architecture Tuned, its Robustness and the performance of model W.

| Output Information | | | Learning Architecture Tuned | | | Learning Architecture Robustness | Model W | |
|---|---|---|---|---|---|---|---|---|
| P | Biological Parameter | p units | $DT_{opt}$ | LVopt | $ECV_{opt}$ (%) | $Q^2$ | RMSEP (p units) | $R^2$ |
| 1 | Cell Count | $10^6$ cells/L | Log | 6 | 7.4 | 0.78 | 1.51 | 0.94 |
| 2 | Chlorophyll *b* | mg/L | Ori | 8 | 14.6 | 0.37 | 1.07 | 0.72 |
| 3 | Chlorophyll a | mg/L | Log | 5 | 28.0 | 0.18 | 0.43 | 0.82 |
| 4 | Carotenoids | mg/L | Ori | 7 | 16.2 | 0.44 | 5.73 | 0.78 |
| 5 | Car/Chl | – | Ori | 6 | 15.3 | 0.74 | 1.26 | 0.69 |
| 6 | Protein | mg/L | Log | 8 | 61.5 | −0.35 | 15.12 | 0.50 |
| 7 | Lutein | mg/L | Ori | 1 | 93.4 | 0.05 | 0.52 | −0.85 |
| 8 | Zeaxanthin | mg/L | Ori | 3 | 102.8 | −0.47 | 0.17 | −0.24 |
| 9 | α-carotene | mg/L | Ori | 3 | 63.8 | 0.08 | 0.23 | 0.62 |
| 10 | β-carotene | mg/L | Log | 7 | 62.1 | −0.02 | 2.84 | 0.56 |
| 11 | 9-cis-β-carotene | mg/L | Ori | 1 | 66.9 | 0.23 | 3.14 | 0.08 |

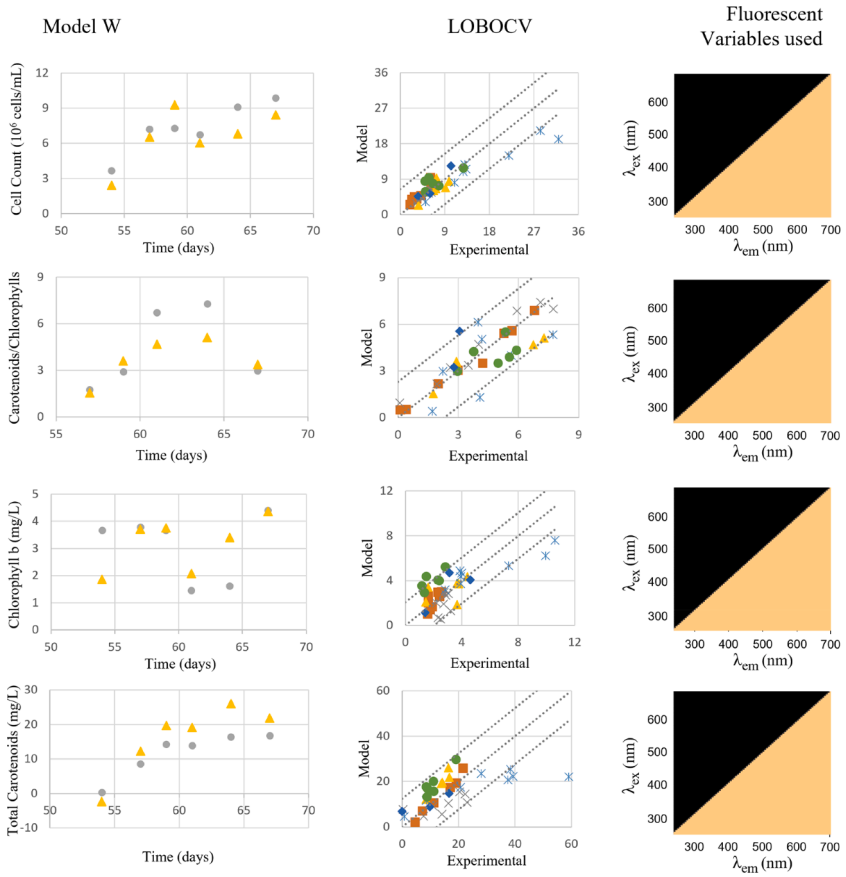


**Fig. 2.** Accuracy plots (i.e., model data vs experimental data) for the learning and testing of model W and for the robustness analysis of the learning architecture for parameters Cell Count, Carotenoids/Chlorophylls, Chlorophyll *b* and Carotenoids for PLSR without variable selection; the plots regarding model W show how far from equality were the predictions of model W of the left out batch C (yellow triangles); the plots regarding the robustness analysis show how consistent is the model building by keeping learning architecture constant varying the left-out batch for testing (A - orange squares; B - violet crosses; C - yellow triangles; D - blue crosses; E - green circles; F – blue diamonds).

- Hypothesis 1: There is a data mismatch problem, meaning that there are batches whose information comes from a different distribution, implying that different models will be learned depending on the training set used (Ng et al., 2022), causing low values of $Q^2$. This can be due to the cultivations not being exact replicates of each other, leaving room for consistent differences between each other that are reflected on the EEMs (e.g., different spectroscopic matrix effects or different intramolecular deactivation/quenching due to slightly different culture media).
- Hypothesis 2: There are fluorescence variables in the EEM that, not only do not contribute with relevant information, but also their information interferes with the execution of PLSR algorithm, namely in the projection of the latent structures (Forina et al., 2004).
- Hypothesis 3: Both above; hypothesis 1 and 2 may be simultaneously verified, meaning that the fluorescence variables whose information is interfering may be exactly the ones which provoke a non-evident batch-to-batch information heterogeneity; a way for verifying this is to perform fluorescence variable selection (see next section).

### 3.2. PLSR with variable selection

The 2nd set of models was generated by combining PLSR with either a model-wise variable selection method, namely variable importance to

projection (VIP), or by moving window PLSR hereby noted as VIP-wise and MWPLSR selection. The results for the learning architecture (now including the number of fluorescence variables selected, $FV_{opt}$) and the best result for the model W of each parameter is presented in Table 5.

According to these results (Table 5), variable selection resulted in learning architectures with higher robustness ($Q^2$) for most of the biological parameters when compared to no variable selection (Table 4). By incorporating this selection method, not only Cell Count and Carotenoids/Chlorophylls, but also Chlorophyll *b*, Chlorophyll a, Carotenoids, and β-carotene now have learning architectures with $Q^2 > 0.5$. The performance of the model W for these parameters is satisfactory ($R^2 > 0.70$) except for β-carotene ($R^2 = 0.54$). In Fig. 3 it is possible to see that the predictions of model W for these five parameters follow the experimental data closely, and that most robustness analysis predictions do not differ more than one standard deviation from the experimental data. These results corroborate Hypothesis 2, meaning that there are fluorescence variables in the EEM that interfere with the execution of PLSR algorithm, and give some support to Hypothesis 3. Thus, it seems that the fluorescence variables removed were responsible for the batch-to-batch heterogeneity. Moreover, Fig. 3 locates in the EEM the fluorescence variables selected for the learning architecture.

Interestingly, regarding variable selection, the results show cases where the excitation-emission wavelengths selected by the modelling procedure for estimating a biological parameter are different than their autofluorescence excitation-emission wavelengths (Fig. 3). For deriving Chlorophyll a and b, the fluorescent variables selected are within the excitation region between 300 and 400 nm and emission region between 475 and 625 nm, which is not expected to correlate with chlorophylls, since their fluorescent emission exists only from 680 nm (Maxwell and Johnson, 2000). Parameter Carotenoids was modelled using 2 windows in the EEM, being one of them within 480 and 570 nm of excitation. Generally, carotenoids absorption is between 440 and 540 nm. Additionally, the Rayleigh scattering lines, which are commonly deemed as interferant in the EEM, seem to provide relevant information to the point of being singularly used for prediction in the case of Carotenoids/-Chlorophylls ratio. A possible explanation for this can be the different interaction with light that chlorophylls and carotenoids have with it: chlorophylls tend to conduct radiation while carotenoids ten to quench it, so it would make sense that the ratio between the two would impact light scattering. Overall, these results suggest that following areas within EEM directly associated with the fluorescence of the compounds of monitoring interest may not provide complete information; other areas must be considered. This is probably due to the effects of physicochemical properties and of the culture broth that can result in interferences such as quenching and inner filter effect, which affect fluorescence profiles in a non-linear manner.

### 3.3. Biological parameters with no robust model and previous work

From the 11 biological parameters, 5 could not be modelled with minimal acceptable robustness according to the criteria stablished (i.e., $Q^2 \geq 0.5$); namely Proteins, Lutein, Zeaxanthin, α-carotene, and 9-cis-β-carotene. However, Protein and α-carotene stand out from having $Q^2 = 0.37$ and 0.41, respectively, although both have model W with $R^2 < 0.75$. The fact that, the great majority of the models obtained using the random data did not overcome $Q^2 = 0.25$, suggests that there is a possibility of obtaining robust models for these parameters if more observations or variables are included, e.g., just as it was done in the previous work by adding climatic data (Sá et al., 2020b). An explanation for the lower robustness may be that the data mismatch problem affects more these parameters (Hypothesis 1) than the successfully modelled ones. Indeed, small variations could not be avoided amongst the batches due mainly to seasonal variation. The previous work accounted for this seasonality by including climatic data, which may explain why some of the left unmodelled biological parameters in this work were considered successfully modelled in the previous one. The evaluation of performance was done differently in the previous work: it resorted to Pearson's coefficient of determination for linear correlation, rather than the coefficient of determination used in this work (equation S.5 in supplementary material). For comparison, coefficients of determination were re-calculated using the same equation. Table 6 shows the Pearson's coefficients of determination for linear correlation obtained from the previous and the present works, and indeed, parameters Cell Count, Chlorophyll a, α-carotene and β-carotene present models with higher accuracy, although no robustness analysis was performed.

## 4. Conclusions

The combination of the PLSR algorithm with variable selection and robustness analysis was shown to be a valuable methodology when trying to derive biological parameters of *Dunaliella salina* cultures from 2D fluorescence spectroscopy. Variable selection not only provided the detection of elements in the excitation-emission matrix (EEM) that negatively impact model robustness and performance when included, but also showed that fluorophore-associated areas in EEM may not be the best way to select wavelengths for monitoring specific biological parameters; the present modelling approach is more reliable. The robustness analysis revealed that models with high-performance during leave-one-out cross-validation and during testing with a left-out batch may still exhibit overfitting and, thus, is a process that should be always included in challenges such as the one in this work. Overall, the methodology developed in this work yielded robust and parsimonious models ready for use for monitoring of Cell Count, Chlorophyll *b*, Chlorophyll a, total Carotenoids, β-carotene concentrations, and Carotenoids/Chlorophylls ratio in *Dunaliella salina* cultures. The models require only up to 22 % of the fluorescent variables present in a 250–700 nm wavelength (with 5 nm step) range of excitation-emission.

**Table 5**
PLSR models obtained using variable selection; each model provides results on the Learning Architecture Tuned, its Robustness and the performance of model W.

| Output Information | | | Learning Architecture Tuned | | | | Learning Architecture Robustness | Model W | |
|---|---|---|---|---|---|---|---|---|---|
| p | Bioparameter | p units | $DT_{opt}$ | $FV_{opt}$ | $LV_{opt}$ | $ECV_{opt}$ (%) | $Q^2$ | RMSEP (p units) | $R^2$ |
| 1 | Cell Count | $10^6$ cells/L | Log-VIP | 888 | 7 | 4.1 | 0.80 | 1.32 | 0.93 |
| 2 | Chlorophyll *b* | mg/L | Log-VIP | 542 | 5 | 3.9 | 0.51 | 1.12 | 0.70 |
| 3 | Chlorophyll a | mg/L | Log-VIP | 231 | 5 | 8.2 | 0.80 | 0.23 | 0.95 |
| 4 | Carotenoids | mg/L | Log-TopMW16 | 663 | 6 | 10.0 | 0.87 | 2.14 | 0.97 |
| 5 | Car/Chl | – | Ori-VIP | 122 | 6 | 5.1 | 0.75 | 1.22 | 0.72 |
| 6 | Protein | mg/L | Log-VIP | 131 | 6 | 23.3 | 0.38 | 16.39 | 0.41 |
| 7 | Lutein | mg/L | Ori-VIP | 210 | 8 | 28.6 | −0.17 | 0.55 | −1.07 |
| 8 | Zeaxanthin | mg/L | Log-VIP | 126 | 8 | 33.5 | −0.20 | 0.23 | −0.76 |
| 9 | α-carotene | mg/L | Ori-MW12 | 144 | 8 | 46.6 | 0.41 | 0.25 | 0.57 |
| 10 | β-carotene | mg/L | Log-VIP | 121 | 8 | 16.7 | 0.69 | 2.91 | 0.54 |
| 11 | 9-cis-β-carotene | mg/L | Ori-MW8 | 64 | 8 | 36.5 | −0.21 | 3.58 | −0.20 |

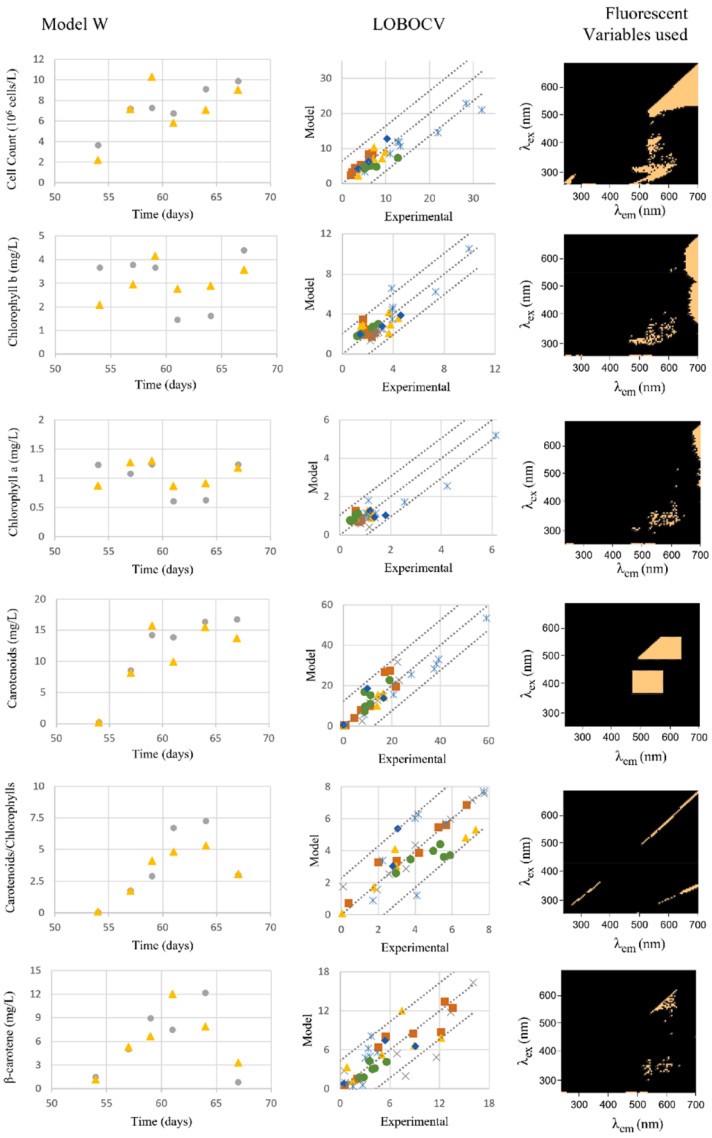


**Fig. 3.** Accuracy plots (i.e., model data vs experimental data) for the learning and testing of model W, and for the robustness analysis of the learning architecture for parameters Cell Count, Chlorophyll *b*, Chlorophyll a, Carotenoids and Carotenoids/Chlorophylls for PLSR coupled with variable selection; the plots regarding model W show how far from equality were the predictions of model W of the left out batch C (yellow triangles); the plots regarding the robustness analysis show how consistent is the model building by keeping learning architecture constant and varying the left-out batch for testing (A - orange squares; B - violet crosses; C - yellow triangles; D - blue crosses; E - green circles; F - blue diamonds).

**Table 6**
Coefficients of determination for linear correlation for the validation ($R^2_v$) obtained from previous work and from the present work; the models from the earlier were computed using 2D fluorescence and climatic data and without fluorescence variable selection.

| Model identification | | Best $R^2_v$ | |
|---|---|---|---|
| p | Parameter | Previous work | Present work |
| 1 | Cell Count ($10^6$ cells/L) | 0.97 | 0.76 |
| 2 | Chlorophyll *b* (mg/L) | 0.85 | 0.90 |
| 3 | Chlorophyll a (mg/L) | 0.75 | 0.67 (not robust) |
| 4 | Carotenoids (mg/L) | 0.79 | 0.91 |
| 8 | Zeaxanthin (mg/L) | 0.69 | 0.97 (not robust) |
| 9 | α-carotene (mg/L) | 0.63 | 0.40 (not robust) |
| 10 | β-carotene (mg/L) | 0.79 | 0.63 |
| 11 | 9-cis-β-carotene (mg/L) | 0.73 | 0.75 (not robust) |

## Funding

This research was supported by the Associate Laboratory for Green Chemistry (LAQV) which is financed by national funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020). This project was also funded by Fundação para a Ciência e Tecnologia/Ministério da Educação e Ciência, Portugal (FCT/MCTES) for the PhD Study grant 2021.07927.BD and program DL 57/2016 – Norma Transitória (SFRH/BPD/95864/2013). This project has received funding from the Bio Based Industries Joint Undertaking (JU) under grant agreement No. 512 887227 - MULTI-STR3AM. The JU receives support from the European Union's Horizon 2020 research and innovation programme and the Bio Based Industries Consortium.

## CRediT authorship contribution statement

**Pedro R. Brandão:** Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Marta Sá:** Investigation,

Validation, Writing – review & editing. **Claudia F. Galinha:** Conceptualization, Supervision, 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.

## Data availability

Data will be made available on request.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2023.108452.

## References


Amigo, J.M., Marini, F., 2013. Multi way methods. Chemometrics in Food Chemistry, , 1st ed.28. Elsevier.

Balabin, R.M., Smirnov, S.V., 2011. Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. Anal. Chim. Acta 692, 63–72. https://doi.org/10.1016/j.aca.2011.03.006.

Barkia, I., Saari, N., Manning, S.R., 2019. Microalgae for high-value products towards human health and nutrition. Mar. Drugs 17, 1–29. https://doi.org/10.3390/md17050304.

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, 26–35. https://doi.org/10.1002/elsc.201900076.

Berglund, A., Wold, S., 2007. INLR (Implicit Non-linear Latent Variable Regression). II. Blockscaling of expanded terms with QSAR examples. Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry, 11. Wiley, pp. 65–79. https://doi.org/10.1002/9783906390406.ch4.

Bernardo, J.M., Smith, A.F.M., 1994. Bayesian theory. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA https://doi.org/10.1002/9780470316870.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B 26, 211–252.

Cawley, G.C., Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107.

Chew, K.W., Yap, J.Y., Show, P.L., Suan, N.H., Juan, J.C., Ling, T.C., Lee, D.J., Chang, J. S., 2017. Microalgae biorefinery: high value products perspectives. Bioresour. Technol. 229, 53–62. https://doi.org/10.1016/j.biortech.2017.01.006.

Cuellar-Bermudez, S.P., Aguilar-Hernandez, I., Cardenas-Chavez, D.L., Ornelas-Soto, N., Romero-Ogawa, M.A., Parra-Saldivar, R., 2015. Extraction and purification of high-value metabolites from microalgae: essential lipids, astaxanthin and phycobiliproteins. Microb. Biotechnol. 8, 190–209. https://doi.org/10.1111/1751-7915.12167.

Efron, B., Gong, G., 1983. A leisurely look at the bootstrap, the jackknife, and a leisurely look at the bootstrap, the Jackknife, and cross-validation. Am. Stat. 37, 36–48.

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. https://doi.org/10.1002/bit.20526. WileyInterScience.

Forina, M., Lanteri, S., Armanino, C., 1987. Chemometrics in food chemistry. Chemometrics and Species Identification. Springer, Berlin, Heidelberg, pp. 91–143. https://doi.org/10.1007/3-540-17308-0_4.

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

Galinha, C.F., Carvalho, G., Portugal, C.A.M., Guglielmi, G., Oliveira, R., Crespo, J.G., Reis, M.A.M., 2011a. Real-time monitoring of membrane bioreactors with 2D-fluorescence data and statistically based models. Water Sci. Technol. 63, 1381–1388. https://doi.org/10.2166/wst.2011.195.

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, 3623–3636. https://doi.org/10.1016/j.watres.2012.04.010.

Galinha, C.F., Carvalho, G., Portugal, C.A.M., Guglielmi, G., Reis, M.A.M., Crespo, J.G., 2011b. Two-dimensional fluorescence as a fingerprinting tool for monitoring wastewater treatment systems. J. Chem. Technol. Biotechnol. 86, 985–992. https://doi.org/10.1002/jctb.2613.

Glindkamp, A., Riechers, D., Rehbock, C., Hitzmann, B., Scheper, T., Reardon, K.F., 2009. Sensors in disposable bioreactors status and trends. Advances in Biochemical Engineering/Biotechnology. Springer, Berlin, Heidelberg. https://doi.org/10.1007/10_2009_10.

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

Graf, A., Claßen, J., Solle, D., Hitzmann, B., Rebner, K., Hoehse, M., 2019. A novel LED-based 2D-fluorescence spectroscopy system for in-line monitoring of Chinese hamster ovary cell cultivations – Part I. Eng. Life Sci. 19, 352–362. https://doi.org/10.1002/elsc.201800149.

Hamed, I., 2016. The evolution and versatility of microalgal biotechnology: a review. Compr. Rev. Food Sci. Food Saf. 15, 1104–1123. https://doi.org/10.1111/1541-4337.12227.

Jiao, L., Bing, S., Zhang, X., Li, H., 2016. Interval partial least squares and moving window partial least squares in determining the enantiomeric composition of tryptophan using UV–vis spectroscopy. J. Serb. Chem. Soc. 81, 209–218. https://doi.org/10.2298/JSC150227065J.

Khan, M.I., Shin, J.H., Kim, J.D., 2018. The promising future of microalgae: current status, challenges, and optimization of a sustainable and renewable industry for biofuels, feed, and other products. Microb. Cell Factories 17, 1–21. https://doi.org/10.1186/s12934-018-0879-x.

Lakowicz, J.R., 2006. Principles of Fluorescence Spectroscopy, 3rd ed. Springer.

Lazraq, A., Cléroux, R., Gauchi, J.P., 2003. Selecting both latent and explanatory variables in the PLS1 regression model. Chemom. Intell. Lab. Syst. 66, 117–126. https://doi.org/10.1016/S0169-7439(03)00027-3.

Lichtenthaler, H., Buschmann, C., 1987. Chlorophyll and carotenoid determination (after Lichtenthaler 1987), a practical instruction. Methods Enzymol. 8, 350–382.

Maxwell, K., Johnson, G.N., 2000. Chlorophyll fluorescence–a practical guide. J. Exp. Bot. 51, 659–668.

Ng, A., Katanforoosh, K., Mourri, Y., 2022. Addressing data mismatch - DeepLearning.AI [WWW Document]. Deep Learn. AI - Coursera. URL https://www.coursera.org/lecture/machine-learning-projects/www.deeplearning.ai-biLiy?utm_source=link&utm_medium=page_share&utm_content=vlp&utm_campaign=top_button (accessed 6.23.22).

Patel, A., Gami, B., Patel, P., Patel, B., 2017. Microalgae: antiquity to era of integrated technology. Renew. Sustain. Energy Rev. 71, 535–547. https://doi.org/10.1016/j.rser.2016.12.081.

Peng, D.X., Lai, F., 2012. Using partial least squares in operations management research: a practical guideline and summary of past research. J. Oper. Manag. 30, 467–480. https://doi.org/10.1016/j.jom.2012.06.002.

Podrazký, O., Kuncová, G., Krasowska, A., Sigler, K., 2003. Monitoring the growth and stress responses of yeast cells by two-dimensional fluorescence spectroscopy: first results. Folia Microbiol. (Praha) 48, 189–192. https://doi.org/10.1007/BF02930954.

Politis, D.N., Romano, J.P., 2003. The stationary bootstrap. J. Am. Stat. Assoc. 98, 585–588. https://doi.org/10.1198/016214503000000468.

Roth, S., 1978. Growth, nutrition, and metabolism of cells in culture. Volume III. George H. Rothblat, Vincent J. Cristofalo. Q. Rev. Biol. 53, 160–161. https://doi.org/10.1086/410492.

De Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. 18, 251–263.

Sá, M., Ferrer-Ledo, N., Wijffels, R., Crespo, J.G., Barbosa, M., Galinha, C.F., 2020a. Monitoring of eicosapentaenoic acid (EPA) production in the microalgae Nannochloropsis oceanica. Algal Res. 45 https://doi.org/10.1016/j.algal.2019.101766.

Sá, M., Monte, J., Brazinha, C., Galinha, C.F., Crespo, J.G., 2019. Fluorescence coupled with chemometrics for simultaneous monitoring of cell concentration, cell viability and medium nitrate during production of carotenoid-rich Dunaliella salina. Algal Res. 44, 101720 https://doi.org/10.1016/j.algal.2019.101720.

Sá, M., Monte, J., Brazinha, C., Galinha, C.F., Crespo, J.G., 2017. 2D Fluorescence spectroscopy for monitoring Dunaliella salina concentration and integrity during membrane harvesting. Algal Res. 24, 325–332. https://doi.org/10.1016/j.algal.2017.04.013.

Sá, M., Ramos, A., Monte, J., Brazinha, C., Galinha, C.F., Crespo, J.G., 2020b. Development of a monitoring tool based on fluorescence and climatic data for pigments profile estimation in Dunaliella salina. J. Appl. Phycol. 32, 363–373. https://doi.org/10.1007/s10811-019-01999-z.

Sartorius Stedim Data Analytics, 2017. Simca® 15 User Guide. Sartorius.

Tartakovsky, B., Sheintuch, M., Hilmer, J.M., Scheper, T., 1996. Application of scanning fluorometry for monitoring of a fermentation process. Biotechnol. Prog. 12, 126–131. https://doi.org/10.1021/bp950045h.

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, 255–260. https://doi.org/10.1016/j.jbiotec.2010.11.015.

Triba, M.N., B, L.L.M., Amathieu, R., Goossens, C., Bouchemal, N., Nahon, P., Rutledge, D.N., Savarin, P., 2015. PLS/OPLS models in metabolomics: impact of permutation of dataset rows on the K-fold cross-validation quality parameters. Mol. Biosyst. 1, 13–19.

Vanthoor-Koopmans, M., Wijffels, R.H., Barbosa, M.J., Eppink, M.H.M., 2013. Biorefinery of microalgae for food and fuel. Bioresour. Technol. 135, 142–149. https://doi.org/10.1016/j.biortech.2012.10.135.

Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinform. 7 https://doi.org/10.1186/1471-2105-7-91.