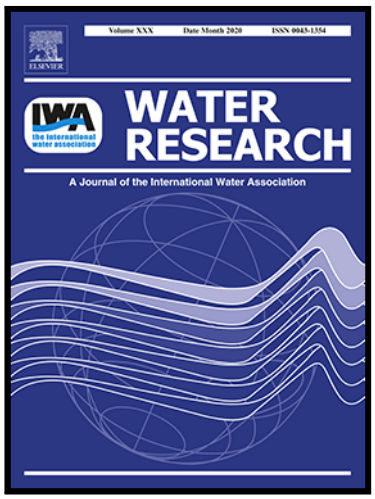
Fast and robust NIRS-based characterization of raw organic waste: Using non-linear methods to handle water effects

Alexandre Mallet, Cyrille Charnier, Éric Latrille, Ryad Bendoula, Jean-Michel Roger, Jean-Philippe Steyer

Journal Pre-proof



Fast and robust NIRS-based characterization of raw organic waste: using non-linear methods to handle water effects

Alexandre Mallet , Cyrille Charnier , Éric Latrille , Ryad Bendoula , Jean-Michel Roger , Jean-Philippe Steyer

PII: S0043-1354(22)01253-2
DOI: https://doi.org/10.1016/j.watres.2022.119308
Reference: WR 119308

To appear in: *Water Research*

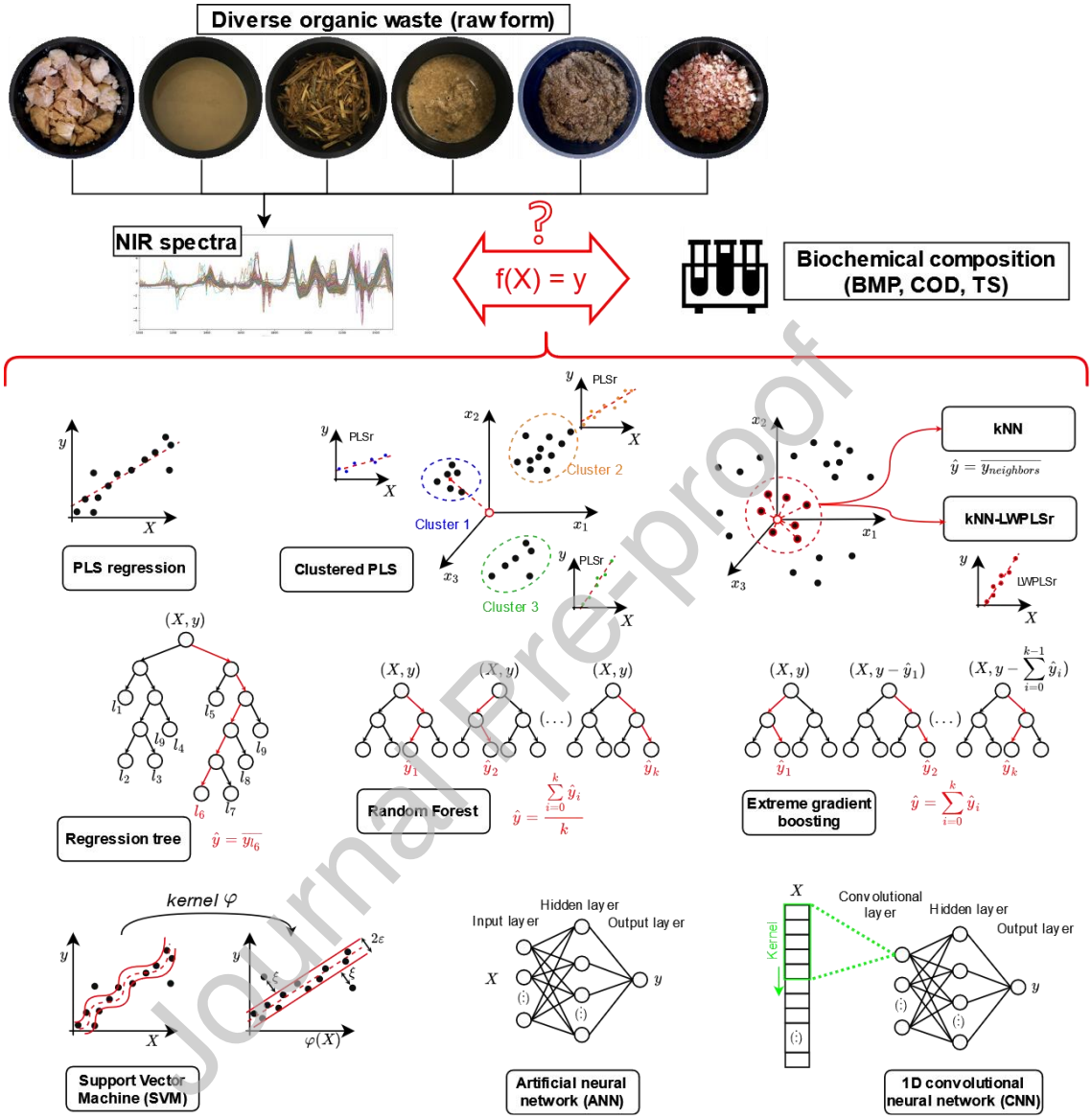
Received date: 23 August 2022
Revised date: 10 October 2022
Accepted date: 27 October 2022

Please cite this article as: Alexandre Mallet , Cyrille Charnier , Éric Latrille , Ryad Bendoula , Jean-Michel Roger , Jean-Philippe Steyer , Fast and robust NIRS-based characterization of raw organic waste: using non-linear methods to handle water effects, *Water Research* (2022), doi: https://doi.org/10.1016/j.watres.2022.119308

Highlights

- Multiple non-linear methods are evaluated for raw organic waste characterization
- Local linear models and non-linear models override water effects
- Calibrations on raw samples (no drying or grinding) were successfully built
- This opens the door to at-site NIRS-based characterization of raw organic waste

Graphical abstract



Graphical Abstract – A comprehensive comparison of non-linear calibration methods for NIRS-based characterization of diverse organic waste in raw form.

Fast and robust NIRS-based characterization of raw organic waste: using non-linear methods to handle water effects

Authors: Alexandre Mallet[a,b,c,d,*], Cyrille Charnier[c], Éric Latrille[a,d], Ryad Bendoula[b], Jean-Michel Roger[b,d], Jean-Philippe Steyer[a]

a) INRAE, LBE, Montpellier University, Narbonne, France *(Full postal address: 102 Avenue des Etangs, 11100, Narbonne, France)*

b) INRAE, ITAP, Montpellier University, Montpellier, France *(Full postal address: 361 rue Jean-François Breton, 34196, Montpellier, France)*

c) BioEnTech, Narbonne, France *(Full postal address: 102 Avenue des Etangs, 11100, Narbonne, France)*

d) ChemHouse Research Group, Montpellier, France *(Full postal address: 361 rue Jean-François Breton, 34196, Montpellier, France)*

* the corresponding author, E-mail: alexandre.mallet@bioentech.eu, Telephone: +33 6 26 42 03 45, Full postal address: 10 cité Chabert, 26000 Valence, FRANCE

Abstract

Fast characterization of organic waste using near infrared spectroscopy (NIRS) has been successfully developed in the last decade. However, up to now, an on-site use of this technology has been hindered by necessary sample preparation steps (freeze-drying and grinding) to avoid important water effects on NIRS. Recent research studies have shown that these effects are highly non-linear and relate both to the biochemical and physical properties of samples. To account for these complex effects, the current study compares the use of many different types of non-linear methods such as partial least squares regression (PLSR) based methods (global, clustered and local versions of PLSR), machine learning methods (support vector machines, regression trees and ensemble methods) and deep learning methods (artificial and convolutional neural networks). On an independent test data set, non-linear methods showed errors 28% lower than linear methods. The standard errors of prediction obtained for the prediction of total solids content (TS%), chemical oxygen demand (COD) and biochemical methane potential (BMP) were respectively 8%, 160 mg(O_2).gTS^{-1} and 92 mL(CH_4).gTS^{-1}. These latter errors are similar to successful NIRS applications developed on freeze-dried samples. These findings hold great promises regarding the development of at-site and online NIRS solutions in anaerobic digestion plants.

Keywords

Near infrared spectroscopy; anaerobic digestion; biochemical methane potential; water effects; non-linear modeling, neural network.

1. Introduction

In bioprocesses such as composting, anaerobic digestion (AD) or pyro-gasification, the ability to effectively characterize the input organic waste is a necessary condition for optimizing the process (Jacobi et al., 2011; Jimenez et al., 2015). Unfortunately, the input feedstock may cover a tremendously wide range of biochemical and physical properties, thus making the development of fast and robust analytical procedures a challenging issue. In the last decade, near infrared spectroscopy (NIRS), in combination with sound multivariate statistical calibrations, has emerged as the most reliable and fast solution for characterizing organic materials. Amongst the developed applications on organic waste, the technology has been used to monitor the maturity of compost (Albrecht et al., 2008; Vergnoux et al., 2009), to assess the biochemical methane potential (BMP) and biodegradability (Doublet et al., 2013; Fitamo et al., 2017; Godin et al., 2015; Lesteur et al., 2011; Mortreuil et al., 2018; Triolo et al., 2014; Yang et al., 2021), but also to predict important variables such as carbohydrates content, nitrogen content and chemical oxygen demand (COD) (Charnier et al., 2017a), cellulose/hemicellulose (Liu et al., 2021), or hydrolysis kinetics (Charnier et al., 2017b). Thanks to these developments, a full description of an organic waste can now be provided by NIRS in less than three days (instead of one to two months for a characteristic like BMP). However, the analytical process still involves cumbersome sample preparation steps (freeze-drying and grinding) to avoid the effects of water and particle size on NIRS, which explains the limited adoption of NIRS for on-line or at-site industrial applications.

Water effects on NIRS were shown to be highly non-linear and resulting from complex interacting physical and chemical effects (Mallet et al., 2021a). These physical effects were shown to result from light path-length modifications directly related to moisture content by a power law (Mallet et al., 2021c). This brings keys to better understand the low performance obtained with linear model-based calibrations made on organic waste with diverse biochemical compositions and humidity levels. While traditional methods such as PLSR have the advantage of simplicity, interpretability and robustness, these methods are not able to cope with non-linear water effects as they rely on the assumption that a linear relationship exists between the predicted characteristic and the spectrum. On the other hand, non-linear methods can consider more complex non-linear relationships and thus may provide usable models (Ni et al., 2014; Pérez-Marín et al., 2007). Different methods can be classified in three categories: PLSR-based methods, machine learning methods and deep learning methods.

Within chemometrics, the founding block of non-linear techniques remains PLSR (Wold et al., 2001). Indeed, clustered and local approaches of PLSR have been developed, where models are built based only on a subset of the dataset. This subset can be chosen based on spectral characteristics or using expert knowledge and metadata. In local methods based on spectra, there is the clustered PLSR approach where a clustering method (k-means (Preda and Saporta, 2005), hierarchical clustering (Tøndel et al., 2011), decision tree (Narayanan et al., 2019)) is used to identify groups on which to train PLSR models. When predicting a new observation, it is assigned to its cluster based on a spectral distance (Euclidean or Mahalanobis (Shen et al., 2019)) or a

trained classification method, and the corresponding cluster's model is used for prediction. In strictly speaking local approaches, such as the local PLSR (Shenk et al., 1997) or the k-nearest neighbors locally weighted PLSR (kNN-LW-PLSR) (Lesnoff et al., 2020), the subset is selected on-the-fly (*i.e.*, just in time for each new observation) based on a spectral distance and the model based on this subset is also built on-the-fly. This type of local approach has been successfully applied for the NIRS-based prediction of BMP in plant biomasses (Godin et al., 2015).

Within machine learning methods, a first group of methods are based on the regression tree concept with a binary recursive partitioning where models are built within each partition (usually a simple average (Holmes et al., 1999), but it can be a PLSR(Eriksson et al., 2009)). The power of regression trees owes to its simplicity and interpretability. Based on this regression tree structure, ensemble methods have been proposed where forests of such tree models are built based on bagging (Random Forest) or boosting (XGBoost). Such models have been assessed and compared for NIRS-based prediction of soil composition (de Santana et al., 2018; Nawar and Mouazen, 2017) or biodiesel blends composition (Cunha et al., 2020).

A second group of methods concerns support vector machines (SVM) regression methods (Drucker et al., 1996), which is essentially a constrained version of linear regression where the L*2*-norm of the coefficient vector is minimized with constraints on the hyperplane (maximal margin). SVM regression is made non-linear using kernel functions (such as polynomial or gaussian radial basis functions) that transform the original data into a higher dimensional feature space to make it possible to perform the linear regres-

sion. SVM regression has been applied in combination with NIRS data (Belousov et al., 2002; Borin et al., 2006; Devos et al., 2009), and more recently for BMP prediction on algae substrates (Yao et al., 2020).

A last category of non-linear methods includes deep learning methods. Artificial neural network (ANN) has been applied to NIRS in the last twenty years (Berzaghi et al., 2002; Marini et al., 2008; Nørgaard et al., 2013), but the increase of dataset size as well as computational capabilities has recently made it a practical predictive tool. In particular, one-dimensional convolutional neural networks (1D-CNN) is an architecture that has been found the most suitable for NIRS data (Acquarelli et al., 2017; Cui and Fearn, 2018; Malek et al., 2018). One reason is that the 1D-convolution layer plays the role of spectral preprocessor (Cui and Fearn, 2018). The power of neural networks lies in its flexible and customizable architecture defined by the number of layers, the type of layers (dense, convolutional), and the layers' parameters (dimensionality, padding/stride, dropout rate, activation function) (Mishra and Passos, 2022). But the challenge of such models is also its complexity and the important number of parameters to tune which makes it prone to overfitting. To avoid this, the learning hyperparameters (mainly batch size, number of epochs, and learning rate) need to be soundly selected. Another challenge of deep learning methods concerns the low interpretability of the obtained models, though some authors propose some numerical tools to investigate the feature importances (Cui and Fearn, 2018).

In this study, the suitability of non-linear models is assessed comprehensively for diverse raw organic waste characterization. Can the higher complexity of machine learning and non-linear algorithms provide models usable directly on raw organic materials? To answer this, non-linear methods from the three communities (chemometrics, machine learning and deep learning) are evaluated for the prediction of total solids (TS%), chemical oxygen demand (COD) and biochemical methane potential (BMP) on different types of organic waste and without any sample preparation (drying, grinding). The calibrated models are evaluated based on their prediction performance (root mean squared error (RMSE), median absolute error (MAD), coefficient of determination (R^2)), but also based on their robustness towards water effects (*i.e.*, moisture content variations).

2. Materials and Methods

2.1. Samples and reference analyses

The dataset consists of 501 different organic waste samples that have been collected in rural, territorial and industrial anaerobic digestion plants in France. These samples cover a very wide range of biochemical and physical properties: solid ligno-cellulosic materials (like silage, cereals, ramial wood chips, and corn cobs), liquid ligno-cellulosic suspensions (such as manure, pig slurry), liquid high-fat content suspensions (catering waste or biowaste), sweet emulsions (such as lactoserum or syrup), or protein and fat solid pastes (such as egg waste, cacao butter, or primary and secondary sludges from wastewater treatment plants). The visual aspect of some of these samples is presented in Appendix A.

Biochemical characterization of samples was obtained by using NIRS-based calibration models (Charnier et al., 2017a) applied to freeze-dried and ground samples. The errors of these models on independent test sets were evaluated at 128 $mg(O_2).gTS^{-1}$ for COD and 78 $mL(CH_4).gTS^{-1}$ for BMP. However, for samples which contained volatile molecules (e.g., volatile fatty acids, ammonia NH_3) that can disappear during the drying process (such as silage or biowaste), standard BMP and COD measurements were made as described in (Angelidaki et al., 2009; Charnier et al., 2017a). The total solids content was measured on all samples using the standard protocol (48 hours of oven-drying at 105°C) and the standard error of laboratory was evaluated at 5%. The histograms of obtained reference values are presented in Figure 1.

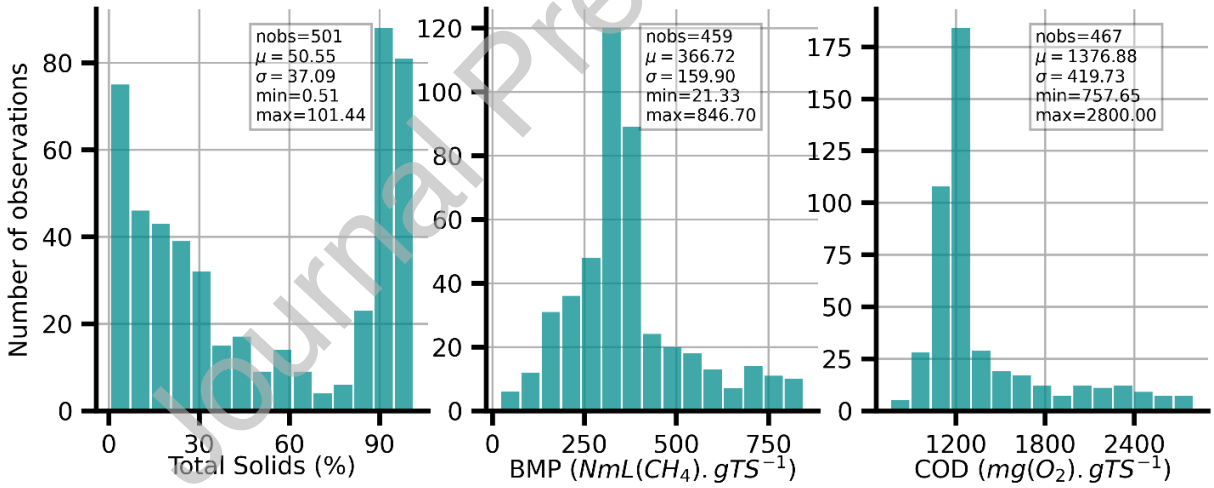


Figure 1 - Histograms of biochemical composition values (total solids - TS%, biochemical methane potential - BMP, chemical oxygen demand – COD). Respective number of observations (labeled as nobs), mean (labeled as μ), standard deviation (labeled as σ), and range (min/max) are provided.

Furthermore, in order to evaluate the robustness of developed NIRS models towards moisture content effects, a dataset of NIRS measurements acquired during N_2-drying experiments of various organic substrates was used as described in (Mallet et al., 2021a). The oven-drying was not used because of possible chemical modifications at

high temperatures (Maillard reactions), and freeze-drying was not used because it requires to freeze the sample and temperature strongly modifies the measured near infrared spectra. This dataset consists of 89 substrates of various biochemical and physical types, covering a wide range of moisture content levels (from 1% to 99%). The predictions made on these NIRS measurements are presented and discussed in Figure 7.

2.2. Spectroscopic system

Triplicate spectra were collected on the raw samples with an NIR-Flex N-500 solids FT-NIR spectrophotometer with a (10 cm diameter) petri dish accessory (Buchi, Flawil, Switzerland), scanning in reflectance mode with a spectral range of 4 000 cm^{-1} to 10 000 cm^{-1} (1000-2500 nm) and a resolution of 4 cm^{-1}. The cost of such instrument ranges between 50 000 € and 80 000 €.An external white reference (Spectralon®) signal $I_0(\lambda)$ is automatically taken every 10 minutes. For each sample, an intensity signal $I(\lambda)$ was collected during the rotation of the sample (average of 96 scans), and the pseudo-absorbance signal $A(\lambda)$ was computed:

$$A(\lambda) = -log_{10}\big(R(\lambda)\big) = -log_{10}\left(\frac{I(\lambda)}{I_0(\lambda)}\right). \quad \text{(Eq. 1)}$$

2.3. Model architectures

Ten different model types were evaluated:

1) A **simple PLSR**(referred here as "**plsr**") using the NIPALS algorithm (Næs and Martens, 1984; Wold, 1973) served as a control linear model, to which non-linear

methods were benchmarked. One model hyperparameter was considered for tuning: the number of latent variables (from 1 to 20).

2) A **k-nearest neighbor regression** (referred here as "**knnr**") which consisted for each sample to select k-nearest neighbors based on minimal Euclidean distance or Mahalanobis distance, and then take the average of the y values of this neighborhood as the predicted value. Two model hyperparameters were considered for tuning: the distance type (Euclidean or Mahalanobis) and the number of neighbors k (3, 5, 10, 50, 100).
3) A local PLSR method called the **k-nearest neighbors locally weighted PLS regression** (referred here as "**knnlwplsr**") (Lesnoff et al., 2020) consisted of the similar procedure as the knnr, but instead of the average, a locally weighted PLSR (Kim et al., 2011) was calculated for each neighborhood. The calculated PLSR was weighted using a normalized (sum to one) version of the Gaussian radial basis function φ applied to distances δ_i between the predicted observation and its neighbors:

$$\varphi(\delta_i) = \exp\left(-\frac{\delta_i}{h\sigma_\delta}\right), \qquad \text{(Eq. 2)}$$

with σ_δ the standard deviation of distances, and h the similarity index.
Simply said, this weighing ensures that the closest neighbors to the predicted observation will influence more the final model. Parameter h controls how much the closest neighbors will weigh more in the model (when h = ∞, this is equivalent to a simple PLSR, with identical weights given to all neighboring observations). Three hyperparameters were considered for tuning: the number of neighbors (10,

50, 100, 300), the similarity h parameter (10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}), and the number of latent variables (1 to 20).

4) A **clustered local approach method** (referred here as "**clusteredplsr**") where a clustering method is first applied (k-means, or hierarchical clustering analysis - HCA) and within each cluster, a simple PLSR is trained. Three model hyperparameters were considered for tuning: the clustering method (k-means or HCA), the number of clusters (2, 3, 4, 5), and the number of latent variables of clusters' PLSR model (1 to 20).
5) A **support vector machines (SVM) regression** (referred here as "**svmr**") was calculated using the radial basis function (RBF) kernel type. Two hyperparameters were considered for tuning: the kernel coefficient Γ (10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}) and the regularization parameter C (10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}).
6) A **decision tree regression** (referred here as "**rtree**") which involves a binary recursive partitioning, where the prediction for a given observation is based on a simple average of the fold (group of observations) in which it lies. Two hyperparameters were considered for tuning: the maximum depth of tree (5, 10, 30, 50, 100), the minimum number of observations in a fold (5, 10, 30).
7) An **ensemble random forest method** (referred here as "**rf**") which consists in growing multiple regression trees based on random sub-selections of features and observations, and then bagging all these tree models by averaging its predictions. Various hyperparameters were fixed: the minimum number of samples within each leaf (5), the number of features drawn for each tree (1501) (Geurts et

al., 2006), and the number of samples drawn for each tree (90% of total number of samples). Two hyperparameters were considered for tuning: the maximum depth of trees (5, 10, 20, 30, 50, 100), and the total number of trees (5, 20, 50, 100, 150).

8) An **ensemble extreme gradient boosting method** (referred here as "**xgb**") which consists in growing multiple regression trees based on random sub-selections of features and observations but based on a boosting principle where each new tree is grown to predict the residuals from the sum of predictions made by the existing trees. Three hyperparameters were considered for tuning: the number of trees (5, 20, 50, 100, 150), the maximum depth of trees (5, 10, 20, 30, 50, 100) and the learning rate (0.01, 0.05, 0.1, 0.5).
9) An **artificial neural network (ANN) model** (referred here as "**nn**") with a simple dense structure. A unique dense layer of 25 neurons was chosen with rectified linear unit (ReLU) activation functions which guarantee the non-linearity of the method. The architecture is presented in Figure 2. Three hyperparameters were considered for tuning: the learning rate (0.001, 0.005, 0.01), the number of epochs (100, 300, 500, 1000), and the batch size (5, 10, 30).
10) A **convolutional neural network (CNN) model** (referred here as "**cnn**") with Three hyperparameters were considered: the learning rate (0.001, 0.005, 0.01), the number of epochs (100, 300, 500, 1000), and the batch size (5, 10, 30).

For each of these methods, the impact of adding a prior dimension reduction step to the model pipeline was evaluated. Indeed, models were built not only on the original varia-

bles, but also on the scores of a global principal components analysis (PCA) or a global PLSR, respectively referred as "**pca**_" and "**pls**_". In the case of PCA, a singular value decomposition (SVD) is applied to the centered spectra X, leading to $X = U_k \Sigma_k V_k^T$, with k the number of components chosen, Σ the singular values, and U and V the left and right singular vectors with $V^T V = I$. The scores T_k are calculated as $T_k = XV_k$. The number of components of the PCA or PLSR was set to 25 components. This prior dimension reduction step allowed to reduce the computation time and to stabilize the tuning of non-linear methods which are prone to overfitting.

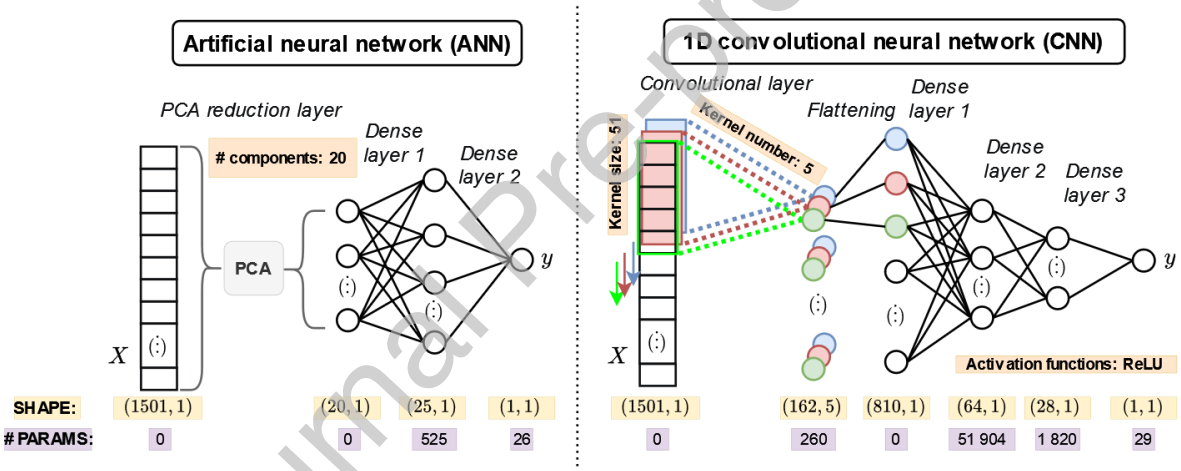


Figure 2 – Architecture of neural networks (models "pca_nn" and "cnn"). The shape of data at each step of the neural network is highlighted in yellow, the number of parameters/coefficients to be trained is highlighted in purple, and the other hyperparameters are highlighted in orange.

2.4. Data analysis and model calibration

All the data analysis (presented in Figure 3) was performed using Python 3.7.11: data wrangling with Pandas 1.3.4, NumPy 1.19.5, SciPy 1.7.1, Scikit-learn 1.0.1, Tensorflow

2.7.0, and plotting with Matplotlib 2.2.2 (Abadi et al., 2016; Hunter, 2007; McKinney, 2010; Oliphant, 2010; Pedregosa et al., 2015; van Rossum and Drake, 2009; Virtanen et al., 2020).

A variety of preprocessing methods can be applied on NIR spectra to reduce the random and systematic variations unrelated to the characteristic of interest (Roger et al., 2020). Here, only simple preprocessing methods were evaluated: a simple standard normal variate (Barnes et al., 1989) (referred as “snv”), the second-order detrend (Barnes et al., 1989) (referred as “dt2”), the first-order Savitzky-Golay (Savitzky and Golay, 1964) derivation (referred as “sg1”), the second-order Savitzky-Golay derivation (referred as “sg2”), and a simple combination of detrend and standard normal variate (referred as “dt2_snv”) to get rid of additive and multiplicative effects (Roger et al., 2020). The raw (absorbance) signal (referred as “a”) was used directly as well, which resulted overall in testing six different preprocessing conditions. The preprocessed spectra are presented in Appendix B.

To evaluate the built models, a validation test set was constituted. With the aim of producing a representative validation test set, the Duplex algorithm (Snee, 1977) was run for each reference characteristic (TS%, COD, BMP). Triplicates remained grouped together within the train or test sets.

For all methods, the tuning of hyperparameters was done using a repeated randomized grouped k-fold cross-validation with $k = 5$ the fold number and $n\_repeats = 30$ the repetition number. Sample triplicates were always kept within one fold to ensure independence between splits within cross-validation. For each cross-validation run, various met-

rics were then calculated: the root-mean-squared-error (RMSE), the mean absolute error (MAE) (Willmott and Matsuura, 2005), the coefficient of determination (R^2). The choice of the hyperparameters was made by analyzing all these metrics together (*i.e.,* choosing the set of hyperparameters minimizing RMSE and MAE, while maximizing R^2). The final performances of the obtained models were evaluated on the validation test set, based on various complementary statistics: the root-mean-squared error (RMSE), the mean absolute error (MAE), the median absolute deviation (MAD), the relative mean absolute error (RMAE), the squared Pearson correlation coefficient (r^2), the determination coefficient (R^2), the bias (b), and the standard error of prediction (SEP). The formulas are provided in Appendix C.

To evaluate the robustness of each model towards TS% variations (*i.e.* moisture content), the models were applied on the dataset of N_2-drying experiments (presented in section 2.1) (Mallet et al., 2021a). The total error of the models can be seen as the result of the average error per substrate (inter-substrate error), and the standard deviation of the errors within each substrate (intra-substrate error). For each substrate, the average of the absolute residuals was calculated, as well as the standard deviation of absolute residuals.

Let s be the substrate number ($s \in [1,89]$), and M_s the space of moisture content levels covered by this substrate. For each substrate number s and moisture content level k (with $k \in M_s$), an absolute residual can be calculated (as in equation 3). The inter-substrates and intra-substrates errors are calculated for each substrate and correspond respectively to the average and the standard deviation of these residuals for $k \in M_s$ (Eq.4 and Eq.5).

$$\mathrm{r_{s,k}} = \left|\widehat{y_{s,k}} - y_{s,k}\right| \quad \text{(Eq. 3)}$$

$$Absolute\ error_{inter-substrates}(\mathrm{s}) = \frac{\sum_{k \in M_s} \mathrm{r_{s,k}}}{card(M_s)}. \quad \text{(Eq. 4)}$$

$$Absolute\ error_{withinintra-substrates}(\mathrm{s}) = \sqrt{\frac{\sum_{k \in M_s} (\mathrm{r_{s,k}} - \overline{\mathrm{r_{s,k}}})}{card(M_s)}}. \quad \text{(Eq. 5)}$$

The inter-substrate error is related to how well the model predicts the biochemical composition, while the intra-substrate error is related to how much the model is sensitive to the varying TS%. The boxplots of these inter-substrate and intra-substrate errors are plotted in Figure 7.

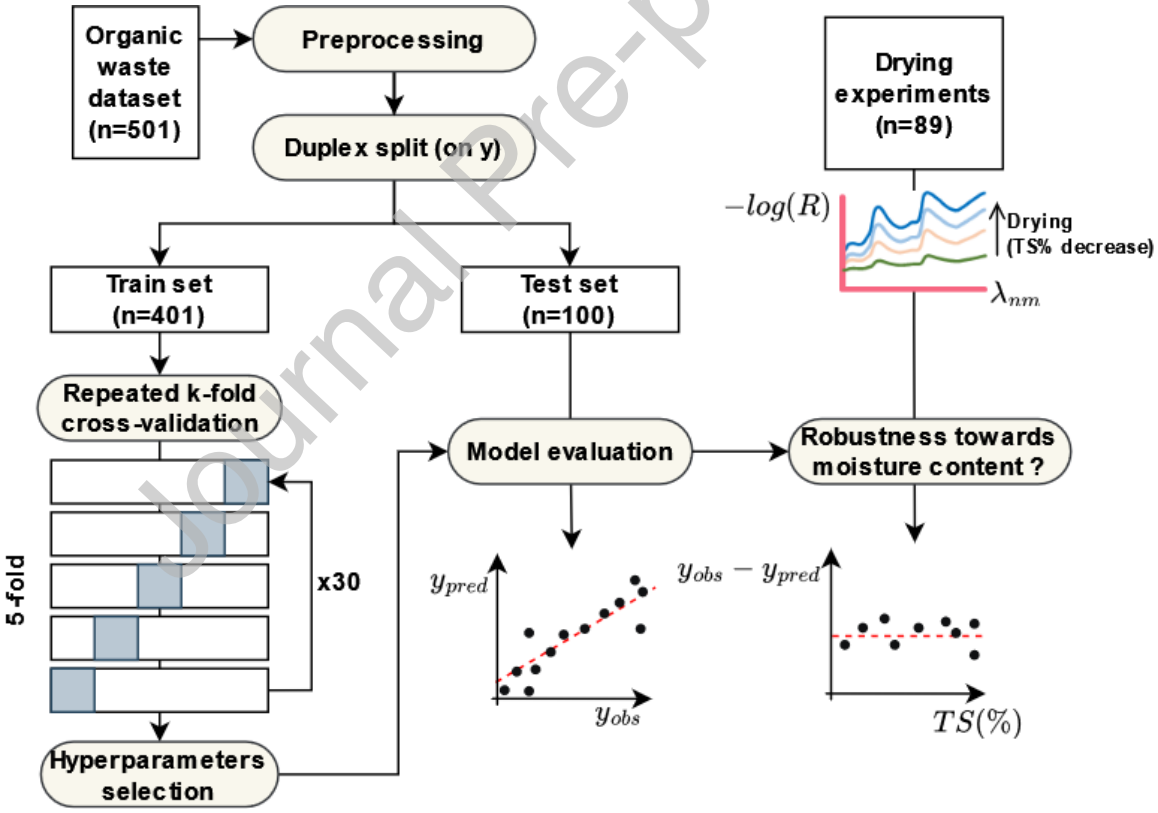


Figure 3 - Flow diagram of data analysis. The process is repeated for all model types (plsr, knnr, knnlwplsr, clusteredplsr, svmr, rtree, rf, xgb, nn, cnn).

3. Results & Discussion

3.1. Data overview

The distributions of the biochemical variables (TS%, COD, BMP) of the substrates are provided in Figure 4 with histograms of train and test sets respectively in blue and orange. The ranges are very wide, with TS% values between 0.5% and 100%, BMP values between 21 and 847 mL(CH_4).gTS^{-1} and COD values between 456 and 2 800 mg(O_2).gTS^{-1}. The distributions of train and test sets are similar, thanks to the Duplex algorithm applied to each reference variables (TS%, COD, BMP). While the distributions of BMP and COD appear to follow normal distributions, the distribution of TS% values appear to follow a bi-normal distribution with a group of substrates with TS% values below 30% and a group of substrates with TS% values above 80%. Indeed, this latter group of substrates corresponds to naturally low moisture content samples such as high-fat content substrates (*e.g.*, oil, slaughterhouse waste), or agro-industrial waste (*e.g.*, flour, sucrose, paper), as well as samples that were simply freeze-dried to extend the range of TS% on which the models can work (*i.e.*, increase robustness).

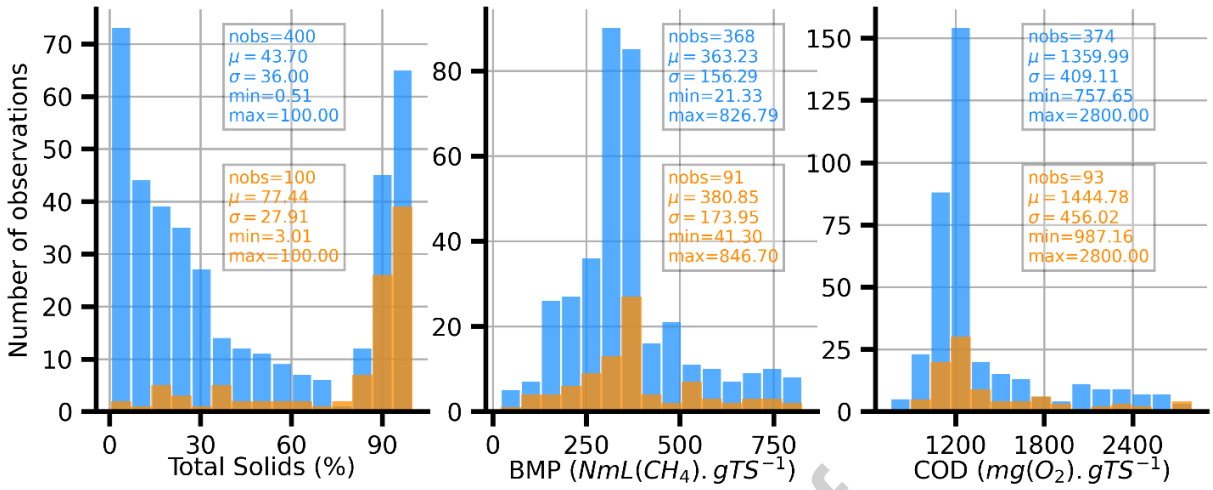


Figure 4 - Histograms of biochemical composition values (total solids - TS%, biochemical methane potential - BMP, chemical oxygen demand – COD) for train (blue) and test sets (orange). Respective number of observations (labeled as nobs), mean (labeled as μ), standard deviation (labeled as σ), and range (min/max) are provided in blue (train set) and orange (test set).

3.2. Model performances: linear methods, PLSR-based methods, machine learning methods and deep learning methods

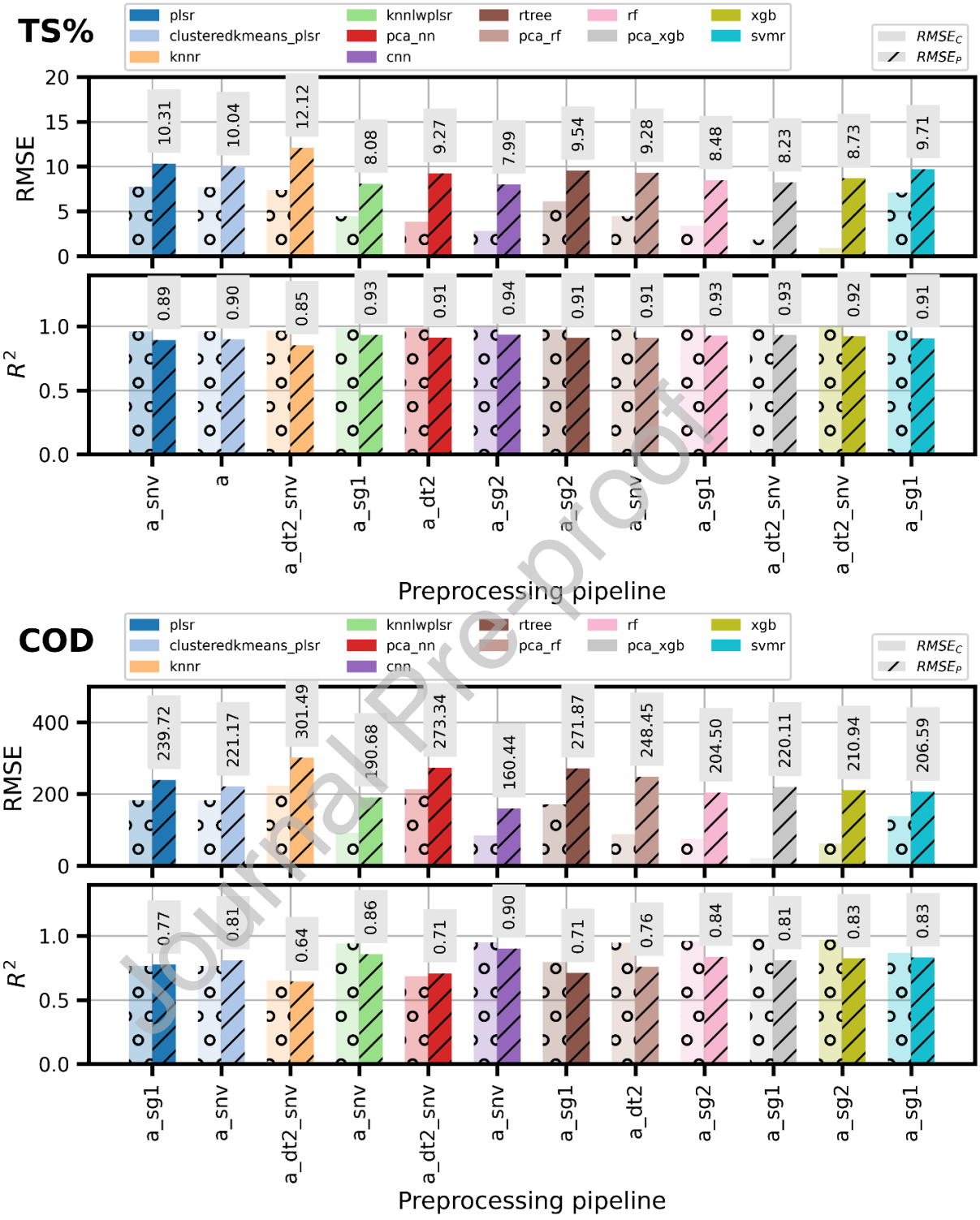
For each model type, the results (RMSE and R^2) on train and test sets for the best obtained models amongst the six preprocessing pipelines that were evaluated (a, a_sg1, a_sg2, a_snv, a_dt2, a_dt2_snv) are presented in Figure 5. The results for all the six preprocessing pipelines are presented in Appendix D and Appendix E. The RMSE and R^2 for the best obtained linear PLSR models were respectively 10.31% and 0.89 for TS% prediction, 240 $mg(O_2).gTS^{-1}$ and 0.77 for COD prediction, and, 127 $mL(CH_4).gTS^{-1}$ and 0.63 for BMP prediction. For TS% prediction, the difference in error between the NIRS-based method and the reference method (+5.31%) is rather low and acceptable. This is especially true knowing the huge benefit of being able to measure such variable in only a couple of minutes with NIRS instead of 24 hours with the reference method.

However, for COD or BMP prediction, the errors obtained with NIRS-based linear models are significantly higher than the errors of the reference methods (+112 $mg(O_2).gTS^{-1}$ for COD, and +49 $mL(CH_4).gTS^{-1}$ for BMP).

In most cases, non-linear methods allowed to build models with significantly lower errors than with the linear methods. Indeed, non-linear models showed RMSE values lowered by up to 23% for TS% prediction, up to 33% for COD prediction and up to 28% for BMP prediction. The best models were obtained with k-nearest neighbors locally weighted PLSR (knnlwplsr) and convolutional neural networks (cnn), with RMSE and R^2 respectively equal to 8% and 0.94 for TS% prediction, respectively equal to 160 $mg(O_2).gTS^{-1}$ and 0.90 for COD prediction, and respectively equal to 92 $mL(CH_4).gTS^{-1}$ and 0.80 for BMP prediction.

For some models, a prior PCA was applied to reduce the dimensionality of the input signals (pca_nn, pca_xgb, pca_rf). In most cases, this resulted in slightly lower model performances compared to models with the raw signal as input (xgb, rf). For example, for TS% prediction, applying a PCA reduction before a random forest (pca_rf) resulted in an RMSE and R^2 of 9.3% and 0.91 instead of 8.5% and 0.93 without PCA reduction (rf). For BMP prediction, applying a PCA reduction before an extreme gradient boosting (pca_xgb) resulted in an RMSE and R^2 of 101 $mL(CH_4).gTS^{-1}$ and 0.76 instead of 96 $mL(CH_4).gTS^{-1}$ and 0.79 without PCA reduction (xgb). However, in some cases, the PCA reduction did not deteriorate the models. For example, for BMP prediction, applying a random forest with (pca_rf) or without (rf) a PCA reduction resulted in equal RMSE and R^2 (respectively 111 $mL(CH_4).gTS^{-1}$ and 0.71). In such cases, one advantage of applying the dimension reduction method is the gain in computation time. Indeed, not

only the method is faster because there are less weights to optimize (20x25 weights instead of 1501x25 weights, which results in 25% less time for 500 samples), but because there are less weights to optimize, the number of iterations needed to find a satisfactory model is also reduced (for a batch size of 10, 100 epochs are needed instead of 500); which results in a final reduction in computation time of 85% for 500 samples. It must be pointed out that a dense neural network without prior dimension reduction (nn) was not evaluated because such model architecture results in too many coefficients to optimize, and therefore an unstable learning process.



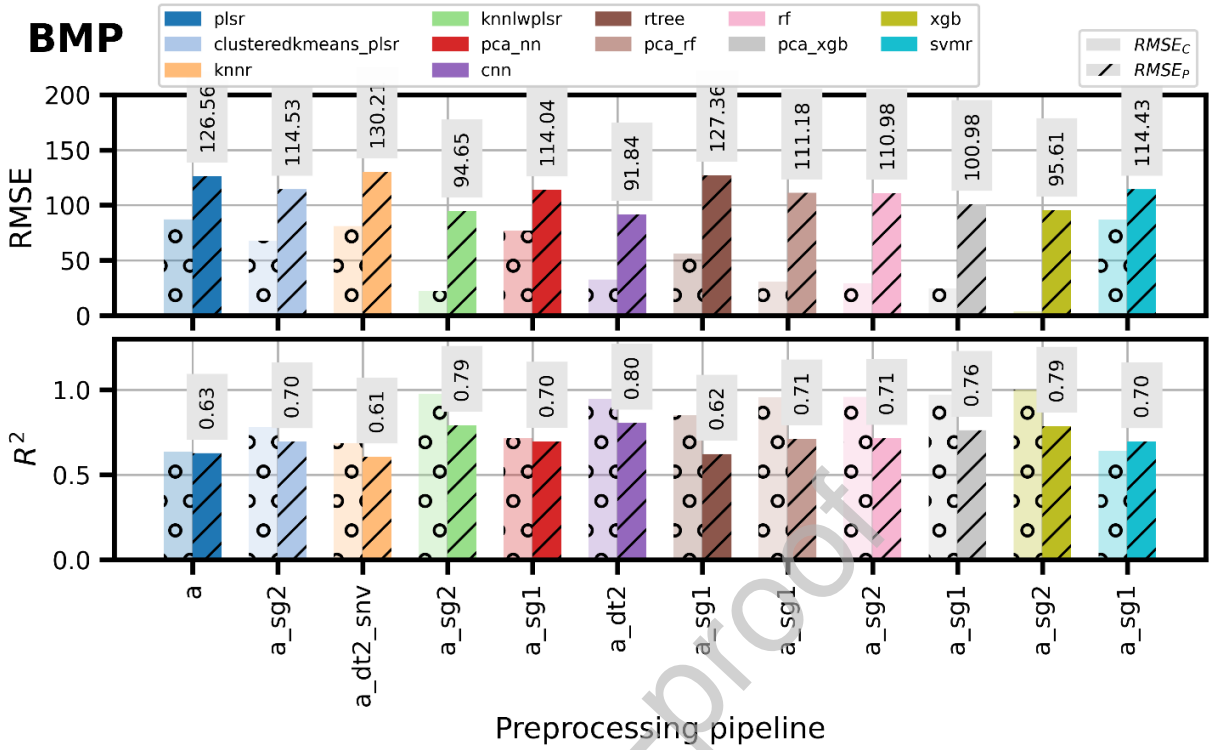


Figure 5 – Barplots of RMSE and R^2 obtained on train and test sets (respectively hatched with circles and lines) for all obtained models for the prediction of TS%, COD, and BMP. Bars are colored by model type (clustered K-Means PLSR – "clusteredkmeans_autoplsr", k-nearest neighbors locally-weighted PLSR – "knnlwplsr", k-nearest neighbors regression – "knnr", PCA followed by dense neural network – "pca_nn", PCA followed by random forest – "pca_rf", PCA followed by extreme gradient boosting – "pca_xgb", regression tree – "rtree", support vector machine regression – "svmr", PLSR – "plsr"). Only the results obtained with the best preprocessing pipeline is presented here (labeled in abscissa: "a", "a_sg1", "a_sg2", "a_snv", "a_dt2", "a_dt2_snv"). Results for all preprocessing pipelines are presented in Appendix E and all the other metrics (RMAE, MAE, MAD, r^2, SEP, Bias) are presented in a table in Appendix D.

A closer look at the models is provided in Figure 6. The observed and predicted values are shown for the three predicted variables (TS%, COD, BMP), for the best obtained models (linear and non-linear respectively to the left and right of the figure). The best non-linear models were obtained using convolutional neural networks (cnn) though, as already mentioned, the models obtained with k-nearest neighbors locally weighted PLSR (knnlwplsr) showed similar performances. These models when built directly on preprocessed data (*i.e.*, sg2, snv or dt2 for respectively TS%, COD and BMP) happen to be more performant than the models built on raw spectra. This puts into perspective some observations made in previous studies (Cui and Fearn, 2018) that highlighted the

fact that the advantage of convolutional neural networks is that they can be applied without any preprocessing steps (and that the preprocessing will be found automatically). Nevertheless, for all predicted variables (TS%, COD, BMP), the prediction points from the non-linear models appear much closer to the diagonal line than linear models which shows how non-linear models are much more suited for raw organic waste characterization. As expected, the train observations (in blue) are closer to this diagonal line than the test observations (in orange), because models were built on these train observations. However, the more significant differences observed between train and test observations for non-linear models could imply that these models are still slightly overfit. This would imply that these models leave further room for improvements (through further hyperparameter tuning). Nevertheless, the obtained errors on the test set (in orange) are already highly promising. The mean absolute deviation (MAD) is provided and complements the RMSE and R^2 by providing an idea of the error to expect for most samples. For BMP prediction, while the RMSEP equals to 91.8 $mL(CH_4).gTS^{-1}$, the MADP equals to only 52 $mL(CH_4).gTS^{-1}$. This shows how well the non-linear models perform, and how in most cases, the error made on BMP prediction is very low.

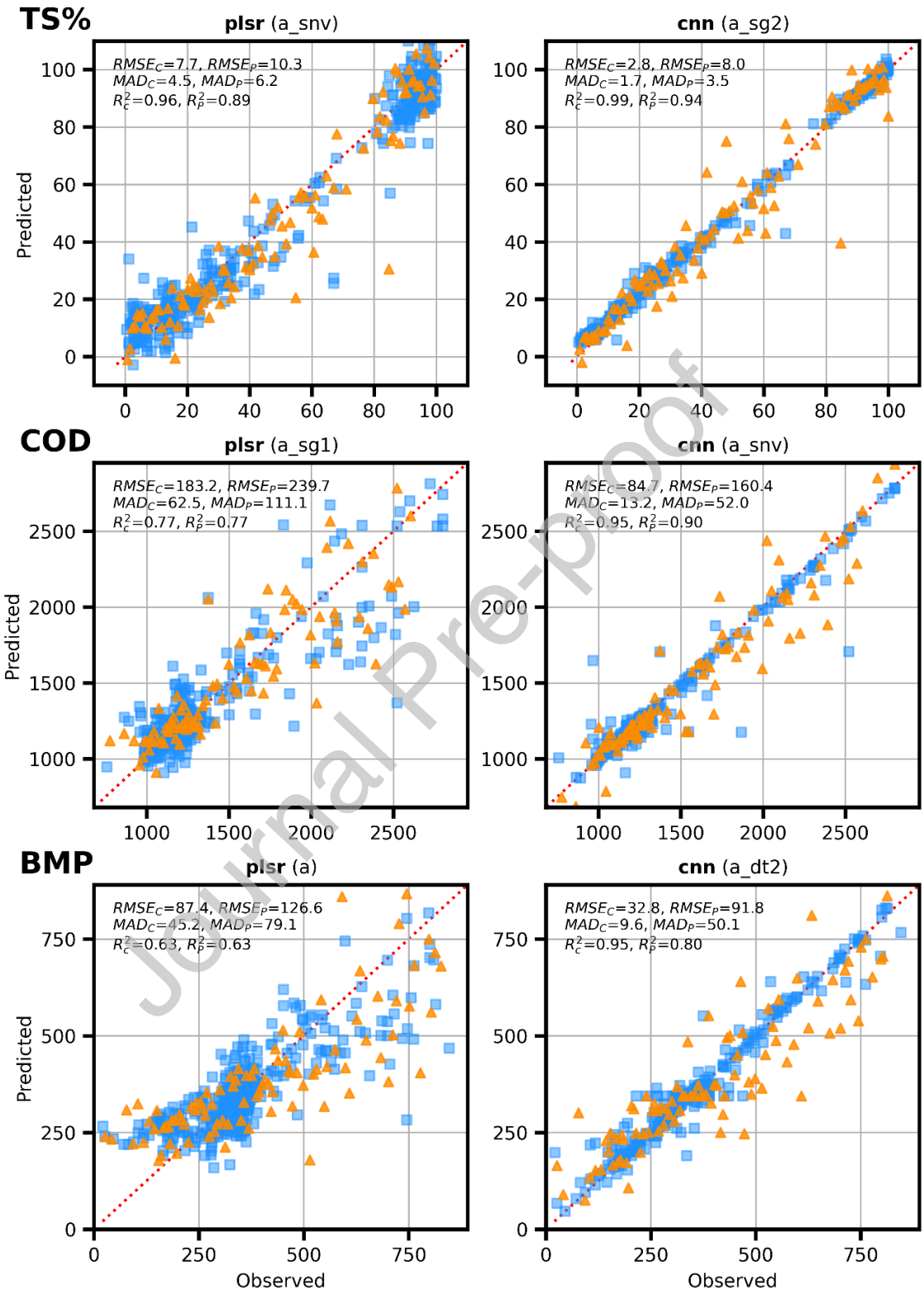


Figure 6 – Scatter plots of predicted and observed values for train and test sets (respectively in blue and orange). For each predicted variable (TS%, COD, BMP), the best obtained linear model (on the left) is compared with the best obtained non-linear model (on the right). The performance metrics (RMSE, MAE, RMAE, MAD, R^2, r^2, SEP, Bias) are provided for train and test sets (respectively with a subscript "C" for calibration/train and a subscript "P" for prediction/test).

3.3. Robustness towards moisture content effects

One of the questions regarding the use of these non-linear models, is whether the accuracy gain that was demonstrated here is due to a lower sensitivity to water effects (TS% differences), or due to better considering the differences in biochemical types when estimating the relationship between the signal and the variable to predict. As presented in section 2.1, in order to evaluate the robustness of developed NIRS models towards moisture content effects, a dataset of NIRS measurements acquired during N_2-drying experiments of various organic substrates was used (Mallet et al., 2021a).

For COD prediction, as indicated by the red arrow in Figure 7, switching from a linear to a non-linear model results in a decrease of 23% of the average inter-substrate error, but an increase of 13% of the average intra-substrate error. Similarly, for BMP prediction, as indicated by the red arrow in Figure 7, switching from a linear to a non-linear model results in a decrease of 14% of the average inter-substrate error, but an increase of 67% of the average intra-substrate error.

In other words, for COD or BMP prediction, the gain in accuracy that was previously observed for non-linear models results from a better modeling of the different substrate types (physical and biochemical differences) more than higher robustness towards TS% variations.

Theoretically, the ideal absorption law (the Beer-Lambert law) states that the relationship between concentration and the signal is linear, but because here there are so

many different substrates, the scattering levels are very different, and this makes this law poorly adapted. For example, the reflectance of highly absorbing samples such as raw biogas slurry, but also forward-scattering transparent liquids (oils) may have low general levels of measured reflectance. These samples will surely not have the same reflectance level than highly scattering samples such as straw, flour or biowaste. While linear models have difficulties coping with these big differences in signal levels, the power of non-linear models lies precisely in the ability to work by clusters of substrates, finding specific relationships between the signal and the predicted variable locally. Indeed, the relationship between BMP (or COD) and the measured signals on raw biogas slurry will be different than the relationship between the BMP (or COD) and the measured signals on straw, flour or biowaste; and using non-linear models this relationship can be modeled differently contrarily to linear models. The non-linear methods allow to account for these scattering differences between substrates, finding different (linear) relationships for each substrate types. In the end, non-linear methods allowed to reduce the errors that were previously made by linear models between substrate types. Another aspect related to moisture content effects concerns the OH absorption region around 1430 nm and 1940 nm. Many studies show that this region is in fact full of indirect information related to chemical composition, due to the multiple OH-bonding types that water makes with the molecules present in the substrate (Mallet et al., 2021a; Tsenkova et al., 2018). Future investigations should be oriented towards interpreting the non-linear models to better understand how this region is used by the models.

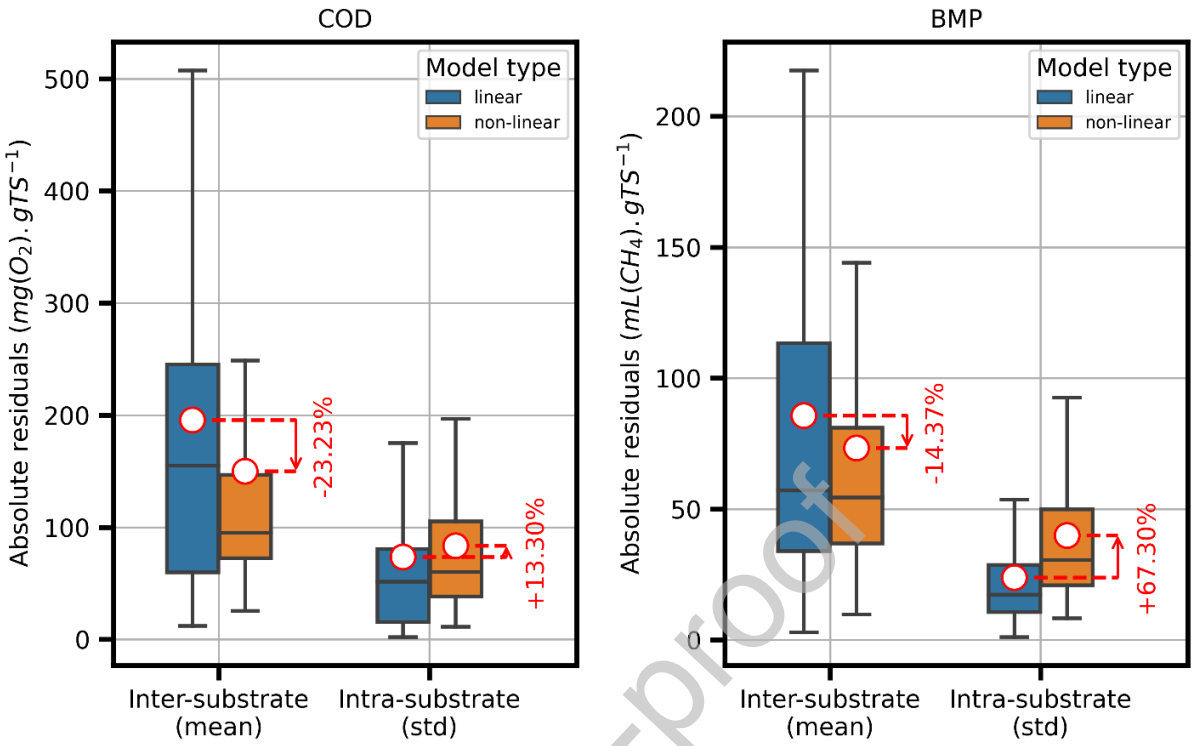


Figure 7 – Boxplots of absolute residuals (between observed and predicted) for the average per substrate (inter-substrate error), and the standard deviation per substrate (intra substrate error). The COD and BMP prediction results correspond respectively to the left and right subplots. The boxplots are colored by model type (linear in blue, non-linear in orange). Median values are presented with bold and black horizontal lines. The box limits represent the first and third quartile values (respectively Q1 and Q3), and the lines that extend from the box show the lowest and largest data points excluding any outliers (respectively Q1−1.5×(Q3−Q1) and Q1−1.5×(Q3−Q1)). Outliers, if existing, are presented in empty black circles. The mean of each boxplot is represented by a white dot surrounded by red.

This result redefines the scientific questions and technical challenges related to building NIR applications on raw (wet) and diverse organic waste. While water effects are certainly a high source of variance with non-linearities to be dealt with (Mallet et al., 2021a, 2021c), it appears that the non-linearity due solely to the diversity of organic waste (*i.e.*, biochemical and physical types) is very high, and that therefore applying non-linear models on such datasets can allow a significant gain in accuracy that allows these models to show similar final errors as linear calibrations built on dry samples.

4. Conclusions

In this study, non-linear methods including PLSR-based methods, machine learning methods and deep learning methods were successfully leveraged to build satisfactory TS%, COD and BMP prediction models based on NIRS and applicable on raw and diverse organic waste. A general gain in accuracy of 28% (based on RMSEP) was obtained compared to models built with linear methods. This significant gain was shown to be mostly due to better modeling the diversity of biochemical and physical types, more than being more robust to moisture content variations. Though not detailed in this study, the presented modeling approach based on non-linear methods could be successfully applied to other important parameters such as N-related parameters (total nitrogen content, proteins content), lipids content or carbohydrates content. This means that today, a full characterization of the organic waste is possible using NIRS and non-linear modeling. The demonstrated feasibility of applying NIRS on raw and diverse substrates, without any required sample preparation (freeze-drying and grinding), has huge implications for the industry. Indeed, this finally opens the door to online and at-site applications in the organic waste recovery industry (AD, composting, pyrogasification). From an applicative and industrial perspective, future steps will be to demonstrate equivalent performance on low-resolution and low-cost portable spectrometers, as already shown on dried samples (Mallet et al., 2021b). Indeed, while the current study has demonstrated the feasibility based on lab-scale spectrometers with costs ranging between 50 000 € and 100 000 €, the feasibility of the approach on low-cost spectrometers (below 10 000 €) should be validated for it to be fully suitable on full-scale biogas plants. Transfer strategies from standard benchtop spectrometers to online spectrometers will also need

to be evaluated. From a more fundamental perspective, future steps should focus on the interpretability of these non-linear models, in particular better understand how the OH absorption regions are used by these non-linear models.

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.

CRediT authorship contribution statement

Alexandre Mallet: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing - original draft, Supervision

Cyrille Charnier: Methodology, Investigation, Writing - review & editing, Supervision

Eric Latrille: Writing - review & editing, Supervision

Ryad Bendoula: Writing - review & editing, Supervision

Jean-Michel Roger: Writing - review & editing, Supervision

Jean-Philippe Steyer: Writing - review & editing, Supervision

Acknowledgements

Financial support from the National Research Institute for Agriculture, Food and Environment (INRAE), the French Agency of National Research and Technology (ANRT) [grant number 2018/0461] and the Biogaz-RIO platform [FEDER-FSE Languedoc Roussillon 2014-2020] is hereby acknowledged.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. https://doi.org/10.48550/arXiv.1603.04467

Acquarelli, J., van Laarhoven, T., Gerretzen, J., Tran, T.N., Buydens, L.M.C., Marchiori, E., 2017. Convolutional neural networks for vibrational spectroscopic data analysis. Anal. Chim. Acta 954, 22–31. https://doi.org/10.1016/j.aca.2016.12.010

Albrecht, R., Joffre, R., Gros, R., Le Petit, J., Terrom, G., Périssol, C., 2008. Efficiency of near-infrared reflectance spectroscopy to assess and predict the stage of transformation of organic matter in the composting process. Bioresour. Technol. 99, 448–455. https://doi.org/10.1016/j.biortech.2006.12.019

Angelidaki, I., Alves, M., Bolzonella, D., Borzacconi, L., Campos, J.L., Guwy, A.J., Kalyuzhnyi, S., Jenicek, P., Van Lier, J.B., 2009. Defining the biomethane potential (BMP) of solid organic wastes and energy crops: A proposed protocol for batch assays. Water Sci. Technol. 59, 927–934. https://doi.org/10.2166/wst.2009.040

Barnes, R.J., Dhanoa, M.S., Lister, S.J., 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 43, 772–777. https://doi.org/10.1366/0003702894202201

Belousov, A.I., Verzakov, S.A., Von Frese, J., 2002. Applicational aspects of support vector machines. J. Chemom. 16, 482–489. https://doi.org/10.1002/cem.744

Berzaghi, P., Flinn, P.C., Dardenne, P., Lagerholm, M., Shenk, J.S., Westerhaus, M.O., Cowe, I.A., NIR Publications, 2002. Comparison of linear and non-linear near infrared calibration methods using large forage databases, in: Near Infrared Spectroscopy: Proceedings of the 10th International Conference. p. 107.

Borin, A., Ferrão, M.F., Mello, C., Maretto, D.A., Poppi, R.J., 2006. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. Anal. Chim. Acta 579, 25–32. https://doi.org/10.1016/j.aca.2006.07.008

Charnier, C., Latrille, E., Jimenez, J., Lemoine, M., Boulet, J.C., Miroux, J., Steyer, J.P., 2017a. Fast characterization of solid organic waste content with near infrared spectroscopy in anaerobic digestion. Waste Manag. 59, 140–148. https://doi.org/10.1016/j.wasman.2016.10.029

Charnier, C., Latrille, E., Jimenez, J., Torrijos, M., Sousbie, P., Miroux, J., Steyer, J.P., 2017b. Fast ADM1 implementation for the optimization of feeding strategy using near infrared spectroscopy. Water Res. 122, 27–35. https://doi.org/10.1016/j.watres.2017.05.051

Cui, C., Fearn, T., 2018. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. Chemom. Intell. Lab. Syst. 182, 9–20. https://doi.org/10.1016/j.chemolab.2018.07.008

Cunha, C.L., Torres, A.R., Luna, A.S., 2020. Multivariate regression models obtained from near-infrared spectroscopy data for prediction of the physical properties of biodiesel and its blends. Fuel 261, 116344. https://doi.org/10.1016/j.fuel.2019.116344

de Santana, F.B., de Souza, A.M., Poppi, R.J., 2018. Visible and near infrared spectroscopy coupled to random forest to quantify some soil quality parameters. Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 191, 454–462. https://doi.org/10.1016/j.saa.2017.10.052

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L., Huvenne, J.P., 2009. Support vector

machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. Chemom. Intell. Lab. Syst. 96, 27–33. https://doi.org/10.1016/j.chemolab.2008.11.005

Doublet, J., Boulanger, A., Ponthieux, A., Laroche, C., Poitrenaud, M., Cacho Rivero, J.A., 2013. Predicting the biochemical methane potential of wide range of organic substrates by near infrared spectroscopy. Bioresour. Technol. 128, 252–258. https://doi.org/10.1016/j.biortech.2012.10.044

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1996. Support Vector Regression Machines, in: Mozer, M.C., Jordan, M., Petsche, T. (Eds.), Advances in Neural Information Processing Systems. MIT Press.

Eriksson, L., Trygg, J., Wold, S., 2009. PLS-Trees®, a top-down clustering approach. J. Chemom. 23, 569–580. https://doi.org/10.1002/cem.1254

Fitamo, T., Triolo, J.M., Boldrin, A., Scheutz, C., 2017. Rapid biochemical methane potential prediction of urban organic waste with near-infrared reflectance spectroscopy. Water Res. 119, 242–251. https://doi.org/10.1016/j.watres.2017.04.051

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Godin, B., Mayer, F., Agneessens, R., Gerin, P., Dardenne, P., Delfosse, P., Delcarte, J., 2015. Biochemical methane potential prediction of plant biomasses: Comparing chemical composition versus near infrared methods and linear versus non-linear models. Bioresour. Technol. 175, 382–390. https://doi.org/10.1016/j.biortech.2014.10.115

Holmes, G., Hall, M., Prank, E., 1999. Generating rule sets from model trees. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 1747, 1–12. https://doi.org/10.1007/3-540-46695-9_1

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55

Jacobi, H.F., Moschner, C.R., Hartung, E., 2011. Use of near infrared spectroscopy in online-monitoring of feeding substrate quality in anaerobic digestion. Bioresour. Technol. 102, 4688–4696. https://doi.org/10.1016/j.biortech.2011.01.035

Jimenez, J., Latrille, E., Harmand, J., Robles, A., Ferrer, J., Gaida, D., Wolf, C., Mairet, F., Bernard, O., Alcaraz-Gonzalez, V., Mendez-Acosta, H., Zitomer, D., Totzke, D., Spanjers, H., Jacobi, F., Guwy, A., Dinsdale, R., Premier, G., Mazhegrane, S., Ruiz-Filippi, G., Seco, A., Ribeiro, T., Pauss, A., Steyer, J.P., 2015. Instrumentation and control of anaerobic digestion processes: a review and some research challenges. Rev. Environ. Sci. Biotechnol. 14, 615–648. https://doi.org/10.1007/s11157-015-9382-6

Kim, S., Kano, M., Nakagawa, H., Hasebe, S., 2011. Estimation of active pharmaceutical ingredients content using locally weighted partial least squares and statistical wavelength selection. Int. J. Pharm. 421, 269–274. https://doi.org/10.1016/j.ijpharm.2011.10.007

Lesnoff, M., Metz, M., Roger, J.M., 2020. Comparison of locally weighted PLS strategies for regression and discrimination on agronomic NIR data. J. Chemom. 10–12. https://doi.org/10.1002/cem.3209

Lesteur, M., Latrille, E., Maurel, V.B., Roger, J.M., Gonzalez, C., Junqua, G., Steyer, J.P., 2011. First step towards a fast analytical method for the determination of Biochemical Methane Potential of solid wastes by near infrared spectroscopy. Bioresour. Technol. 102, 2280–2288. https://doi.org/10.1016/j.biortech.2010.10.044

Liu, J., Jin, S., Bao, C., Sun, Y., Li, W., 2021. Rapid determination of lignocellulose in corn stover based on near-infrared reflectance spectroscopy and chemometrics methods. Bioresour. Technol. 321, 124449. https://doi.org/10.1016/j.biortech.2020.124449

Malek, S., Melgani, F., Bazi, Y., 2018. One-dimensional convolutional neural networks for spectroscopic signal regression. J. Chemom. 32, 1–17. https://doi.org/10.1002/cem.2977

Mallet, A., Charnier, C., Latrille, É., Bendoula, R., Steyer, J.P., Roger, J.M., 2021a. Unveiling

non-linear water effects in near infrared spectroscopy: A study on organic wastes during drying using chemometrics. Waste Manag. 122, 36–48. https://doi.org/10.1016/j.wasman.2020.12.019

Mallet, A., Pérémé, M., Awhangbo, L., Charnier, C., Roger, J.M., Steyer, J.P., Latrille, É., Bendoula, R., 2021b. Fast at-line characterization of solid organic waste: Comparing analytical performance of different compact near infrared spectroscopic systems with different measurement configurations. Waste Manag. 126, 664–673. https://doi.org/10.1016/j.wasman.2021.03.045

Mallet, A., Tsenkova, R., Muncan, J., Charnier, C., Latrille, éric, Bendoula, R., Steyer, J.P., Roger, J.M., 2021c. Relating Near-Infrared Light Path-Length Modifications to the Water Content of Scattering Media in Near-Infrared Spectroscopy: Toward a New Bouguer-Beer-Lambert Law. Anal. Chem. 93, 6817–6823. https://doi.org/10.1021/acs.analchem.1c00811

Marini, F., Bucci, R., Magrì, A.L., Magrì, A.D., 2008. Artificial neural networks in chemometrics: History, examples and perspectives. Microchem. J. 88, 178–185. https://doi.org/10.1016/j.microc.2007.11.008

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference. pp. 56–61. https://doi.org/10.25080/majora-92bf1922-00a

Mishra, P., Passos, D., 2022. Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy. Postharvest Biol. Technol. 183. https://doi.org/10.1016/j.postharvbio.2021.111741

Mortreuil, P., Baggio, S., Lagnet, C., Schraauwers, B., Monlau, F., 2018. Fast prediction of organic wastes methane potential by near infrared reflectance spectroscopy: A successful tool for farm-scale biogas plant monitoring. Waste Manag. Res. 36, 800–809. https://doi.org/10.1177/0734242X18778773

Næs, T., Martens, H., 1984. Multivariate calibration. II. Chemometric methods. Trends Anal. Chem. 3, 266–271. https://doi.org/10.1016/0165-9936(84)80044-8

Narayanan, H., Sokolov, M., Butté, A., Morbidelli, M., 2019. Decision Tree-PLS (DT-PLS) algorithm for the development of process: Specific local prediction models. Biotechnol. Prog. 35, 1–11. https://doi.org/10.1002/btpr.2818

Nawar, S., Mouazen, A.M., 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. Sensors (Switzerland) 17, 1–22. https://doi.org/10.3390/s17102428

Ni, W., Nørgaard, L., Mørup, M., 2014. Non-linear calibration models for near infrared spectroscopy. Anal. Chim. Acta 813, 1–14. https://doi.org/10.1016/j.aca.2013.12.002

Nørgaard, L., Lagerholm, M., Westerhaus, M., 2013. Artificial Neural Networks and Near Infrared Spectroscopy -A case study on protein content in whole wheat grain. Focus 5.

Oliphant, T.E., 2010. Guide to NumPy, Methods. Trelgol Publishing USA.

Pedregosa, F., Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Mueller, A., 2015. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 19, 29–33.

Pérez-Marín, D., Garrido-Varo, A., Guerrero, J.E., 2007. Non-linear regression methods in NIRS quantitative analysis. Talanta. https://doi.org/10.1016/j.talanta.2006.10.036

Preda, C., Saporta, G., 2005. Clusterwise PLS regression on a stochastic process. Comput. Stat. Data Anal. 49, 99–108. https://doi.org/10.1016/j.csda.2004.05.002

Roger, J., Boulet, J.-C., Zeaiter, M., Rutledge, D.N., 2020. Pre-processing Methods, in: Comprehensive Chemometrics. Elsevier, pp. 1–75. https://doi.org/10.1016/b978-0-12-409547-2.14878-4

Savitzky, A., Golay, M.J.E., 1964. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Anal. Chem. 36, 1627–1639. https://doi.org/10.1021/ac60214a047

Shen, G., Lesnoff, M., Baeten, V., Dardenne, P., Davrieux, F., Ceballos, H., Belalcazar, J.,

Dufour, D., Yang, Z., Han, L., Fernández Pierna, J.A., 2019. Local partial least squares based on global PLS scores. J. Chemom. 33, 1–3. https://doi.org/10.1002/cem.3117
Shenk, J.S., Westerhaus, M.O., Berzaghi, P., 1997. Investigation of a LOCAL calibration procedure for near infrared instruments. J. Near Infrared Spectrosc. 5, 223–232. https://doi.org/10.1255/jnirs.115
Snee, R.D., 1977. Validation of Regression Models: Methods and Examples. Technometrics 19, 415–428. https://doi.org/10.1080/00401706.1977.10489581
Tøndel, K., Indahl, U.G., Gjuvsland, A.B., Vik, J.O., Hunter, P., Omholt, S.W., Martens, H., 2011. Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models. BMC Syst. Biol. 5, 1–17. https://doi.org/10.1186/1752-0509-5-90
Triolo, J.M., Ward, A.J., Pedersen, L., Løkke, M.M., Qu, H., Sommer, S.G., 2014. Near Infrared Reflectance Spectroscopy (NIRS) for rapid determination of biochemical methane potential of plant biomass. Appl. Energy 116, 52–57. https://doi.org/10.1016/j.apenergy.2013.11.006
Tsenkova, R., Pollner, B., Kovacs, Z., 2018. Essentials of Aquaphotomics and Its Chemometrics Approaches 6, 1–25. https://doi.org/10.3389/fchem.2018.00363
van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual, CreateSpace. Scotts Valley, CA. https://doi.org/10.5555/1593511
Vergnoux, A., Guiliano, M., Le Dréau, Y., Kister, J., Dupuy, N., Doumenq, P., 2009. Monitoring of the evolution of an industrial compost and prediction of some compost properties by NIR spectroscopy. Sci. Total Environ. 407, 2390–2403. https://doi.org/10.1016/j.scitotenv.2008.12.033
Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. Pietro, Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C.N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D.A., Hagen, D.R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G.A., Ingold, G.L., Allen, G.E., Lee, G.R., Audren, H., Probst, I., Dietrich, J.P., Silterra, J., Webber, J.T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J.L., de Miranda Cardoso, J.V., Reimer, J., Harrington, J., Rodríguez, J.L.C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N.J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P.A., Lee, P., McGibbon, R.T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T.J., Robitaille, T.P., Spura, T., Jones, T.R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.O., Vázquez-Baeza, Y., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2
Wold, H., 1973. Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments, Multivariate Analysis–III. Academic press, Inc. https://doi.org/10.1016/b978-0-12-426653-7.50032-6
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
Yang, G., Li, Y., Zhen, F., Xu, Y., Liu, J., Li, N., Sun, Y., Luo, L., Wang, M., Zhang, L., 2021. Biochemical methane potential prediction for mixed feedstocks of straw and manure in anaerobic co-digestion. Bioresour. Technol. 326, 124745. https://doi.org/10.1016/j.biortech.2021.124745
Yao, Y., Shen, X.M., Qiu, Q., Wang, J., Cai, J.H., Zeng, J.S., Lang, X.Y., 2020. Predicting the

Biochemical Methane Potential of Organic Waste with Near-Infrared Reflectance Spectroscopy Based on GA-SVM. Guang Pu Xue Yu Guang Pu Fen Xi/Spectroscopy Spectr. Anal. 40, 1857–1861. https://doi.org/10.3964/j.issn.1000-0593(2020)06-1857-05