Composites Science and Technology 264 (2025) 111127

Contents lists available at ScienceDirect

Composites Science and Technology

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

Machine learning-driven property predictions of polypropylene composites using IR spectroscopy

Szilvia Klébert [a], Róbert Várdai [a,b], Anita Rácz [a,*]

[a] Institute of Materials and Environmental Chemistry, HUN-REN Research Centre for Natural Sciences, Magyar Tudósok Körútja 2, Budapest, H-1117, Hungary
[b] Laboratory of Plastics and Rubber Technology, Department of Physical Chemistry and Materials Science, Faculty of Chemical Technology and Biotechnology, Budapest University of Technology and Economics, Műegyetem Rkp. 3, Budapest, H-1111, Hungary

ARTICLE INFO

Keywords:
Regression
PLS
Neural network
Mechanical properties

ABSTRACT

There is a growing need for environmentally friendly alternatives to the determination of the mechanical properties, thermal stability and other functional characteristics of polymer composites, which led to the use of machine learning modeling combined with fast, non-destructive measurements like Fourier-transform infrared spectroscopy (FTIR). In this study, we have successfully classified almost 200 in-house polypropylene composites according to the applied reinforcements with the above-mentioned combination of methods. The balanced accuracy of test validation was over 0.9 for the extreme gradient boosting (XGBoost)-based model. With the same IR spectra, we have developed consensus machine learning models for predicting the modulus, tensile strength and elongation at break – which are important mechanical properties from the application point of view. The three-step validation protocol has verified that the models were appropriate for the prediction of the mechanical features of the polymer composites and their classification based on the applied reinforcements.

1. Introduction

Future challenges will demand innovative solutions that push technological boundaries, necessitating the development of materials with diverse functionalities, tailored properties designed for specific applications and performance. Throughout history, the development of new materials has been one of the most significant accomplishments of the given era, paving the way for the creation of new technologies. Conversely, at times, a leap in technological advancement requires the discovery of new materials [1]. Today, the likelihood of exploring an entirely new material is quite low. Instead, researchers are focusing on creating novel materials by combining existing ones, emphasizing properties and performance [2]. As a result, modern advanced materials are primarily composites and hybrid composites exhibiting superior properties that the original materials did not possess [3]. Composites typically consist of a matrix and reinforcement, but hybrid ones involve more than one kind of reinforcing material. Composites are not solely a human creation, as natural materials like wood or bone are also composites. However, the advent of modern synthetic composites, especially those based on polymer matrices, began in the mid-20th century. The primary benefits of polymer composites besides outstanding properties, include their ease of processing into various structural forms, their ability to substitute traditional materials with significant weight reduction, enhanced reliability, or chemical inertness [4]. Among the various types of composites or hybrid composites, those based on polypropylene (PP) have garnered significant attention due to their versatile properties and wide-ranging applications [5]. In polypropylene-based hybrid composites, the polymer matrix of polypropylene is often combined with different fibers such as glass, carbon, polymer or natural fibers to enhance mechanical strength, thermal stability, and other functional characteristics [6]. However, the growing emphasis on sustainability and environmental considerations is driving the development of bio-based and recyclable polypropylene composites [7]. The use of natural fibers as reinforcements aligns with the global push towards greener materials and circular economy principles. The concept behind fiber reinforced composites is straightforward: rigid and strong fibers bear the load, while the polymer matrix distributes it among the fibers. The stiffness and strength of the fibers typically employed in these composites are even 100 times greater than those of the matrix [8]. The characteristics of all heterogeneous polymers are influenced by four key factors: the properties of the components, their composition, the structure, and the interfacial adhesion. To achieve the best performance and cost-efficiency, it is essential to optimize all these equally crucial factors [9]. Each type of filler imparts unique properties

* Corresponding author.
E-mail address: racz.anita@ttk.hu (A. Rácz).

https://doi.org/10.1016/j.compscitech.2025.111127
Received 20 October 2024; Received in revised form 24 February 2025; Accepted 25 February 2025
Available online 28 February 2025

to the composite, broadening the scope of polypropylene's applications. As a member of the polyolefin family, polypropylene is hydrophobic, without any functional groups, which often results in poor adhesion with polar or hydrophilic fillers [10]. To overcome this challenge, various surface treatments and compatibilizers are employed. Surface treatments [11–13] can modify the filler surface to enhance compatibility with the polypropylene matrix. Compatibilizers, like maleic anhydride-grafted polypropylene (MAPP), are added to the composite to improve interfacial adhesion by acting as a bridge between the matrix and the filler [14]. However, these improvements in mechanical properties often come with trade-offs. For example, while the stiffness and strength increase, the impact resistance of the composite might decrease compared to pure polypropylene [15]. To address this shortcoming, hybrid composites are often prepared by incorporating a third component, typically an elastomer [16,17]. Fibers and fillers increase stiffness while elastomer responsible for the increased impact resistance, but in case of natural fiber composites this principle does not work, which can be attributed to several reasons [18]. In wood hybrid composites, reinforcement with polymer fibers results in satisfactory impact resistance [19].

With newer technologies prioritizing to reduce environmental burden, researchers are increasingly turning to chemometric and machine learning models to replace experimental methods that are time-consuming, energy-intensive, and costly [20]. The growing interest has led to the use of chemometric modeling methods in material characterization, particularly utilizing easy, fast, and non-destructive measurements like FTIR. Numerous studies have shown that with enough samples and a suitable algorithm, a correlation can be established between the recorded FTIR spectra and the desired property, allowing for easy prediction from a single FTIR spectrum [21]. This method has become a standard workflow for quality control, particularly in the food and pharmaceutical industries. Regarding composites, the multivariate modeling techniques have effectively accelerated the design process itself [22,23], moreover different features (e.g., mechanical, thermal, tribological and electrical) can be predicted as well [24]. In most cases, the input parameters for the prediction of the mechanical properties are physical characteristics of the starting materials, like density and thickness [25], particle size of fillers [26], thermal conductivity [27]. S. M. Khan [28] predicted the mechanical properties by using material type, composition and number of reinforcement and matrix layers as input parameters for cross-ply laminated fiber-reinforced polymer composites. In the study of M.S. Kabbani and coworkers [29], the effect of the cooling rate on the mechanical properties of glass fiber–polypropylene composites was predicted based on the composition of the material or the number of matrix and reinforcement layers. T. Nguyen [30] used FTIR spectra as input for predicting the molecular weight.

In this study, we have used almost 200 different in-house PP-based composites (or hybrid composites), with various types and concentrations of polymer fibers and natural or synthetic reinforcements. The preparation and the mechanical characterization of these samples were lengthy and exhausting, therefore our aim was to develop multivariate regression models for predicting modulus, tensile strength and elongation at break with the help of a fast analytical method, FTIR. Moreover, we have aimed for the accurate classification of the composites according to the used reinforcements. To the best of our knowledge, this is the first application of FTIR spectra for the determination of the above-mentioned mechanical features and classification of polypropylene composites.

2. Materials and methods

2.1. Materials

Polypropylene (PP) Daplen HJ 325 MO, manufactured by Borealis GmbH, Austria, was utilized, featuring a melt flow rate (MFR) of 50 g/10 min (2.16 kg, 230 °C) and a density of 0.91 g/cm^3. To enhance interfacial adhesion, maleic anhydride-grafted polypropylene (MAPP), specifically the Scona 6102 grade from Byk-Chemie GmbH, Germany, was employed. This MAPP has an MFR of 3.5 g/10 min (2.16 kg, 190 °C) and a maleic anhydride content of 0.9–1.2 %. The functionalized PP constituted 10 wt% of the total reinforcement. Poly(vinyl alcohol) (PVA) from Kuraray, Japan, and poly(ethylene terephthalate) (PET, T713, Performance Fibers GmbH) were used as polymer fibers. For reinforcement, carbon fibers, glass fibers, wood flour, and talcum were employed. The carbon fiber, Panex PX 35 from Zoltek Zrt, Hungary, had a diameter of 7.2 μm and a length of 8.4 mm, and was coated with an epoxy/polyurethane blend, further over-coated with polyester resin. The glass fiber, ThermoFlow 636 from Johns Manville, USA, had a diameter of 13 μm and a length of 4 mm, coated with silane sizing. Wood flour Filtracell EFC 1000 from Rettenmaier and Söhne GmbH, Germany, had fibers averaging 363 μm in length and 64 μm in diameter. The Jetfine 3CA talcum from Imerys Performance Minerals (Paris, France) had an average particle size of 6 μm and a 2.78 g/cm^3 density. The composition of composites containing only one type of fiber ranged from 0 to 60 wt% in 5 wt% increments. Hybrid composites consistently included 20 wt% of reinforcing fiber (carbon, glass, or wood) and talc, with the polymer fiber (PET, PVA) content varying from 0 to 40 wt% in 5 wt% increments.

2.2. Sample preparation

The components were homogenized using a twin-screw compounder (Brabender DSK 42/7, Brabender, Germany) at set temperatures of 180-190-200-210 °C and a screw speed of 40 rpm. Prior to extrusion, wood flour was dried at 105 °C for 4 h, and PVA and PET fibers were dried at 80 °C for 4 h in a vacuum oven. To enhance homogeneity, the extrusion process was repeated once. The resulting granulated composites were injection molded into standard ISO 527 1A tensile bars, each 4 mm thick, using a Demag IntElect 50/330-100 injection molding machine (Demag, Germany). The processing parameters were set as follows: temperatures of 40-170-180-190-200 °C, injection pressures ranging from 300 to 1200 bar depending on fiber type and content, a back pressure of 50 bar, an injection speed of 50 mm/s, a holding time of 25 s, and a cooling time of 30 s. The mold temperature was maintained at 40 °C. The specimens were then stored at ambient temperature (25 °C, 50 % relative humidity) for one week before further testing.

2.3. Mechanical characterization

Tensile tests were conducted using an Instron 5566 universal testing machine (Instron, USA) with a gauge length of 115 mm and a crosshead speed of 5 mm/min. Young's modulus was determined within the strain range of 0.05–0.25 % by fitting a straight line to the recorded data. The values for tensile strength and elongation-at-break were obtained from the stress-strain curves recorded during the tests. Modulus is the least susceptible to micromechanical deformation, therefore the original reference measurements of elongation at break and tensile strength have higher errors compared to those for modulus.

2.4. Infrared spectroscopy

Fourier transform infrared spectra were recorded in attenuated total reflectance (ATR) mode with a Spectrum 3 spectrometer (PerkinElmer, Inc.) equipped with a Universal ATR (UATR) accessory (PerkinElmer, Inc.). Spectra were recorded from 4000 cm^{-1} to 550 cm^{-1} at a spectral resolution of 4 cm^{-1} with 32 scans applying a contact pressure of 100 N. The average of five parallel spectra were used for the further evaluation. Baseline correction (Automatic Whittaker transformation) with 1st derivative and optionally standard normal variate (SNV) transformation were used as pre-scaling for the models. We will specify the exact combinations later on in each model.

2.5. Statistical evaluation

2.5.1. Classification methods

In the case of classification models, we have used two machine learning based algorithms: random forest (RF), and extreme gradient boosting (XGBoost). Both algorithms are tree-based techniques, and they are well-known especially for classification studies. Tree-based algorithms create binary splits of the data into branches with the most important variables and their limit values in the split points. The decision trees can be considered as the models, and their aim is to find the best splits, in order to accurately classify the samples in their groups. The difference between the two applied algorithms is the following: while random forest consists of individual trees that form a voting-based consensus to determine the class memberships of the samples [31], XGBoost sequentially improves the single trees, based on the errors of the preceding trees [32].

In RF, the number of models (trees) was set to 100, and the information gain ratio was applied as the split criterion. In XGBoost, as there were multiple classes, softmax probability (softprob) was used as the objective function and similarly to RF, a hundred models (boosting rounds) were generated. The KNIME (4.6.4) software environment was used for the machine learning modeling [33]. The complete KNIME workflow for classification is shown in the Supplementary Information (Fig. S1).

The classification models were evaluated based on their accuracy and balanced accuracy metrics:

$$Accuracy\ (ACC) = \frac{correctly\ predicted\ samples}{total\ number\ of\ samples} \tag{1}$$

$$Balanced\ accuracy\ (BACC) = \frac{\sum_j^k \frac{n_{j,corr}}{n_{j,total}}}{k} \tag{2}$$

where $n_{j,corr}$ is the number of correctly classified samples in class j, $n_{j,actual}$ is the total number of samples in class j, and k means the number of classes. In addition to accuracy, balanced accuracy is a recommendable and sensitive performance parameter for imbalanced classes [34]. The definition of F1 score value is explained in detail in our previous work [34].

For the validation of the models, we have used a combination of i) iterative randomized (stratified) 5-fold cross-validation (CV), with ii) 80:20 split ratio for test validation (20 % of the data is separated randomly for test) and iii) a permutation test. For the permutation test, we have used the same distribution of the classes as in the original dataset.

2.5.2. Regression methods

In case of the regression models, we have combined the traditional partial least squares regression method with a machine learning-based neural network. We have selected these two algorithms out of several classical and machine learning tools, based on our preliminary results and experience. Moreover, these algorithms can complement each other as they can capture both linear and non-linear relationships in the data.

PLS is a classical and well-known multivariate linear technique for the prediction of continuous reference variables. The basic idea of PLS is the relationship between input variables (here spectral variables, **X**), and reference variables (**Y**), such as the mechanical properties of the polypropylene composite samples. Both **X** and **Y** are expressed in terms of score and load matrices, but there is also a connection between the **X** and **Y** matrices: $\mathbf{Y} = \mathbf{X}b + \mathbf{E}$, where b is the regression coefficient. These coefficients are coming from the inner relationship between the two score matrices. An important step in PLS is the optimization of the number of PLS components (latent variables, LVs). If we are using too much, we can risk overfitting, but with too few LVs, the models can be underfitted. Here, we have used the global minimum of the root mean squared error of cross-validation (RMSECV) values for the determination of the optimal number of LVs. These values were used together with the other root mean squared errors (RMSE) for the calibration (RMSEC), cross-validation (RMSECV) and test validation (RMSEP) in order to determine the performance of the models. Moreover, R^2 values (determination coefficients) and mean absolute errors (MAE) were also calculated for the proper evaluation of the models.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \tag{4}$$

$$MAE = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n} \tag{5}$$

where y_i is the measured value (reference), $\hat{y}_i$ is the predicted value, $\bar{y}_i$ is the average of the measured values and n is the number of samples.

The other regression method was the machine learning-based neural network (NN), which is a nonlinear algorithm for the prediction of continuous reference variables [35]. Neural networks are nonlinear models inspired by the human brain, consisting of hidden layers of interconnected neurons (nodes) that process input data (here the spectral dataset) to detect patterns and make predictions in the output. It learns to perform tasks by adjusting the weights of connections through training. We have used KERAS neural network (with python deep learning integration in KNIME) for the tasks with three densely connected hidden layers. The number of neurons was optimized separately between 10 and 1000 in a hyperparameter optimization loop. In the past few years, exploring the overparametrized area for the finding of global optima has been a commonly accepted paradigm, which is also beneficial for model generalization [36].

In the case of regression models, genetic algorithm [37] was used as a variable selection technique for the selection of the most important spectral variables. This statistical tool helps to maximize the connection between the spectra and the response variable, which can sometimes depend on visually smaller, but important and information-rich details of the spectra. Here, RMSECV was used as a typical fitness function in searching for the local optima based on the PLS models. The applied parameter settings are summarized in Table 1.

It is highly effective to use variable windows instead of unique variables, especially in the case of spectral datasets, where the neighboring variables are often correlated [38], therefore we have split the datasets into 50 or 25 variable-size windows during the variable selection process with genetic algorithms. These spectral intervals were created in a continuous way from the data and there was no overlap between them. PLS regression and genetic algorithm was carried out in MATLAB 2019a (MathWorks, Inc.) with PLS Toolbox (v. 8.8.1, Eigenvector Research, Inc.). The selected variables in each spectral dataset were used as input variables for the further modeling phases with PLS and neural networks.

For validation, the same triad of validation methods were used as in the case of classification: i) an iterative 5-fold randomized cross-validation, ii) with 75:25 split ratio test validation and iii) a permutation test. Here, the Kennard-Stone algorithm was used to split the dataset

Table 1
Parameter settings for the genetic algorithm variable selection.

GA parameter settings	
Population size	100
Window width	25 or 50
Initial terms (%)	30
Max generations	100
% at convergence	50
Mutation rate	0.005
Crossover	Double
Fitness	RMSECV

into training and test sets. The consensus model was based on the average of the predicted values of the PLS and NN models. The measured vs. predicted plots of the regression models were edited in OriginPro 2018 software (OriginLab Corporation). The complete KNIME workflow for the neural network and consensus models is shown in the Supplementary Information (Fig. S2).

3. Results and discussion

In both classification and regression modeling, the pretreated spectra were used. Fig. 1 shows the original (raw) and the transformed (baseline correction + 1st derivative + SNV) average spectra. The exact spectrum pre-scaling combinations will be specified in the classification and regression sections.

3.1. Classification of the reinforcements in polypropylene composites

In the classification phase, we have split the IR dataset into four groups, based on the major reinforcement fibers. The samples in the groups contained the reinforcement materials in different ratios, moreover additional fillers were also in the composite mixes. Here, our aim was to explore the possibility of separation based on the applied reinforcements even if we have a noisy background with other components in the composites. The grouping system is shown in detail in Table 2. We have excluded the original reference composites, which contained only PVA or PET in the polypropylene or polypropylene/elastomer composites. Therefore, we have used 151 samples out of the total 199 for the classification model. The average raw spectra of the four classes can be seen in the Supplementary information (Fig. S3). Based on the figure, we can safely say that three classes out of four have very similar spectra, which are practically indistinguishable visually. Only one class (talcum) stands out, but it would not be enough to accurately classify the whole dataset.

We have used RF and XGBoost for the classification modeling with the same validation protocols. At first, the dataset of the pre-scaled spectra (Baseline correction + 1st derivative + SNV) was split into training and test set in an 80:20 ratio. In the next step, 5-fold stratified and randomized cross-validation (CV) was applied in 10 iterations on the training set. The average of the probability values for each class and for each composite sample were calculated based on the iterations. Performance parameters, such as accuracy (ACC) and balanced accuracy (BACC) were reported for the cross-validation and test validation part as well. Balanced accuracy was important to calculate due to the slightly imbalanced distribution between the sizes of the classes. The summary of the performances is shown in Tables 3 and 4.

Table 2
The contents of the different classes in the model with the sizes of the groups (n).

Class	Reinforcements	Additional components	n
1	Carbon fiber	PVA/MAPP	30
2	Wood flour	PET/PVA/MAPP	44
3	Glass fiber	PET/PVA/MAPP	45
4	Talcum	PET/PVA/MAPP	32

Permutation tests were carried out in order to exclude the possibility of overfitting. In the case of permutation test, the cross-validation of the models was repeated ten times with randomized class memberships. The accuracy of the RF and XGBoost models in the permutation test scenario was decreased to 0.33 and 0.39 respectively, which is even worse than the performance of the fully random scenario (accuracy = 0.5). This result confirmed that the RF and XGBoost models are not overfitted.

Both models could provide outstanding classification based on the reinforcements, however XGBoost model was a bit better in cross-validation, therefore we have selected it as the best one. The ATR-FTIR measurements coupled with tree-based machine learning models are clearly capable of the determination of the different reinforcements in the polypropylene composites.

3.2. Regression models to determine the quantitative features

In this section, the following three reference measurement data of the composites were modeled: i) modulus, ii) toughness and iii) elongation at break. These three reference measurements were the output variables; therefore, the models predicted these reference values as the outcome (as predicted values). The same pre-scaled spectra were used as in the case of classification, and they can be considered as the input

Table 3
Summary of the performance parameters for the two models.

Model	CV		Test validation	
	ACC	BACC	ACC	BACC
RF	0.983	0.980	0.935	0.917
XGBoost	**0.992**	**0.990**	**0.935**	**0.917**

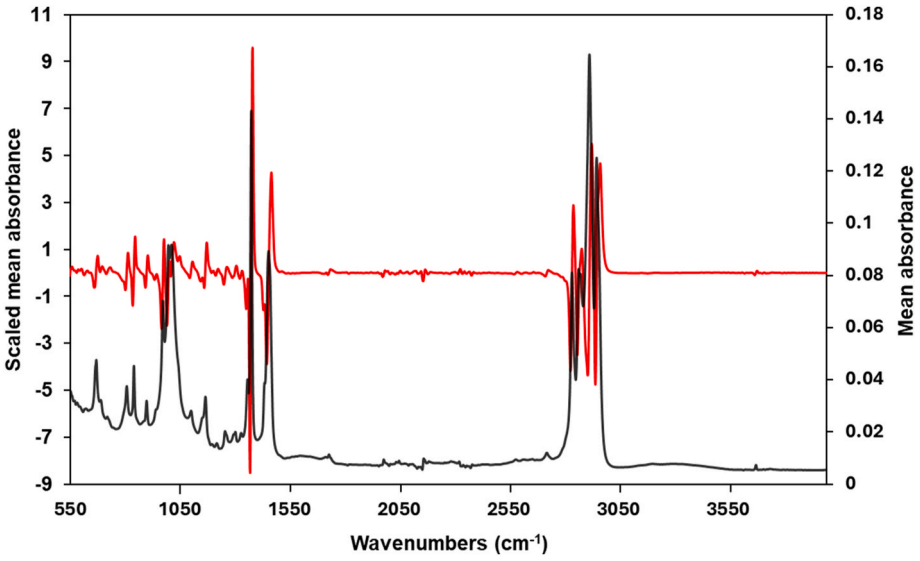

Fig. 1. The mean raw (black) and pre-scaled (Baseline correction + 1st derivative + SNV) (red) ATR-FTIR spectra of the whole spectral dataset.

Table 4
Confusion matrices and F1 score values for the two models.

XGBoost model							RF model						
CV	Class	Predicted				F1 score	CV	Class	Predicted				F1 score
		1	2	3	4				1	2	3	4	
Actual	**1**	23	1	0	0	0.979	**Actual**	**1**	23	0	1	0	0.979
	2	0	35	0	0	0.986		**2**	0	35	0	0	0.986
	3	0	0	36	0	1.000		**3**	0	0	36	0	0.986
	4	0	0	0	25	1.000		**4**	0	1	0	24	0.980
Test	**Class**	**Predicted**				**F1 score**	**Test**	**Class**	**Predicted**				**F1 score**
		1	**2**	**3**	**4**				**1**	**2**	**3**	**4**	
Actual	**1**	4	0	2	0	0.800	**Actual**	**1**	4	0	2	0	0.800
	2	0	9	0	0	1.000		**2**	0	9	0	0	1.000
	3	0	0	9	0	0.900		**3**	0	0	9	0	0.900
	4	0	0	0	7	1.000		**4**	0	0	0	7	1.000

variables for modeling. PLS regression and KERAS neural networks were used for regression modeling and we have combined the results of the two models. Genetic algorithm was used as variable selection for models with an initial population of 100 and 100 iterations. The validation protocol was the same for all three cases: iterative randomized 5-fold cross-validation (CV) was used in both PLS and NN models and the datasets were split into training and test sets in a 75:25 ratio to verify the performance of the models with an additional test validation. Furthermore, permutation tests were used to rule out overfitting in both the PLS and NN models. The complete modeling workflow is shown in Fig. 2.

3.2.1. *Prediction of modulus*

In the prediction of modulus, the dataset originally contained 195 composite samples in the rows, and 3451 spectral variables in the columns. Here, we have used baseline correction with the 1st derivative for the PLS model and an additional SNV transformation for the NN model as pre-scaling. The covered range of the reference modulus in the model was between 1.3 and 13.4 GPa.

In the modeling workflow, PLS model was generated at first. Test samples were separated in the beginning out of the 195 samples, therefore the PLS modeling was carried out on 141 training samples. In the variable selection part, the variable window size was 50. This way the genetic algorithm was selected 950 variables out of 3451 for the final modeling process. The selected variables are shown in Fig. S4A in the Supplementary information. Five-fold randomized cross-validation and a test validation part have been applied, and the optimal PLS model contained 10 LVs according to RMSECV values of cross-validation. The performance of the PLS model is summarized together with the NN and combined models at the end of the section in Table 5. The permutation test has successfully proved that the PLS model was not overfitted.

In the next step, for the NN model, the previously selected 950 variables were used as input data. At first, hyperparameter optimization was applied for the determination of the optimal number of hidden neurons in the dense hidden layers. We have used three dense layers, and the final numbers of neurons were 227, 683 and 127, respectively. The same test set was used for the NN model, as for the PLS model. For the validation of the model, we have used both five-fold cross-validation and an iterative test validation (changing the random seed in the neural network node). Finally, the additional permutation test showed that the NN-based model was not overfitted. The performance parameters of the NN model are shown in Table 5.

In the final step, the predicted modulus values of the PP composites based on PLS and NN models have been merged, and the average of the predicted values was calculated for both cross-validation and test validation samples. The final performance parameters of the consensus (or combined) model are shown in Table 5.

The measured and predicted modulus values based on the consensus model are plotted in Fig. 3.

The performance parameters and the measured vs. predicted plot can verify that the prediction of moduli is highly reliable based on the ATR-FTIR spectra of the composites. While PLS model was a bit better in cross-validation, the Consensus model was as good as PLS in test validation. Moreover, both models were clearly acceptable based on their performance.

3.2.2. *Prediction of tensile strength*

In the case of tensile strength, we had 179 samples in the dataset, with reference values between 17.6 and 62.3 MPa. This exclusion was inevitable due to the insufficient number of composites above 62.3 MPa, as the model would have been biased and not accurate enough in the upper range of the tensile strength. Thus, we have used 136 samples for the training and cross-validation of the model out of the 179 samples.

We have used baseline correction with the 1st derivative for the PLS model and an additional SNV transformation for the NN model in the pre-scaling section, similarly to the modulus prediction. Additionally, we have applied standardization to the reference tensile strength values in both cases.

For PLS modeling, 950 variables were selected with the same settings as in the case of the modulus model based on genetic algorithm. The selected variables are shown in Fig. S4B in the Supplementary

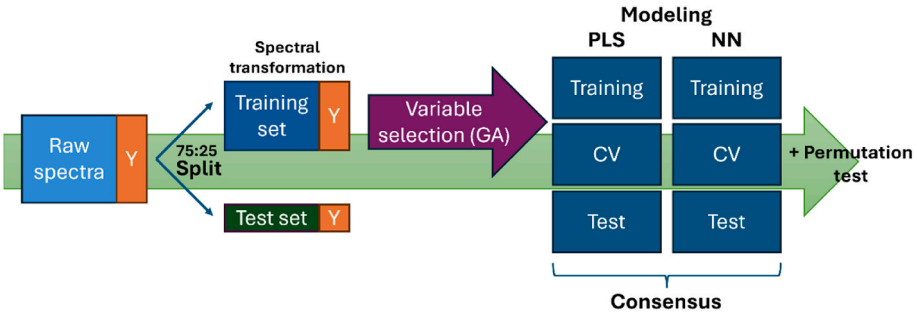

Fig. 2. The complete workflow of the regression models. GA = Genetic algorithm, CV = 5-fold randomized cross-validation, Y = reference measurements.

Table 5
Summary of the performance parameters in the case of modulus prediction. Top values are marked with green background and bold text.

	CV			Test		
Model	**PLS**	**NN**	**Consensus**	**PLS**	**NN**	**Consensus**
R^2	**0.95**	0.80	0.92	**0.96**	0.92	**0.96**
RMSE	**0.49**	0.96	0.60	**0.52**	0.75	0.55
MAE	**0.38**	0.59	0.39	0.37	0.53	**0.35**

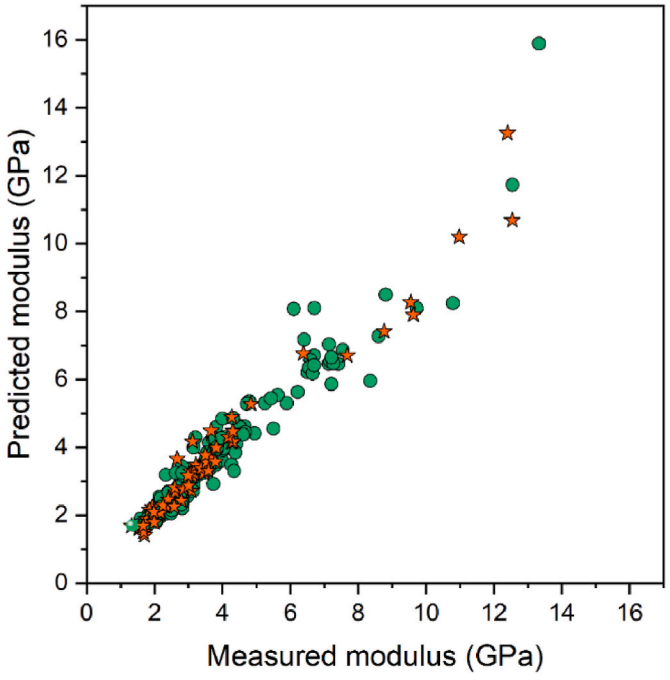

Fig. 3. Measured and predicted moduli are plotted against each other. Cross-validation samples are marked with green circles, while test validation samples are marked with orange stars.

information. Finally, fourteen LVs were applied in the PLS model according to the RMSECV values of cross-validation. Model performances are summarized together jointly with NN and the consensus model in Table 6. The model was successfully verified by the permutation test against the overfitting.

In the case of NN, we have used the previously selected 950 variables for modeling. Hyperparameter optimization selected the number of neurons in the three dense layers as 122, 962 and 679, respectively. The same repeated test validation protocol was used as in the case of moduli, and the test validation samples were the same as in the PLS model. This step can be very helpful to check the model stability. The permutation test showed that the model was not overfitted. The performance parameters are summarized in Table 6. In the last step, we have combined the PLS and NN models into a consensus model in the same way as for the prediction of moduli. These performance values are shown in Table 6 as well.

The measured and predicted tensile strength values based on the consensus model are plotted in Fig. 4.

Based on the performance parameters of the models and the measured vs. predicted plot of the consensus model, we can conclude that the consensus model outperformed the single variants in both CV and test validation. The model is acceptable for the prediction of the tensile strength of the polypropylene composites in the examined range based on ATR-IR spectra.

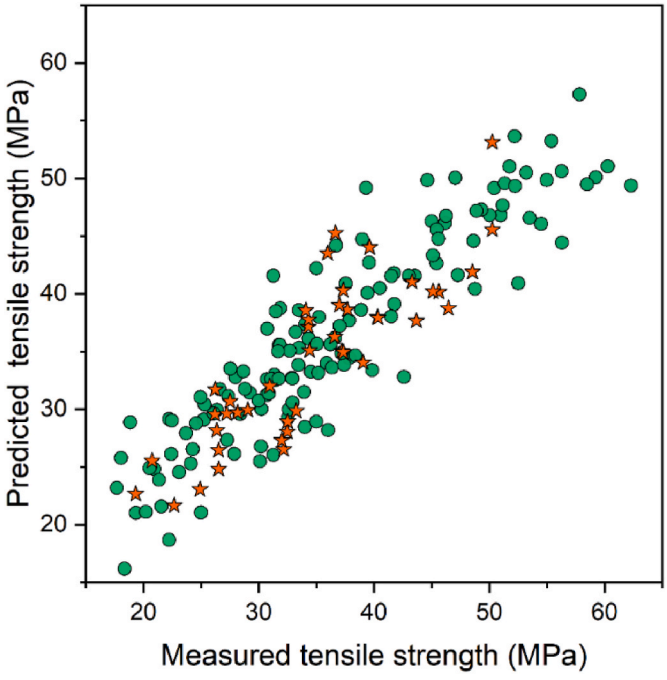

Fig. 4. Measured and predicted tensile strengths of the polymer composites. Cross-validation samples are marked with green circles, while test validation samples are marked with orange stars.

Table 6
Summary of the performance parameters for tensile strength prediction. Top values are marked with a green background and bold text.

	CV			Test		
Model	**PLS**	**NN**	**Consensus**	**PLS**	**NN**	**Consensus**
R^2	0.73	0.79	**0.82**	0.63	0.64	**0.70**
RMSE	5.53	4.95	**4.56**	4.73	4.65	**4.22**
MAE	4.33	3.72	**3.56**	3.89	3.66	**3.60**

3.2.3. Elongation at break

Finally, the spectral dataset of polymer composites was applied to the prediction of elongation at break. Here, we had 192 samples in total with appropriate reference values, and we have used 145 samples for the training and cross-validation of the models. The spectral pretreatment was a bit different compared to the other two cases: here, SNV transformation was used for both the NN and the PLS models. Thus, baseline correction along with the 1st derivative and SNV transformation was used for both modeling algorithms for the pre-scaling of the data. The range of the reference values was between 0.6 and 24.5 %.

For PLS modeling, we have selected 775 optimal variables out of the original set based on the genetic algorithm variable selection with the window size of 25 variables. The selected variables are shown in Fig. S4C in the Supplementary information. Eleven LVs were selected for the best PLS model with the RMSECV values of cross-validation. The test validation and permutation test verified that the model was robust enough and not overfitted.

For NN, the previously selected 775 variables were used as input data for modeling. The selected numbers of hidden neurons based on hyperparameter optimization were 686, 551 and 445. After the test validation and permutation test protocol (same as in the previous cases with the same sample set as for the PLS modeling part), the model was successfully verified against overfitting. The performance of the final PLS and NN models, along with the consensus model (which was calculated analogously to the previous cases) are shown in Table 7.

The measured and predicted elongation at break values based on the consensus model are plotted in Fig. 5.

The performance parameters and measured vs. predicted plot clearly show that the consensus model should be selected as the best one, because it has clearly outperformed the single models in every performance metric. The model is robust and accurate enough for the appropriate prediction of the elongation at break values of the polypropylene composites based on their ATR-IR spectra.

Our presented regression models for the prediction of modulus, tensile strength and elongation at break clearly supported the view that the use of consensus modeling has a great advantage especially in those cases, where i) we can detect nonlinearity in the relationship between the input and output data, ii) or the applied reference experiments can simply have larger errors (which can happen frequently, not just in material science), therefore the performance of the models are limited. Among the examined mechanical properties, the modulus is the least sensitive to micromechanical deformation: during its determination, the samples are subjected to very small forces, making local deformation processes less pronounced. This procedure reduces measurement errors, which can positively influence the accuracy of the prediction. Additionally, we have created the plots of the measured vs. predicted values for all three cases with the marking of the different reinforcement groups (see in Supplementary Information Fig. S5). Based on these figures, we can conclude that the mechanical properties of composites containing stronger reinforcing fibres exhibit higher values, although the accuracy of prediction was not different between these groups.

With these three models, we have shown that the physical characteristics of the samples can have an impact on light scattering, therefore it will also influence the spectral intensities in certain regions. These influences cannot be detected visually as well as the typical intensity peaks in an IR spectrum, but chemometrics can help us overcome this problem and easily find the important, but sometimes smaller details in the spectra. To the best of our knowledge, there is no other study in the literature which has found a proper correlation between the IR spectra of polymers and their mechanical properties. This can pave the way towards many indirect analyses, which have not been considered feasible in the past decades, but machine learning and combined techniques could help to reach this point.

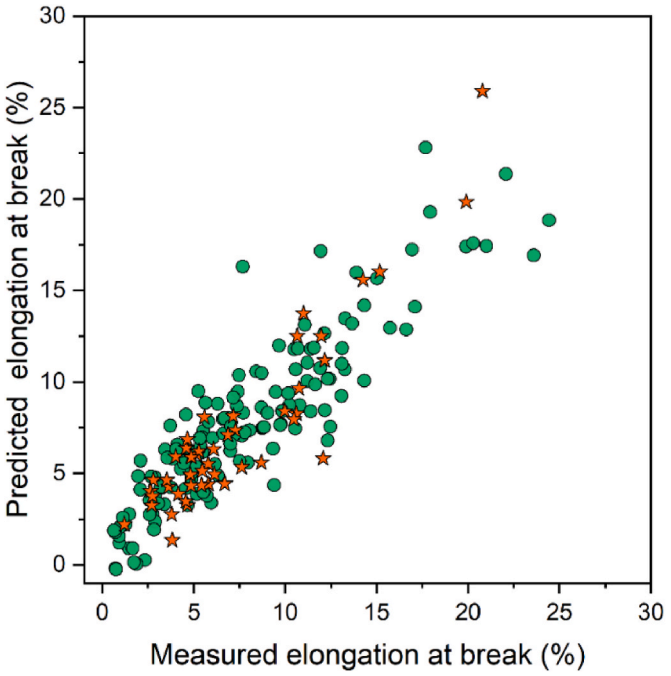

Fig. 5. Measured and predicted elongation at break of the composites. Cross-validation samples are marked with green circles, while test validation samples are marked with orange stars.

This study is a proof of concept that spectral data in material science are more valuable today than ever before, and they can be successfully combined with machine learning algorithms to find fast, easy and environmentally friendly alternatives to time-demanding experiments in many cases.

4. Conclusion

In this work, we have combined IR spectroscopy with machine learning tools for the determination of three mechanical characteristics of polypropylene composites. We have successfully classified the composite samples according to their reinforcements. Our XGBoost-based classification model has achieved a balanced accuracy greater than 0.9, even for the test samples. In the case of mechanical features, the three regression models combined the advantages of the classical PLS and the machine learning-based neural networks in a consensus

Table 7
Summary of the performance parameters in the case of elongation at break. Top values are marked with green background and bold text.

	CV			Test		
Model	**PLS**	**NN**	**Consensus**	**PLS**	**NN**	**Consensus**
R^2	0.67	0.78	**0.78**	0.68	0.58	**0.79**
RMSE	2.94	2.40	**2.40**	2.43	2.76	**1.97**
MAE	2.27	1.87	**1.84**	1.94	2.01	**1.49**

modeling workflow, which could outperform the single models in the prediction of tensile strength ($R^2_{Test} = 0.70$) and elongation at break ($R^2_{Test} = 0.79$). In the prediction of modulus values, the consensus model performed similarly to the single PLS model ($R^2_{Test} = 0.96$). In summary, our classification and regression models verified that the combination of IR spectroscopy with multivariate chemometric tools can open new directions in the qualitative and quantitative determination of the specific characteristics of the polymer composites.

CRediT authorship contribution statement

Szilvia Klébert: Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **Róbert Várdai:** Writing – original draft, Resources, Investigation, Data curation, Conceptualization. **Anita Rácz:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization.

Funding

The work was supported by project no. OTKA K 134260 of the Ministry of Innovation and Technology of Hungary from the National Research, Development, and Innovation Office, financed under the K type funding scheme. The work was supported by the Hungarian Academy of Sciences: János Bolyai Research Scholarship (A.R.).

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.

Acknowledgement

The authors thank to Dr. David Bajusz for the helpful discussions and language editing of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compscitech.2025.111127.

Data availability

I have shared the data as supplementary material.

References

[1] J. Wood, The top ten advances in materials science, Mater. Today 11 (2008) 40–45, https://doi.org/10.1016/S1369-7021(07)70351-6.
[2] S. Siengchin, A review on lightweight materials for defence applications: present and future developments, Def. Technol. 24 (2023) 1–17, https://doi.org/10.1016/j.dt.2023.02.025.
[3] G. Yang, M. Park, S.-J. Park, Recent progresses of fabrication and characterization of fibers-reinforced composites: a review, Compos. Commun. 14 (2019) 34–42, https://doi.org/10.1016/j.coco.2019.05.004.
[4] C. Baley, P. Davies, W. Troalen, A. Chamley, I. Dinham-Price, A. Marchandise, V. Keryvin, Sustainable polymer composite marine structures: developments and challenges, Prog. Mater. Sci. 145 (2024) 101307, https://doi.org/10.1016/j.pmatsci.2024.101307.
[5] G. Bayramoğlu, M. Mudu, Processing and characterization of polypropylene nanocomposite films reinforced with hexagonal boron nitride nanosheets, Period. Polytech. - Chem. Eng. 67 (2023) 94–103, https://doi.org/10.3311/PPch.20589.
[6] K.V. Balaji, K. Shirvanimoghaddam, R. Yadav, R. Mahmoodi, M.R. Ghandehari Ferdowsi, M. Naebe, Hybrid heterophasic polypropylene composites with basalt fibers and Magnesium oxysulfate reinforcements for sustainable automotive materials, J. Mater. Res. Technol. 28 (2024) 546–559, https://doi.org/10.1016/j.jmrt.2023.12.043.
[7] K. Ben Hamou, H. Kaddami, F. Elisabete, F. Erchiqui, Synergistic association of wood/hemp fibers reinforcements on mechanical, physical and thermal properties of polypropylene-based hybrid composites, Ind. Crops Prod. 192 (2023) 116052, https://doi.org/10.1016/j.indcrop.2022.116052.
[8] J. Karger-Kocsis, T. Bárány (Eds.), Polypropylene Handbook, Springer International Publishing, Cham, 2019, https://doi.org/10.1007/978-3-030-12903-3.
[9] R. Várdai, Polypropylene Hybrid Composites : Structure and Properties, Budapest University of Technology and Economics, 2021.
[10] H.Y. Seo, K.Y. Cho, D. Im, Y.J. Kwon, M. Shon, K.-Y. Baek, H.G. Yoon, High mechanical properties of covalently functionalized carbon fiber and polypropylene composites by enhanced interfacial adhesion derived from rationally designed polymer compatibilizers, Composites, Part B 228 (2022) 109439, https://doi.org/10.1016/j.compositesb.2021.109439.
[11] H. Demir, U. Atikler, D. Balköse, F. Tıhmınlıoğlu, The effect of fiber surface treatments on the tensile and water sorption properties of polypropylene–luffa fiber composites, Compos. Part A Appl. Sci. Manuf. 37 (2006) 447–456, https://doi.org/10.1016/j.compositesa.2005.05.036.
[12] H.A. Raslan, E.S. Fathy, R.M. Mohamed, Effect of gamma irradiation and fiber surface treatment on the properties of bagasse fiber-reinforced waste polypropylene composites, Int. J. Polym. Anal. Char. 23 (2018) 181–192, https://doi.org/10.1080/1023666X.2017.1405535.
[13] S.H. Han, H.J. Oh, S.S. Kim, Evaluation of fiber surface treatment on the interfacial behavior of carbon fiber-reinforced polypropylene composites, Composites, Part B 60 (2014) 98–105, https://doi.org/10.1016/j.compositesb.2013.12.069.
[14] L. Dányádi, T. Janecska, Z. Szabó, G. Nagy, J. Móczó, B. Pukánszky, Wood flour filled PP composites: compatibilization and adhesion, Compos. Sci. Technol. 67 (2007) 2838–2846, https://doi.org/10.1016/j.compscitech.2007.01.024.
[15] K. Shirvanimoghaddam, K.V. Balaji, R. Yadav, O. Zabihi, M. Ahmadi, P. Adetunji, M. Naebe, Balancing the toughness and strength in polypropylene composites, Composites, Part B 223 (2021) 109121, https://doi.org/10.1016/j.compositesb.2021.109121.
[16] J.E. Stamhuis, Mechanical properties and morphology of polypropylene composites II. Effect of polar components in talc-filled polypropylene, Polym. Compos. 9 (1988) 72–77, https://doi.org/10.1002/pc.750090110.
[17] M. Fasihi, H. Mansouri, Effect of rubber interparticle distance distribution on toughening behavior of thermoplastic polyolefin elastomer toughened polypropylene, J. Appl. Polym. Sci. 133 (2016), https://doi.org/10.1002/app.44068.
[18] G. Keledi, A. Sudar, C. Burgstaller, K. Renner, J. Moczo, B. Pukanszky, Tensile and impact properties of three-component PP/wood/elastomer composites, Express Polym. Lett. 6 (2012) 224–236, https://doi.org/10.3144/expresspolymlett.2012.25.
[19] R. Vardai, T. Lummerstorfer, C. Pretschuh, M. Jerabek, M. Gahleitner, B. Pukanszky, K. Renner, Impact modification of PP/wood composites: a new approach using hybrid fibers, Express Polym. Lett. 13 (2019) 223–234, https://doi.org/10.3144/expresspolymlett.2019.19.
[20] S.S. Sorour, C.A. Saleh, M. Shazly, A review on machine learning implementation for predicting and optimizing the mechanical behaviour of laminated fiber-reinforced polymer composites, Heliyon 10 (2024) e33681, https://doi.org/10.1016/j.heliyon.2024.e33681.
[21] V. Sharma, M.-L. Tsai, C.-W. Chen, P.-P. Sun, P. Nargotra, C.-D. Dong, Advances in machine learning technology for sustainable biofuel production systems in lignocellulosic biorefineries, Sci. Total Environ. 886 (2023) 163972, https://doi.org/10.1016/j.scitotenv.2023.163972.
[22] G.X. Gu, C.-T. Chen, M.J. Buehler, De novo composite design based on machine learning algorithm, Extrem. Mech. Lett. 18 (2018) 19–28, https://doi.org/10.1016/j.eml.2017.10.001.
[23] G. Ammasai Sengodan, Prediction of two-phase composite microstructure properties through deep learning of reduced dimensional structure-response data, Composites, Part B 225 (2021) 109282, https://doi.org/10.1016/j.compositesb.2021.109282.
[24] A. Rácz, D. Tátraaljai, S. Klébert, Determination of melt flow index and polymer additives in polyethylene based on IR spectra and multivariate modeling, Mater. Today Chem. 33 (2023) 101671, https://doi.org/10.1016/j.mtchem.2023.101671.
[25] S.K. Kamane, N.K. Patil, B.R. Patagundi, Use of artificial neural network to predict the bending behavior of steel I beam externally attached with FRP sheets, Mater. Today Proc. 39 (2021) 17–21, https://doi.org/10.1016/j.matpr.2020.04.893.
[26] B.R.S. Reddy, M. Premasudha, B.B. Panigrahi, K. Cho, N.G.S. Reddy, Modeling <scp>constituent–property</scp> relationship of polyvinylchloride composites by neural networks, Polym. Compos. 41 (2020) 3208–3217, https://doi.org/10.1002/pc.25612.
[27] B. Liu, N. Vu-Bac, T. Rabczuk, A stochastic multiscale method for the prediction of the thermal conductivity of Polymer nanocomposites through hybrid machine learning algorithms, Compos. Struct. 273 (2021) 114269, https://doi.org/10.1016/j.compstruct.2021.114269.
[28] S.M. Khan, S.A. Malik, N. Gull, S. Saleemi, A. Islam, M.T.Z. Butt, Fabrication and modelling of the macro-mechanical properties of cross-ply laminated fibre-reinforced polymer composites using artificial neural network, Adv. Compos. Mater. 28 (2019) 409–423, https://doi.org/10.1080/09243046.2019.1573448.
[29] M.S. Kabbani, H.A. El Kadi, Predicting the effect of cooling rate on the mechanical properties of glass fiber–polypropylene composites using artificial neural networks, J. Thermoplast. Compos. Mater. 32 (2019) 1268–1281, https://doi.org/10.1177/0892705718792351.
[30] T. Nguyen, A. Arabi Shamsabadi, M. Bavarian, Coupling ATR-FTIR spectroscopy with multivariate analysis for polymers manufacturing and control of polymers' molecular weight, Digit, Chem. Eng. 7 (2023) 100089, https://doi.org/10.1016/j.dche.2023.100089.
[31] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[32] T. Chen, C. Guestrin, XGBoost, in: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, New York, NY, USA, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[33] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, Knime - the Konstanz information miner, ACM SIGKDD Explor, Newsl. 11 (2009) 26–31, https://doi.org/10.1145/1656274.1656280.
[34] A. Rácz, D. Bajusz, K. Héberger, Multi-level comparison of machine learning classifiers and their performance metrics, Molecules 24 (2019) 1–18, https://doi.org/10.3390/molecules24152811.
[35] R. Hecht-Nielsen, Theory of the backpropagation neural network, in: Neural Networks Percept, Elsevier, 1992, pp. 65–93, https://doi.org/10.1016/B978-0-12-741252-8.50010-8.
[36] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proc. Natl. Acad. Sci. USA 116 (2019) 15849–15854, https://doi.org/10.1073/pnas.1903070116.
[37] R. Leardi, A. Lupiáñez González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometr. Intell. Lab. Syst. 41 (1998) 195–207, https://doi.org/10.1016/S0169-7439(98)00051-3.
[38] C.M. Andersen, R. Bro, Variable selection in regression-a tutorial, J. Chemom. 24 (2010) 728–737, https://doi.org/10.1002/cem.v24.11/12.