scientific reports

OPEN

Efficient and accurate determination of the degree of substitution of cellulose acetate using ATR-FTIR spectroscopy and machine learning

Frank Rhein[1]✉, Timo Sehn[2] & Michael A. R. Meier[2,3]✉

Multiple linear regression models were trained to predict the degree of substitution (DS) of cellulose acetate based on raw infrared (IR) spectroscopic data. A repeated k-fold cross validation ensured unbiased assessment of model accuracy. Using the DS obtained from ^{1}H NMR data as reference, the machine learning model achieved a mean absolute error (MAE) of 0.069 in DS on test data, demonstrating higher accuracy compared to the manual evaluation based on peak integration. Limiting the model to physically relevant areas unexpectedly showed the $\mathrm{C-H}$ peak to be the strongest predictor of DS. By applying a n-best feature selection algorithm based on the F-statistic of the Pearson correlation coefficient, several relevant areas were identified and the optimized model achieved an improved MAE of 0.052. Predicting the DS of other cellulose acetate data sets yielded similar accuracy, demonstrating that the developed models are robust and suitable for efficient and accurate routine evaluations. The model solely trained on cellulose acetate was further able to predict the DS of other cellulose esters with an accuracy of $\approx 0.1 - 0.2$ in DS and model architectures for a more general analysis of cellulose esters were proposed.

Keywords Machine learning, Degree of substitution, Infrared spectroscopy, Cellulose ester, Cellulose acetate

Fossil resource depletion combined with the increased pollution of our environment are the main reasons for a change towards biobased polymeric materials. Cellulose, as the most abundant biopolymer on earth (1.5×10^{12} tons p.a.)[1,2], plays a key role in establishing biobased alternatives to fossil based polymers due to its unique properties. Besides biocompatibility and biodegradability, cellulose also possesses high thermal and mechanical resistance, both being important requirements for potential industrial applications[3,4]. Up to the present day, cellulose esters (CEs) are the most important cellulose derivatives from an industrial point of view[5]. The synthesis of CEs can be performed according to several synthetic procedures in a homogeneous[1,4,6–8] or heterogeneous fashion[9–12]. The material properties and processability of CEs depend on the applied synthesis procedures as well as on their structural composition, i.e., the nature of the ester and the degree of substitution (DS)[6,13]. The DS is defined as the average number of substituted hydroxyl groups per anhydroglucose (AGU) unit and can adopt values between 0 and 3. CEs find applications in coatings, drug delivery, food packaging, membrane and fiber industry, or bio-medicine[3,4,14,15]. Especially cellulose acetate (CA), which is most commonly known as the material used for cigarette filters, can be recognized as the CE with the highest industrial interest so far[5].

The material properties of CA can be adjusted by varying the DS. Thus, it is essential to have reliable, simple and fast analytical tools for the determination of the DS of CAs and other CEs. The most common method for DS determination, which is also recognized as the most time- and labor-intensive one, is the hydrolysis of CEs in an alkaline medium and subsequent titration[8,16]. Additionally, this method suffers from a large sample demand and is also categorized as a destructive analysis method. A more straightforward non-destructive alternative for DS determination of CEs based on nuclear magnetic resonance spectroscopy (NMR) are ^{1}H-

[1]Institute of Mechanical Process Engineering and Mechanics (MVM), Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany. [2]Institute of Biological and Chemical Systems – Functional Molecular Systems (IBCS-FMS), Karlsruhe Institute of Technology (KIT), Karlsruhe 76344, Germany. [3]Institute of Organic Chemistry (IOC), Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany. ✉email: frank.rhein@kit.edu; m.a.r.meier@kit.edu

and ^{13}C-NMR[1,14,15,17–20]. These approaches are limited by the solubility of the CEs in solvents suitable for NMR analysis. This is especially challenging for CEs with low DS, and requires the synthesis of more elaborate solvent systems and adjusted NMR protocols[21]. A supplemental requirement for accurate DS determination via ^{1}H- and ^{13}C-NMR is that the magnetic resonances of the introduced substituents do not overlap with the signals of the anhydro glucuse units (AGU). King et al.[22] introduced a more advanced technique that involves the derivatization of unmodified hydroxyl groups with a phosphorous reagent and a subsequent quantitative ^{31}P NMR evaluation. Other analytical techniques, such as gas chromatography[23], elemental analysis[24,25] or UV–Vis spectroscopy[26] are reported, but limited with respect to precision and applicability. Wolfs et al.[27] reported a supplemental and more convenient DS determination for CEs via attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy by simple integration of the $\mathrm{C{=}O}$, $\mathrm{C{-}O_{ester}}$, and $\mathrm{C{-}H}$ stretching vibration bands. After fitting of calibration curves, DS values could be determined with an average relative error between 5.5 % and 7.3 % depending on the evaluated peaks. These values are similar to the reported accuracy of the ^{1}H NMR method and—compared to the labor-intensive NMR-based characterization methods—this approach does not require elaborate sample preparation and measurements are very fast (≈ 30 s). Nevertheless, as this DS determination via ATR-FTIR still requires the manual selection and integration of the corresponding peaks in the spectra, this study raises the question whether the procedure can be automated, streamlined and/or enhanced by applying machine learning techniques.

Machine learning (ML) is a branch of artificial intelligence that aims at inferring solutions to problems by applying statistical methods to data instead of explicitly programming the solution. Regression is a subset of ML tasks and describes the prediction of dependent variables (outcomes) based on independent variables (inputs, features). Different algorithms can be used to achieve this, including multiple linear regression (MLR) or neural networks (NN). The adjustment (optimization) of internal parameters to the specific data set is referred to as learning and the obtained parameter set in conjunction with the algorithm is considered a model. If learning is based on provided, static data of input-output pairs, this is referred to as supervised learning.

ML, although currently an omnipresent term, is by no means a new idea and has been tied to spectroscopy from the very beginning. It can be argued that one of the earliest examples is the well-known work of Beer[28], who—although without computers at the time—fitted extinction coefficients of a linear model to experimental observations. More contemporary examples, actually based on algorithmic learning without human evaluation, are often summarized under the term chemometrics[29] and are applied to various spectroscopic measurements[30], like Raman spectroscopy[31,32], UV/Vis absorption spectroscopy[33,34], NMR spectroscopy[35,36], or IR spectroscopy[37,38]. The combination of IR and regression analysis was applied to determine the degree of oxidation of dialdehyde cellulose[39] and the DS of both carboxymethyl starch[40] and methylesterified pectic polysaccharides[41]. The DS of CA specifically, was accurate determined by regression analysis of FT Raman spectra[42]. Further, ML was applied to determine the cellulose and lignin content of lignocellulose originating from different sources with high accuracy[43]. ML techniques applied to NIR measurements were able to accurately predict cellulose pulp dryness in real time[44], and the hemicellulose, cellulose and lignin content of moso bamboo[45].

The aim of this work is to find accurate ML models for predicting the DS of CA based on raw IR spectra, which has not been reported before. Such a model would be fast and not require any manual data processing or evaluation, two properties that are highly relevant for quality control in industry and unbiased analytical procedures in academia. Multiple linear regression models are trained, while a repeated k-fold cross-validation allows for an unbiased evaluation of model accuracy. The results are compared to the values obtained from ^{31}P NMR, ^{1}H NMR and the integration-based method proposed by Wolfs et al.[27] to assess accuracy. However, this study is not solely concerned with predicting DS values. Features (wavenumbers) are selected based on physical knowledge and algorithms to reveal the most relevant areas of the spectra, i.e. areas that contain most information. Finally, an optimized model is applied to CA data of different experimenters and different synthesis routes as well as to IR spectra of other CEs to quantify the robustness and extrapolation capabilities of the model. This has not been reported before and paves the way for more generalized analytical models.

Data overview

All data sets used in this study were obtained by us in our laboratory and have been previously published[1,6,27]. Table 1 provides an overview of the different data sets with corresponding reference, number of data points N, DS-range and additional details. Data sets are marked with monospace font and named after the first author (e.g. `Wolfs.A`). All data sets contain ^{1}H NMR spectroscopy and IR spectroscopy data. The reference set `Wolfs.A` additionally contains ^{31}P NMR spectroscopy data for $\mathrm{DS} > 1.21$ (insolubility below). DS values were

Name	Refs.	N	$\mathrm{DS}_{1\mathrm{H}}$-range	Details
`Wolfs.A`	[27]	16	$0.41 - 2.96$	Cellulose acetate, reference data set[a].
`Wolfs.B`	[1]	5	$1.04 - 2.5$	Cellulose acetate, same chemistry[a], partially different experimenter
`Sehn.A`	[6]	7	$0.80 - 2.34$	Cellulose acetate, different chemistry[b], different experimenter
`Sehn.B`	[6]	5	$2.05 - 2.17$	Cellulose esters of different chain lengths with DS $\approx$ 2, different chemistry[b], different experimenter
`Sehn.C`	[6]	5	$1.99 - 2.16$	Cellulose esters of different chain lengths (mixtures) with DS $\approx$ 2, different chemistry[b], different experimenter

Table 1. Data overview. [a] Modification of MCC with vinyl acetate in a DMSO/DBU/CO_2 switchable solvent system at 60 °C (conventional heating) within 6 h (see supporting information). [b] Modification of MCC with vinyl acetate in a DMSO/TMG/CO_2 switchable solvent system at 140 °C (microwave heating) within 10 min (see supporting information)

determined in various ways: $\mathrm{DS}_{1\mathrm{H}}$ and $\mathrm{DS}_{31\mathrm{P}}$ are obtained by applying the routine described in the methods section and[22] to ^{1}H NMR and ^{31}P NMR data, respectively. $\mathrm{DS}_{\mathrm{IR,int}}$ is obtained via peak integration of IR data and corresponds to $\mathrm{DS}_{\mathrm{IR,mean}}$ of[27]. $\mathrm{DS}_{\mathrm{IR,ML}}$ is obtained by the ML evaluation methods proposed in this study. More information on the sample preparation and analyses is found in the supporting information for this article or the respective publications. All compiled data is provided in open access[46].

Results and discussion

^{31}P NMR versus ^{1}H NMR versus IR integration accuracy based on data set `Wolfs.A`

The obtained $\mathrm{DS}_{1\mathrm{H}}$ values based on the described integration routine and reported $\mathrm{DS}_{\mathrm{IR,int}}$ values are compared to the reported $\mathrm{DS}_{31\mathrm{P}}$ values of `Wolfs.A`. It is assumed that ^{31}P NMR provides the most accurate DS values. Please note that only data points with $\mathrm{DS} > 1.21$ can be considered due to insolubility of the samples below this DS value. ^{1}H NMR data achieved a better MAE of 0.032 ($\mathrm{MRE} = 1.53\%$), while IR integration reported previously showed a MAE of 0.078 ($\mathrm{MRE} = 3.64\%$). The results are visualized in Fig. 1.

To extend the study also to DS values smaller than 1.21, all following investigations are performed with $\mathrm{DS}_{1\mathrm{H}}$ values as ground truth, i.e. as training data. Based on Fig. 1 and considering that ^{31}P NMR data is also susceptible to experimental errors, this assumption appears justified.

ML prediciton accuracy on data set `Wolfs.A`

Baseline accuracy

A repeated *k*-fold cross validation (CV, see methods section) was performed with randomized data assignment between repetitions on the $\mathrm{DS}_{1\mathrm{H}}$ data of `Wolfs.A`. A multiple linear regression (MLR) model is used with $N = 1000$ repetitions and $k = 8$ folds, meaning that roughly $12.5\,\%$ of the data points were reserved for testing in each iteration. The results are shown in Fig. 2. A $\mathrm{MAE}_{\mathrm{IR,ML}}$ of 0.069 ($\mathrm{MRE}_{\mathrm{IR,ML}} = 5.73\%$) is obtained. Additionally, the $\mathrm{DS}_{\mathrm{IR,int}}$ values are plotted against the $\mathrm{DS}_{1\mathrm{H}}$ values. A $\mathrm{MAE}_{\mathrm{IR,int}}$ of 0.083 ($\mathrm{MRE}_{\mathrm{IR,int}} = 5.34\%$) is obtained. Hence, the ML evaluation achieved a slightly smaller MAE and slightly larger MRE than the manual IR integration method, although both evaluation methods are precise and achieve $\mathrm{MAE} < 0.1$. This immediately demonstrates that a simple linear regression model is able to make good predictions on the DS based on provided raw IR spectra. It should be highlighted that only the prediction accuracy on the test data, i.e. unseen data that has not been used in training, is shown, while for IR integration data, no test/train split was performed. It is apparent that the integration method consistently overpredicts the DS values, while the ML evaluation appears less biased. The opposing behavior in MAE and MRE can be explained by the large prediction error of the ML model at the relatively small $\mathrm{DS}_{1\mathrm{H}}$ value of 0.70, resulting in a larger relative error.

Influence of k-fold

To investigate the influence of the amount of training data, the parameter *k* was varied between $k \in \{2, \ldots, 16\}$. These values correspond to $\{50\,\%, \ldots, 6.25\,\%\}$ of the data being used for testing, i.e. higher k values result in less data reserved for the test set and more data being used for training. The results are given in Fig. 3 as a box plot. The box spans from the first (Q1) to the third quartile (Q3) and contains the central $50\,\%$ of the data, while

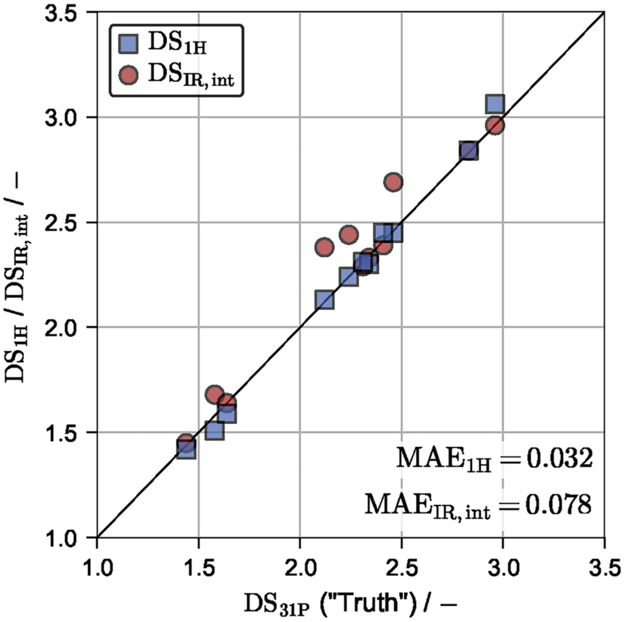

Fig. 1. $\mathrm{DS}_{31\mathrm{P}}$ values versus $\mathrm{DS}_{1\mathrm{H}}$ and $\mathrm{DS}_{\mathrm{IR,int}}$ for the `Wolfs.A` data set. Only $\mathrm{DS} > 1.21$ values are considered due to insolubility of the samples below this value.

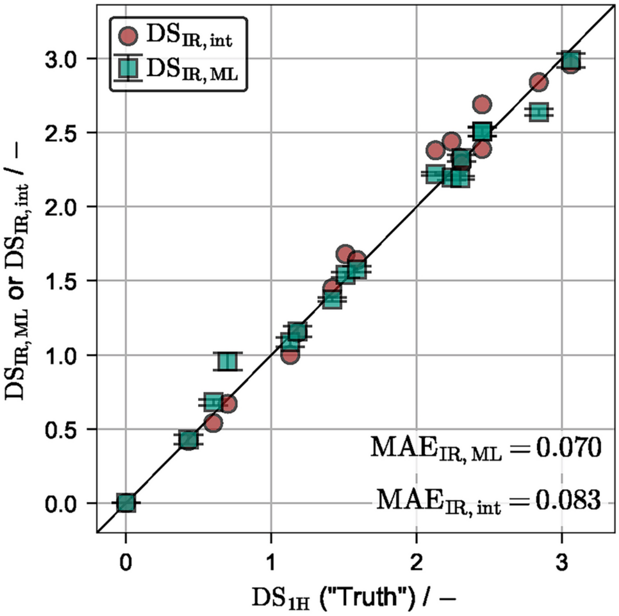

Fig. 2. DS_{1H} vs. $\mathrm{DS}_{IR,ML}$ and $\mathrm{DS}_{IR,int}$ values for the `Wolfs.A` data set. The mean values and corresponding standard deviation of all test data points from *N* repetitions and *k* folds are shown.

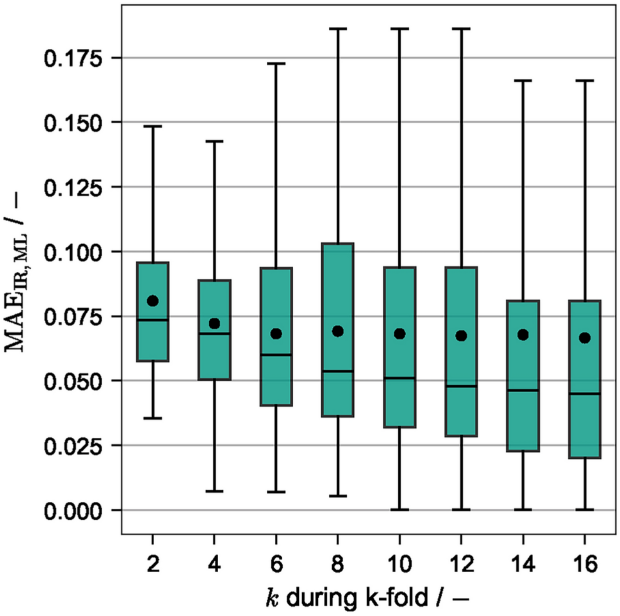

Fig. 3. $\mathrm{MAE}_{IR,ML}$ for varying *k* during k-fold. $k \in \{2, 4, 8, 16\}$ corresponds to $50\,\%$, $25\,\%$, $12.5\,\%$ and $6.25\,\%$ of the data being used for testing.

the whiskers extend to the farthest data point lying within 1.5 times the inter-quartile range (IQR). The median value is indicated by a horizontal line and the mean value by a point inside the box.

Increasing *k*, i.e. increasing the amount of data being used for training versus testing, lowers both the mean and median MAE across all $N \times k$ folds. However, even for $k = 2$, i.e. only $50\,\%$ of the data being used for training, an $\mathrm{MAE} < 0.1$ is achieved in the majority of folds. This shows that the linear regression model is robust with respect to the required amount of training data and is not prone to overfitting. The span of the whiskers are quite large, which means that the MAE exceeds 0.1 in some folds and predictions are less accurate than the IR-based method. However, these cases are retraced to extreme test/train splits, like e.g. only training for small DS, while predicting only large DS values. Please note that this is only due to the applied cross validation with randomized test/train splits and will not be relevant for production models, i.e. routine evaluation. Here, either all or sensibly selected data is used for training, effectively eliminating these extreme cases.

Influence of wavenumeber ranges and feature selection

To evaluate if and how different parts of the spectrum hold more or less information relevant for model training, model training was performed on specific wavenumber ranges, while any data outside the specific range was omitted. Fig. 4c visualizes the investigated ranges together with an exemplary spectrum: A non-fingerprint region is defined for $\nu > 1500\,\mathrm{cm}^{-1}$ and a fingerprint region for $\nu < 1500\,\mathrm{cm}^{-1}$. Additionally, peak specific ranges for $\mathrm{C=O}$ ($1600\,\mathrm{cm}^{-1} < \nu < 1800\,\mathrm{cm}^{-1}$), $\mathrm{C-H}$ ($1325\,\mathrm{cm}^{-1} < \nu < 1425\,\mathrm{cm}^{-1}$), $\mathrm{C-O}$ ($1150\,\mathrm{cm}^{-1} < \nu < 1300\,\mathrm{cm}^{-1}$) and a combined range $\mathrm{C=O, C-H, C-O}$ ($1150\,\mathrm{cm}^{-1} < \nu < 1800\,\mathrm{cm}^{-1}$) are defined. For reference, both the full range and a randomly selected area with no apparent information ($2000\,\mathrm{cm}^{-1} < \nu < 2500\,\mathrm{cm}^{-1}$) are investigated. For each area, a repeated CV was performed with $N = 1000$ and $k = 8$. The resulting MAE values are given as box-plot in Fig. 4a and compared to the integration-based result from above indicted by a horizontal line.

As expected, the randomly selected area was not able to yield reliable predictions of the DS. Similarly, predicting the DS solely based on the $\mathrm{C=O}$ or $\mathrm{C-O}$ peaks produced large deviations and is less accurate than the integration-based method. The best predictions were obtained for the full and $\mathrm{C-H}$ range, followed by the fingerprint region. These results are surprising in the sense that Wolfs et al.[27] reported highest accuracy for evaluation of the $\mathrm{C=O}$, followed by the $\mathrm{C-O}$ and lastly the $\mathrm{C-H}$ peak. These differences can be attributed to the different model architecture used, with the $\mathrm{C-H}$ peak being more suited towards a linear regression. Fig. 4a highlights the counter-intuitive fact that manually selecting specific wavenumber ranges, a routinely applied procedure in academia and industry, does not necessarily improve model accuracy. In fact, it might lead to drastically worse predictions.

However, the search for the region that contains most information and yields the best predictions can also be performed automatically via an n-best feature selection algorithm described in the methods section. A parameter study was performed varying n between $n \in [50, 1750]$ in steps of 100 performing a repeated CV for each with $N = 1000$ and $k = 8$. The resulting MAE values are shown as box plots in Fig. 4b. The full data

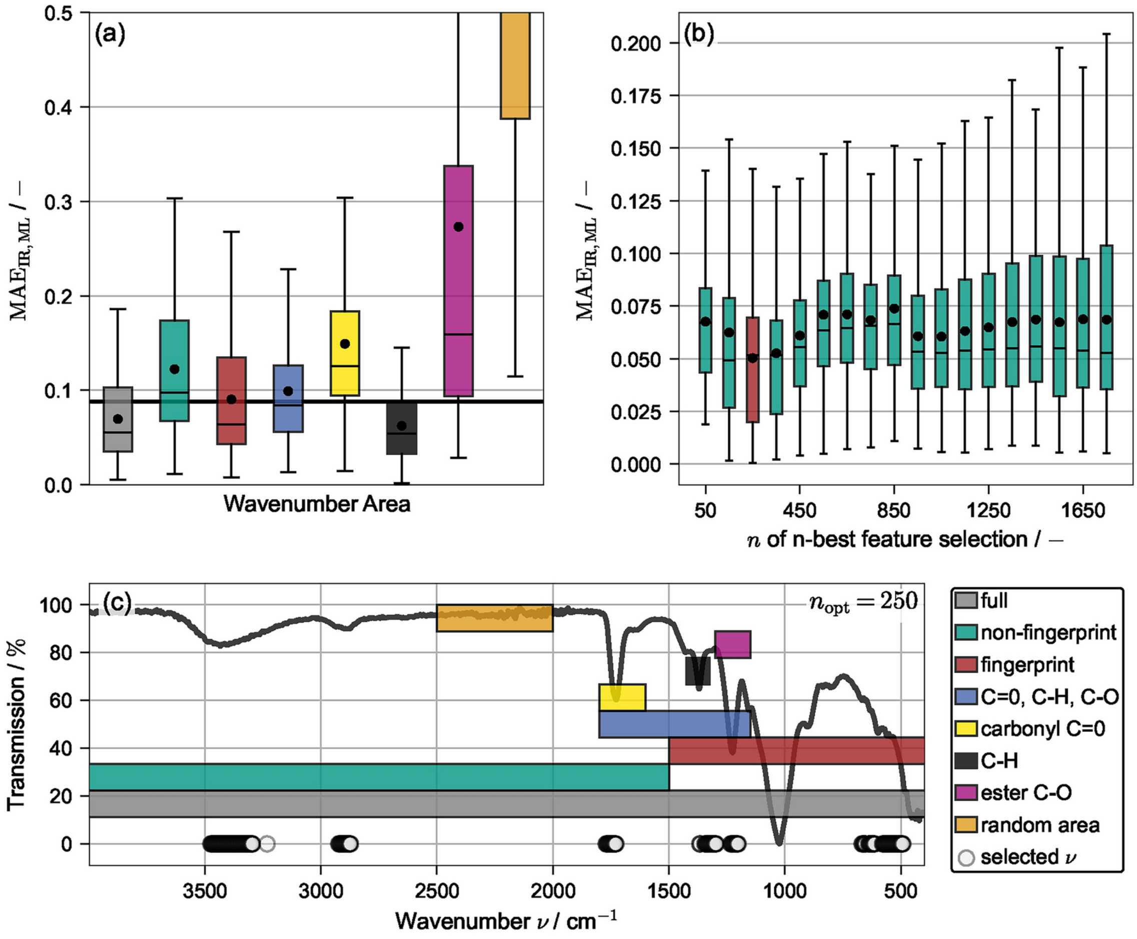

Fig. 4. (**a**) Box-plots of $\mathrm{MAE}_{\mathrm{IR,ML}}$ for varying wavenumber ranges viszualized in (**c**). (**b**) Box-plots of $\mathrm{MAE}_{\mathrm{IR,ML}}$ for varying n of n-best feature selection. The optimum value ($n = 250$) is colored in red and the selected wavelengths are visualized in (**c**).

set contains 1750 features and is shown on the far right. Reducing the number of features reduces both the MAE and its spread until $n = 950$, while further reduction initially increases the prediction error. Reducing n further yields a global minimum at $n_{\mathrm{opt}} = 250$, where $\mathrm{MAE}_{\mathrm{opt}} = 0.052$ is achieved. This represents a significant increase in prediction accuracy compared to the integration-based evaluation method. Increasing prediction accuracy by limiting the amount of features, i.e. information, seems counter-intuitive, however, this can be explained by a reduced amount of overfitting. The initially present 1750 features contain plenty of wavenumbers that do not store information on the DS. Excluding these from training makes the resulting models more robust and more accurate on unseen data. Vividly, the model is forced to learn the underlying relationships rather than experiment-specific random deviations. Fig. 4c shows the selected wavenumbers as scattered points at $T = 0$. Besides the expected peaks for $\mathrm{C = O}$, $\mathrm{C - H}$ and $\mathrm{C - O}$, the model also evaluates in the range $3250\,\mathrm{cm}^{-1} < \nu < 3500\,\mathrm{cm}^{-1}$, $2850\,\mathrm{cm}^{-1} < \nu < 2950\,\mathrm{cm}^{-1}$ and, most surprisingly, in the range $500\,\mathrm{cm}^{-1} < \nu < 650\,\mathrm{cm}^{-1}$ of the fingerprint region. For conventional, i.e. manual ATR-FTIR characterization of organic and polymeric compounds (including cellulose derivatives), especially vibrational bands at higher wavenumbers ($\geq 1500\,\mathrm{cm}^{-1}$) outside the fingerprint region are essential. Particularly, the finger print region is not frequently examined in data evaluation due to its complexity. The ML approach thus seems to provide benefits for data analysis a researcher would not easily identify. Hence, not only did the n-best feature selection increase the overall prediction accuracy, it simultaneously returned relevant information about where information on the DS is stored in the IR spectrum. It should be noted that only raw IR data was used for model training and the feature select algorithm identified the chemically relevant areas without additional mechanistic information or guidance.

Application to other data sets

To assess the applicability of the ML evaluation method, i.e. the extrapolation to other data sets, two linear regression models were trained on the entire `Wolfs.A` data set. No cross validation is required as all points are included in training. One model is trained on the optimum amount of features found in the previous section ($n = 250$) and one is trained on the entire range of wavelengths. Both models are used to predict the $\mathrm{DS}_{\mathrm{IR,ML}}$ for samples contained in the `Wolfs.B`, `Sehn.A`, `Sehn.B` and `Sehn.C` data sets, which are compared to the respective $\mathrm{DS}_{1\mathrm{H}}$ values. Fig. 5a shows the results for the model with feature selection and Fig. 5b for the one without. The resulting MAE and MRE values are given in Tab. 2.

Both `Wolfs.B` and `Sehn.A` are predicted with a high accuracy, demonstrating that the MLR model is robust and can be applied to IR data of different experimenters (`Wolfs.B`) and even different synthesis routes (`Sehn.A`). Additionally, the selected features seem to be universal for CA, as using the optimized model with feature select results in overall lower errors. These results show that routine evaluations of the DS of CAs can be performed without the need for laborious and time-consuming ^{1}H or even ^{31}P NMR measurements and without the need for manual peak integration and evaluation of IR data as was done by Wolfs et al.[27]. It should be noted that all IR data was measured on the same instrument and future studies should include the influence of instrument-specific deviations on the robustness of the evaluation. However, the applied normalization in Eq. 1 alleviates this effect as long as the relative intensities between wavenumbers stay similar.

Applying the trained model to `Sehn.B` and `Sehn.C`, i.e. to other cellulose esters than CA, should be regarded as a large extrapolation task, as no IR data of different CEs was included during training. However, Fig. 5b shows that when the MLR model has access to all wavenumbers it is indeed capable of predicting the DS to an accuracy of $\approx 0.1 - 0.2$, with only cellulose hexanoate of `Sehn.B` being predicted poorly ($\Delta\mathrm{DS} = 0.62$). This shows that to some degree, the DS influences the IR spectra of different CEs quite similarly and that a machine learning model is capable of learning these effects. However, the model with feature select shown in Fig. 5a is unable to predict other esters, with the MAE even exceeding 1 for `Sehn.B`. To discuss this drastically different behavior, exemplary IR spectra of all data sets are visualized in Fig. 5c. The samples closest to $\mathrm{DS}_{1\mathrm{H}} = 2$ were selected, meaning that for the linear model to yield similar predictions, all curves need to be similar. Again the selected wavenumbers ($n = 250$) are indicated with scattered points. In general, all spectra are similar, especially in the expected ranges for the $\mathrm{C = O}$, $\mathrm{C - H}$ and $\mathrm{C - O}$ peaks, but also at $\nu \approx 3500\,\mathrm{cm}^{-1}$. However, both the peak at $\nu \approx 2900\,\mathrm{cm}^{-1}$ and the fingerprint region differ quite strongly for different CEs. As these regions contain many of the selected wavenumbers, it is not surprising that the reduced model has difficulties in predicting the DS. By not applying any feature selection, these areas are less important and prediction accuracy is increased for these different CEs.

It should be noted that the objective of the trained model was not to predict the DS of CEs other than CA and that it lacked the required data to accurately do so. Nevertheless, Fig. 5c suggests that this may indeed be possible with enough relevant training data. Two adjusted model structures can be conceptualized: First, a combined model with two output parameters—one being the DS and one being e.g. the chain length of the ester or smiles string of the substituent—could be trained. Secondly, it is sensible to train multiple ester-specific models for prediction of the DS and combine them with a superordinate classification model, which predicts the type of ester and hence, which model to apply.

A note on model architecture

This work did not apply other model architectures than the MLR, although in the contemporary machine learning trend, neural networks (NN) are often displayed as somehow superior. It is important to stress that NN are only one of many different model architectures and that all have specific benefits and downsides. The MLR achieved a prediction accuracy close to ^{1}H-based evaluation, especially after feature selection. Considering that $\mathrm{DS}_{1\mathrm{H}}$ values also contain experimental errors, it is argued that a NN has not much room for improvement and even if slightly lower numbers were obtained, this might not represent an improvement with respect to the true DS. Therefore, if multiple models do not differ significantly in accuracy, one should always choose the one with

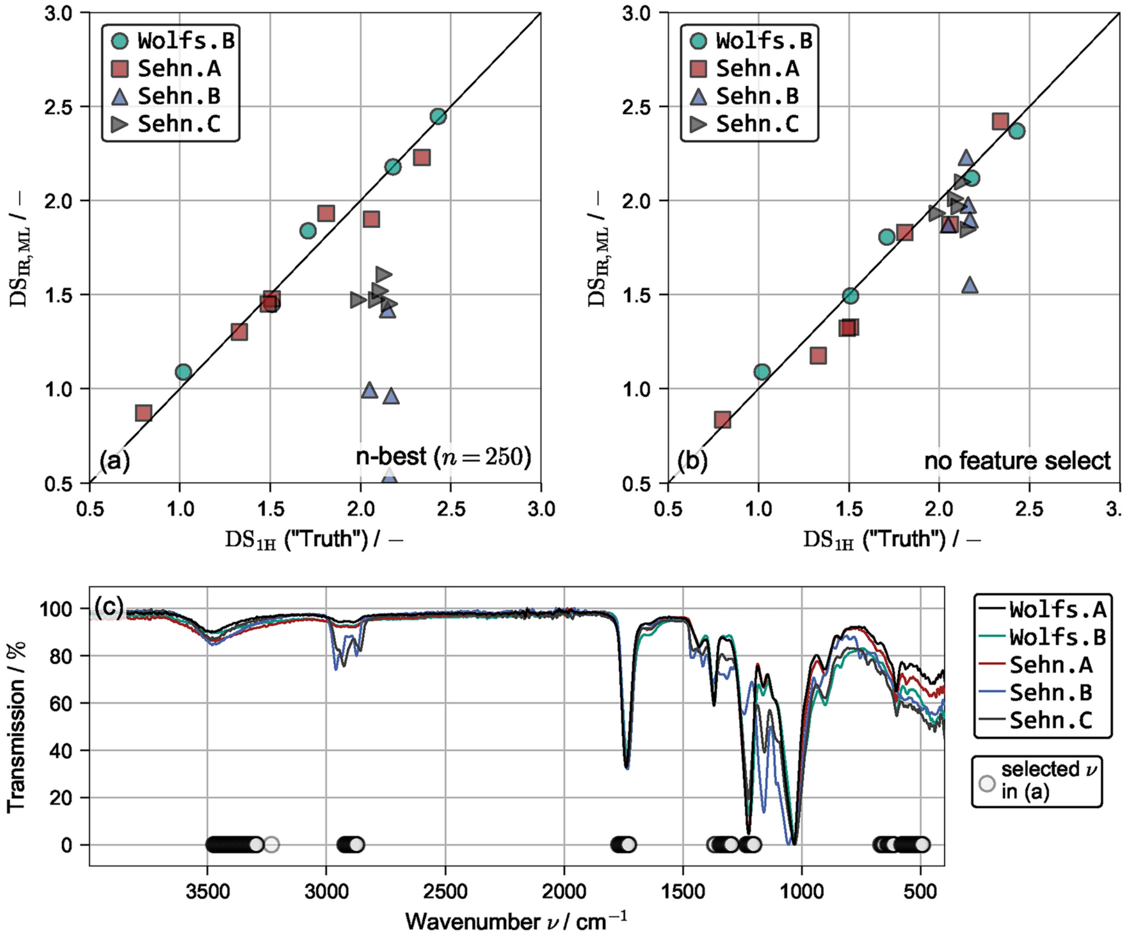

Fig. 5. DS_{1H} versus $DS_{IR,ML}$ values for the Wolfs.B, Sehn.A, Sehn.B and Sehn.C data set with applied n-best feature select ($n = 250$) (**a**) and without feature selection (**b**). All models were trained on the full Wolfs.A data set. Exemplary IR spectra at $DS_{1H} \approx 2$.

	Feature Select ($n = 250$)		No Feature Select	
Name	MAE [–]	MRE [%]	MAE [–]	MRE [%]
Wolfs.B	0.056	3.82	0.061	3.76
Sehn.A	0.080	5.00	0.118	7.58
Sehn.B	1.281	59.73	0.266	12.37
Sehn.C	0.592	28.19	0.125	5.88

Table 2. MAE and MRE values for the Wolfs.B, Sehn.A, Sehn.B and Sehn.C data set with applied n-best feature select ($n = 250$) (left) and without feature selection (right). All models were trained on the full Wolfs.A data set.

fewer hyper-parameters. For a NN to perform well, the number of layers, number of nodes in each layer and activation functions have to be chosen correctly, i.e. optimized with respect to the training data. Otherwise NN are highly prone to overfitting and are known for their poor extrapolation capabilities. A previous study showed how this can be done directly during training[47], however, the increased work and complexity did not seem to be justified in this context. Additionally, a NN is a black box without any apparent meaning to the individual weights, while the magnitude of the specific weights of the MLR herein directly corresponds to the relative importance of a specific wavenumber.

Conclusions

This work demonstrates that a simple multiple linear regression model can be trained based on a small data set of raw IR spectra of CAs and corresponding DS values, which is capable of predicting the DS of unseen spectra with high accuracy, i.e. mean absolute errors of < 0.1. This approach does neither require any manual

IR data processing like baseline corrections nor manual evaluations like integration or calibration and is therefore highly interesting for fast and unbiased routine analyses in industry and academia. Applying a feature selection algorithm did not only result in lower prediction errors, but simultaneously provided insights into the wavenumber areas, where information on the DS is stored. The fingerprint region was identified as relevant, which would commonly be neglected when evaluating manually. This methodology can generally be applied to other spectroscopic measurements, as long as information on the DS is stored within. The trained model was able to predict the DS based on raw IR data of other experimenters applying a different synthesis route, demonstrating the robustness of the approach. It should be emphasized that the data used in this study was not specifically generated to train regression models. When training a model for production, i.e. routine evaluations of DS, more and most importantly evenly spaced data should be recorded to build an even more reliable model without bias towards certain DS ranges.

The trained model was applied to different CEs and achieved overall good prediction accuracy. This is surprising considering the degree of extrapolation, as only CAs were used during training. A closer inspection of the IR spectra suggests that with enough data, not only DS prediction, but also ester identification should become feasible. Future work should focus on applying design of experiment techniques to systematically extend the data set to different CEs. With this data, the proposed model architectures should be tested, to obtain a more generalized evaluation model for cellulose esters. If necessary, the models could be extended with other forms of input data, such as images or different spectroscopic measurements, to allow for a more differentiated prediction.

Methods

^{1}H integration routine

As discussed in the introduction, DS determination via ^{1}H NMR is a frequently employed tool. However, the method may suffer from individual baseline correction and integration, i.e. the process depends on individual researchers and thus may include individual bias. Hence, the following step-by-step protocol for the DS determination of CA via ^{1}H NMR using *MestReNova* is introduced and was used in this study to avoid such errors:

1. Reference ppm values to solvent signal (DMSO-d_6 = 2.50 ppm)
2. Choose *"Multipoint Baseline Correction"*
3. Apply *"Multipoint Baseline Correction"* in the following areas:
   (a) From 0.00 ppm to 0.50 ppm
   (b) From 6.00 ppm to 10.0 ppm
4. Integration of magnetic resonances in the following areas:
   (a) From 1.40 ppm to 2.25 ppm (acetyl signal)
   (b) From 2.73 ppm to 6.00 ppm (AGU signal)

IR data normalization

Before use, the raw IR transmission spectrum of each sample *j*, denoted by $\mathbf{T}_{\mathrm{raw},j} \in \mathbb{R}^{1\times M}$, where *M* is the number of wavenumbers, is scaled according to the min-max normalization in Eq. 1. Throughout this publication, arrays are notated in bold to distinguish them from integers.

$$\mathbf{T}_j = \frac{\mathbf{T}_{\mathrm{raw},j} - \min(\mathbf{T}_{\mathrm{raw},j})}{\max(\mathbf{T}_{\mathrm{raw},j}) - \min(\mathbf{T}_{\mathrm{raw},j})} \tag{1}$$

ML model structure and implementation

The goal of this work was to train and evaluate regression models that are able to predict the DS based on a provided IR spectrum. Training was achieved using static data, so supervised learning was applied. The main success-defining factor is the chosen model architecture. In this study, one of the simplest model architectures, a multiple linear regression (MLR) model was used that is based on ordinary least squares. For any given training set, the DS data is structured in a $\mathbf{D} \in \mathbb{R}^{S\times 1}$ array, while the corresponding spectroscopic data is structured in a $\mathbf{T} \in \mathbb{R}^{S\times M}$. *S* is the number of samples and *M* is the amount of wavenumbers (features), while in this case the amount of predicted values (outputs) is one (i.e. the DS). A MLR model $\mathcal{M}$ is defined by a set of weights $\mathbf{w} \in \mathbb{R}^{M\times 1}$, so that

$$\mathbf{D}_{\mathrm{IR,ML}} = \mathcal{M}(\mathbf{T}) = \mathbf{T}\mathbf{w} \quad . \tag{2}$$

During training, the weights are optimized according to

$$\min_{w} ||\mathbf{T}\mathbf{w} - \mathbf{D}||_2^2 \tag{3}$$

and the model was implemented with the scikit-learn[48] class `LinearRegression` in Python. As DS values can only be positive, the ReLU (rectified linear unit) function $f(x) = \max(0, x)$ is applied to the MLR output to enforce this physical constraint. The beauty of a MLR model is that it has only one relevant hyper parameter: A constant term can be included in the features and the corresponding weights correspond to the wavenumber-specific intercepts. This option `fit_intercept` was set to False in this study.

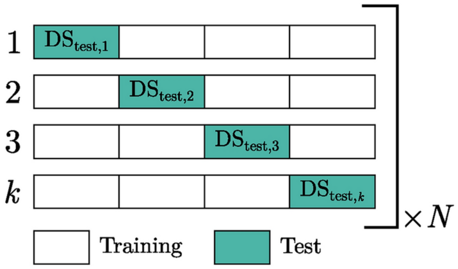

Fig. 6. Visualization of the repeated cross validation procedure.

Quality metrics

The mean absolute error (MAE) between any reference $\mathbf{D}_{\text{ref}}$ and evaluation data set $\mathbf{D}_{\text{eval}}$ of size $\mathbb{R}^{S\times1}$ is defined as

$$\text{MAE} = \frac{1}{S}\sum_{i=1}^{S}|\mathbf{D}_{\text{eval},i} - \mathbf{D}_{\text{ref},i}| \tag{4}$$

and has the same dimension as DS. Similarly, the mean relative error (MRE) is defined as

$$\text{MRE} = \frac{1}{S}\sum_{i=1}^{S}\frac{|\mathbf{D}_{\text{eval},i} - \mathbf{D}_{\text{ref},i}|}{\mathbf{D}_{\text{ref},i}} \quad . \tag{5}$$

Repeated cross validation

To quantify the predictive performance of any supervised learning model, it is important to test it on unseen data that has not been used during training. Hence, the full data set available is initially split into a train and test subset, with self-explanatory intention. However, this is problematic because the model performance, e.g. the MAE in Eq. 4, is only evaluated on the test subset. To quantify the model performance on the entire data set, a k-fold cross validation (CV) is routinely used that is visualized in Fig. 6. The full data set is split into k subsets, ensuring that every data point is exclusively contained in one subset. Subsequently, these subsets are used as test data, while the remaining points are used for model training. In essence, k individual models are trained and each is evaluated on the respective test data. Subsequently, Eq. 4 can be applied to the combined predictions of the k test subsets, yielding an objective metric for model accuracy. As the assignment into k subsets is generally done randomly, there are various possible data combinations that differ in model performance. To eliminate this stochastic behavior, the k-fold CV is repeated N times with randomized data assignment between repetitions. The `RepeatedKFold` class of the scikit-learn[48] package was used.

Feature select

The idea of a feature select (FS) algorithm is to only train the model on the most relevant features (wavenumbers). This can generally improve model performance and make a model more robust, as noisy or non-informative features are omitted. Obviously, one hast to define what *relevant* means in this context. In this work, the `SelectKBest` class based on the `f_regression` metric of scikit-learn[48] was used. Here, the Pearson correlation coefficient

$$\boldsymbol{\rho}_i = \frac{\text{cov}\,(\mathbf{T}[:,i],\mathbf{D})}{\sigma(\mathbf{T}[:,i])\sigma(\mathbf{D})} = \frac{\frac{1}{S-1}\sum_{j=1}^{S}(\mathbf{T}_{j,i} - \bar{T}_i)(\mathbf{D}_j - \bar{D})}{\sqrt{\frac{1}{S-1}\sum_{j=1}^{S}(\mathbf{T}_{j,i} - \bar{T}_i)}\sqrt{\frac{1}{S-1}\sum_{j=1}^{S}(\mathbf{D}_j - \bar{D})}} \tag{6}$$

is calculated for each individual feature i that measures the linear correlation between any wavenumber and the DS. $\boldsymbol{\rho}$ is a normalized measure of covariance and can take values between -1 and 1. Subsequently, the F-statistic

$$\mathbf{F}_i = \frac{\boldsymbol{\rho}_i^2}{1-\boldsymbol{\rho}_i^2}(n-2) \tag{7}$$

is computed, which ranks the features between 0 and 1 and therefore also ranks strong negative correlations highly. Providing an integer n, the algorithm selects the n-best features from $\mathbf{F}$, i.e. the wavenumbers with the highest F values.

Data availibility

All data used in this study is available at KITopen[46]. This includes both raw ATR-FTIR spectra as well as DS values based on IR integration, ^{1}H NMR and ^{31}P NMR. The Python scripts for reproducing the results presented in this study are publicly available in the Github repository pdhs-group/DS_IR_ML.

Received: 25 September 2024; Accepted: 10 January 2025
Published online: 23 January 2025

References

1. Wolfs, J. & Meier, M. A. R. A more sustainable synthesis approach for cellulose acetate using the DBU/CO2 switchable solvent system. *Green Chem.* **23**, 4410–4420. https://doi.org/10.1039/D1GC01508G (2021).
2. Klemm, D., Heublein, B., Fink, H.-P. & Bohn, A. Cellulose: Fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed.* **44**, 3358–3393. https://doi.org/10.1002/anie.200460587 (2005).
3. Zhao, X. et al. Thermal and barrier characterizations of cellulose esters with variable side-chain lengths and their effect on PHBV and PLA bioplastic film properties. *ACS Omega* **6**, 24700–24708. https://doi.org/10.1021/acsomega.1c03446 (2021).
4. Esen, E., Hädinger, P. & Meier, M. A. R. Sustainable fatty acid modification of cellulose in a CO2-based switchable solvent and subsequent Thiol-Ene modification. *Biomacromol* **22**, 586–593. https://doi.org/10.1021/acs.biomac.0c01444 (2020).
5. Erdal, N. B. & Hakkarainen, M. Degradation of cellulose derivatives in laboratory, man-made, and natural environments. *Biomacromol* **23**, 2713–2729. https://doi.org/10.1021/acs.biomac.2c00336 (2022).
6. Sehn, T. & Meier, M. A. R. Structure-property relationships of short chain (mixed) cellulose esters synthesized in a DMSO/TMG/CO2 switchable solvent system. *Biomacromol* **24**, 5255–5264. https://doi.org/10.1021/acs.biomac.3c00762 (2023).
7. Heinze, T. et al. Effective preparation of cellulose derivatives in a new simple cellulose solvent. *Macromol. Chem. Phys.* **201**, 627–631. https://doi.org/10.1002/(SICI)1521-3935(20000301)201:6<627::AID-MACP627>3.0.CO;2-Y (2000).
8. Wu, J. et al. Homogeneous acetylation of cellulose in a new ionic liquid. *Biomacromol* **5**, 266–268. https://doi.org/10.1021/bm034398d (2004).
9. Fischer, S. et al. Properties and applications of cellulose acetate. *Macromol. Symp.* **262**, 89–96. https://doi.org/10.1002/masy.200850210 (2008).
10. Steinmeier, H. 3. Acetate manufacturing, process and technology 3.1 Chemistry of cellulose acetylation. *Macromol. Symp.* **208**, 49–60. https://doi.org/10.1002/masy.200450405 (2004).
11. Huang, L., Wu, Q., Wang, Q. & Wolcott, M. One-step activation and surface fatty acylation of cellulose fibers in a solvent-free condition. *ACS Sustain. Chem. Eng.* **7**, 15920–15927. https://doi.org/10.1021/acssuschemeng.9b01974 (2019).
12. Li, M.-L. et al. One-step solvent-free strategy to efficiently synthesize high-substitution cellulose esters. *ACS Sustain. Chem. Eng.* **12**, 9669–9681. https://doi.org/10.1021/acssuschemeng.4c00953 (2024).
13. Duchatel-Crépy, L. et al. Substitution degree and fatty chain length influence on structure and properties of fatty acid cellulose esters. *Carbohyd. Polym.* **234**, 115912. https://doi.org/10.1016/j.carbpol.2020.115912 (2020).
14. Tanaka, S., Iwata, T. & Iji, M. Long/short chain mixed cellulose esters: Effects of long acyl chain structures on mechanical and thermal properties. *ACS Sustain. Chem. Eng.* **5**, 1485–1493. https://doi.org/10.1021/acssuschemeng.6b02066 (2017).
15. Chen, Z., Zhang, J., Xiao, P., Tian, W. & Zhang, J. Novel thermoplastic cellulose esters containing bulky moieties and soft segments. *ACS Sustain. Chem. Eng.* **6**, 4931–4939. https://doi.org/10.1021/acssuschemeng.7b04466 (2018).
16. Jogunola, O. et al. Ionic liquid mediated technology for synthesis of cellulose acetates using different co-solvents. *Carbohyd. Polym.* **135**, 341–348. https://doi.org/10.1016/j.carbpol.2015.08.092 (2016).
17. Gao, X. et al. Rapid transesterification of cellulose in a novel DBU-derived ionic liquid: Efficient synthesis of highly substituted cellulose acetate. *Int. J. Biol. Macromol.* **242**, 125133. https://doi.org/10.1016/j.ijbiomac.2023.125133 (2023).
18. Yao, Z. et al. Rapid homogeneous acylation of cellulose in a CO2 switchable solvent by microwave heating. *ACS Sustain. Chem. Eng.* **10**, 17327–17335. https://doi.org/10.1021/acssuschemeng.2c05872 (2022).
19. Tezuka, Y. & Tsuchiya, Y. Determination of substituent distribution in cellulose acetate by means of a 13C NMR study on its propanoated derivative. *Carbohyd. Res.* **273**, 83–91. https://doi.org/10.1016/0008-6215(95)00107-5 (1995).
20. Kono, H., Hashimoto, H. & Shimizu, Y. NMR characterization of cellulose acetate: Chemical shift assignments, substituent effects, and chemical shift additivity. *Carbohyd. Polym.* **118**, 91–100. https://doi.org/10.1016/j.carbpol.2014.11.004 (2015).
21. Fliri, L. et al. Solution-state nuclear magnetic resonance spectroscopy of crystalline cellulosic materials using a direct dissolution ionic liquid electrolyte. *Nat. Protoc.* **18**, 2084–2123. https://doi.org/10.1038/s41596-023-00832-9 (2023).
22. King, A. W. et al. A new method for rapid degree of substitution and purity determination of chloroform-soluble cellulose esters, using 31P NMR. *Anal. Methods* **2**, 1499–1505. https://doi.org/10.1039/C0AY00336K (2010).
23. Freire, C., Silvestre, A., Pascoal Neto, C. & Rocha, R. An efficient method for determination of the degree of substitution of cellulose esters of long chain aliphatic acids. *Cellulose* **12**, 449–458. https://doi.org/10.1007/s10570-005-2203-2 (2005).
24. Peydecastaing, J., Vaca-Garcia, C. & Borredon, E. Accurate determination of the degree of substitution of long chain cellulose esters. *Cellulose* **16**, 289–297. https://doi.org/10.1007/s10570-008-9267-8 (2009).
25. Wolfs, J., Nickisch, R., Wanner, L. & Meier, M. A. R. Sustainable one-pot cellulose dissolution and derivatization via a tandem reaction in the DMSO/DBU/CO2 switchable solvent system. *J. Am. Chem. Soc.* **143**, 18693–18702. https://doi.org/10.1021/jacs.1c08783 (2021).
26. Casarano, R., Fidale, L. C., Lucheti, C. M., Heinze, T. & El Seoud, O. A. Expedient, accurate methods for the determination of the degree of substitution of cellulose carboxylic esters: Application of UV-vis spectroscopy (dye solvatochromism) and FTIR. *Carbohyd. Polym.* **83**, 1285–1292. https://doi.org/10.1016/j.carbpol.2010.09.035 (2011).
27. Wolfs, J., Scheelje, F. C. M., Matveyeva, O. & Meier, M. A. R. Determination of the degree of substitution of cellulose esters via ATR-FTIR spectroscopy. *J. Polym. Sci.* **61**, 2697–2707. https://doi.org/10.1002/pol.20230220 (2023).
28. Beer, A. Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Ann. Phys.* **162**, 78–88. https://doi.org/10.1002/andp.18521620505 (1852).
29. Kowalski, B. R. Chemometrics: Views and propositions. *J. Chem. Inf. Comput. Sci.* **15**, 201–203. https://doi.org/10.1021/ci60004a002 (1975).
30. Meza Ramirez, C. A., Greenop, M., Ashton, L. & Rehman, I. u. Applications of machine learning in spectroscopy. *Appl. Spectrosc. Rev.* **56**, 733–763. https://doi.org/10.1080/05704928.2020.1859525 (2021).
31. Ralbovsky, N. M. & Lednev, I. K. Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chem. Soc. Rev.* **49**, 7428–7453. https://doi.org/10.1039/D0CS01019G (2020).
32. Guo, S., Popp, J. & Bocklitz, T. Chemometric analysis in Raman spectroscopy from experimental design to machine learning-based modeling. *Nat. Protoc.* **16**, 5426–5459. https://doi.org/10.1038/s41596-021-00620-3 (2021).
33. Winkler, M., Gleiss, M. & Nirschl, H. Soft sensor development for real-time process monitoring of multidimensional fractionation in tubular centrifuges. *Nanomaterials.*[SPACE]https://doi.org/10.3390/nano11051114 (2021).

34. Maris, M. A., Brown, C. W. & Lavery, D. S. Nonlinear multicomponent analysis by infrared spectrophotometry. *Anal. Chem.* **55**, 1694–1703. https://doi.org/10.1021/ac00261a013 (1983).
35. Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* **58**, 512–519. https://doi.org/10.1002/mrc.4989 (2020).
36. Kern, S. et al. Artificial neural networks for quantitative online NMR spectroscopy. *Anal. Bioanal. Chem.* **412**, 4447–4459. https://doi.org/10.1007/s00216-020-02687-5 (2020).
37. Lansford, J. L. & Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat. Commun.* **11**, 1513. https://doi.org/10.1038/s41467-020-15340-7 (2020).
38. Nogueira, M. S. et al. Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. *Sci. Rep.* **11**, 15409. https://doi.org/10.1038/s41598-021-93511-2 (2021).
39. Simon, J. et al. A fast method to measure the degree of oxidation of dialdehyde celluloses using multivariate calibration and infrared spectroscopy. *Carbohyd. Polym.* **278**, 118887. https://doi.org/10.1016/j.carbpol.2021.118887 (2022).
40. Liu, J., Chen, J., Dong, N., Ming, J. & Zhao, G. Determination of degree of substitution of carboxymethyl starch by Fourier transform mid-infrared spectroscopy coupled with partial least squares. *Food Chem.* **132**, 2224–2230. https://doi.org/10.1016/j.foodchem.2011.12.072 (2012).
41. Barros, A. S. et al. Determination of the degree of methylesterification of pectic polysaccharides by FT-IR using an outer product PLS1 regression. *Carbohyd. Polym.* **50**, 85–94. https://doi.org/10.1016/S0144-8617(02)00017-6 (2002).
42. Zhang, K., Feldner, A. & Fischer, S. Ft Raman spectroscopic investigation of cellulose acetate. *Cellulose* **18**, 995–1003. https://doi.org/10.1007/s10570-011-9545-8 (2011).
43. Pancholi, M. J., Khristi, A., Athira, K. M. & Bagchi, D., Comparative analysis of lignocellulose agricultural waste and pre-treatment conditions with FTIR and machine learning modeling. *BioEnergy Res.* **16**, 123–137. https://doi.org/10.1007/s12155-022-10444-y (2023).
44. Costa, L. R., Tonoli, G. H. D., Milagres, F. R. & Hein, P. R. G. Artificial neural network and partial least square regressions for rapid estimation of cellulose pulp dryness based on near infrared spectroscopic data. *Carbohyd. Polym.* **224**, 115186. https://doi.org/10.1016/j.carbpol.2019.115186 (2019).
45. Li, X., Sun, C., Zhou, B. & He, Y. Determination of hemicellulose, cellulose and lignin in moso bamboo by near infrared spectroscopy. *Sci. Rep.* **5**, 17210. https://doi.org/10.1038/srep17210 (2015).
46. Rhein, F., Sehn, T. & Meier, M. A. R. Degree of substitution of cellulose acetate and other esters: Raw ATR-FTIR spectra and DS data obtained from IR integration, 1H NMR and 31P NMR [dataset]. *RADAR4KIT*. https://doi.org/10.35097/tvwlylbMvDXhEcRt (2024).
47. Rhein, F., Hibbe, L. & Nirschl, H. Hybrid modeling of hetero-agglomeration processes: A framework for model selection and arrangement. *Eng. Comput.*[SPACE]https://doi.org/10.1007/s00366-023-01809-8 (2023).
48. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

The authors acknowledge Jonas Wolfs for data provision.

Author contributions

Conceptualization, F.R., M.M.; methodology, F.R., T.S., M.M.; formal analysis, F.R., T.S.; investigation, T.S.; data curation, F.R., T.S.; writing—original draft preparation, F.R., T.S.; writing—review and editing, M.M.; visualization, F.R.; software, F.R.; supervision, F.R., M.M.; project administration, F.R., M.M.. All authors have read and agreed to the published version of the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-86378-0.

Correspondence and requests for materials should be addressed to F.R. or M.A.R.M.