Received: 21 April 2023 | Revised: 30 June 2023 | Accepted: 3 July 2023

DOI: 10.1002/mrc.5381

RESEARCH ARTICLE

WILEY

Automated nuclear magnetic resonance fingerprinting of mixtures

Thomas Specht | **Justus Arweiler** | **Johannes Stüber** | **Kerstin Münnemann** | **Hans Hasse** | **Fabian Jirasek**

Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern, Kaiserslautern, Germany

Correspondence
Fabian Jirasek, Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663 Kaiserslautern, Germany.
Email: fabian.jirasek@rptu.de

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant/Award Number: JI 401/1-1; Carl-Zeiss-Stiftung

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a powerful tool for qualitative and quantitative analysis. However, for complex mixtures, determining the speciation from NMR spectra can be tedious and sometimes even unfeasible. On the other hand, identifying and quantifying structural groups in a mixture from NMR spectra is much easier than doing the same for components. We call this group-based approach "NMR fingerprinting." In this work, we show that NMR fingerprinting can even be performed in an automated way, without expert knowledge, based only on standard NMR spectra, namely, ^{13}C, ^{1}H, and ^{13}C DEPT NMR spectra. Our approach is based on the machine-learning method of support vector classification (SVC), which was trained here on thousands of labeled pure-component NMR spectra from open-source data banks. We demonstrate the applicability of the automated NMR fingerprinting using test mixtures, of which spectra were taken using a simple benchtop NMR spectrometer. The results from the NMR fingerprinting agree remarkably well with the ground truth, which was known from the gravimetric preparation of the samples. To facilitate the application of the method, we provide an interactive website (https://nmr-fingerprinting.de), where spectral information can be uploaded and which returns the NMR fingerprint. The NMR fingerprinting can be used in many ways, for example, for process monitoring or thermodynamic modeling using group-contribution methods—or simply as a first step in species analysis.

KEYWORDS

^{13}C, ^{1}H, benchtop NMR, machine learning, mixtures, NMR, SMARTS, support vector classification

1 | INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is an established technique for elucidating the structure of unknown components. However, this usually requires expert knowledge and may be tedious, especially if mixtures are studied. Therefore, computer-assisted structure elucidation (CASE) programs have been developed to facilitate this process.[1] As input, these programs typically require a set of 1D and 2D NMR spectra of the

unknown sample and information on the molecular formula of the components to be identified, usually obtained by high-resolution mass spectrometry.[2–5] CASE programs then propose the most probable molecular structures to the user by applying a set of rules and logic and comparing the recorded spectra to the predicted spectra of the candidate molecules. Unfortunately, CASE programs are currently restricted to pure components, which strongly limits their applicability in practice.

NMR spectroscopy has also successfully been applied for the component-specific analysis of mixtures, yielding both qualitative and quantitative information on the composition.[6–11] However, in complex mixtures, elucidating all components and determining their concentrations is only tedious in the best case; in the worst case, it is infeasible. Suppose all components in a mixture are known, and only their concentrations are unknown. In that case, the situation is comparatively simple, and various methods for the automated quantification exist, even for situations where peaks overlap.[12–17] If, on the other hand, the mixture contains unknown components, already the first step in the evaluation, namely, the decision which peaks to consider, is generally ambiguous and subject to the user, although first machine-learning (ML) approaches for automation have been developed.[18]

A data-driven approach to facilitate the evaluation of NMR spectra applicable to mixtures and employed, for example, in metabolomics,[19] is dereplication.[20–22] In dereplication, individual components in a mixture are identified by comparison of the experimental NMR spectrum of the mixture with the NMR spectra of pure components retrieved from a data bank. While this approach is, in principle, straightforward, its main problem is that even the largest NMR data banks contain only a tiny fraction of all possible components,[4] so dereplication approaches are limited to specific applications. Moreover, methods that rely on comparing measured NMR spectra to those of data banks are often sensitive to differences in conditions of the experiments.[4] Hence, there is currently no broadly applicable, reliable, and robust way for automatically elucidating the components in mixtures from NMR spectra.

Compared with the elucidation of the *components* of a mixture by NMR spectroscopy, identifying the *structural groups* that make up the components is a much simpler task. We call the respective procedure of obtaining the group-specific information from NMR spectra "NMR fingerprinting" in the following. NMR fingerprints can be of great practical relevance, for example, for monitoring reaction and separation processes or as a basis for predicting the properties of mixtures by thermodynamic group-contribution methods.[23–28] Another potential application field is the so-called biofluid analysis using benchtop NMR spectrometers,[29–31] for example, for detecting diseases in combination with ML methods.

NMR fingerprinting is facilitated by the fact that NMR spectroscopy directly provides information about the local environment of the studied nucleus, that is, the structural group to which it belongs. Therefore, chemical shift tables exist that depict characteristic ranges of structural groups in different NMR spectra.[32] However, because the characteristic ranges of structural groups often overlap, the peak assignment based on chemical shift tables is not unambiguous. To overcome the problems associated with working with chemical shift tables, in prior work,[33] we have proposed to use a supervised ML method, namely, support vector classification (SVC), for identifying and assigning structural groups to chemical shift regions in ^{13}C NMR spectra, which also uses information from ^{1}H NMR spectra. During the training, the SVC learned to map the information from the NMR spectra (the input) to different structural groups (the output). Although the SVC was trained only on pure-component data, the method yields good results also for mixtures.[33]

In this work, the NMR fingerprinting concept is substantially extended to also include information from ^{13}C DEPT NMR spectra in addition to ^{1}H and ^{13}C NMR spectra of the sample. The ^{13}C DEPT NMR spectra provide direct information on the substitution degree of the carbon atoms and are used here to differentiate, for example, between "CH_3" and "CH_2" groups in the respective sample.

Furthermore, in this work, we introduce using SMARTS in the NMR fingerprinting framework. SMARTS is an acronym for SMILES arbitrary target specification strings,[34] which are based on the simplified molecular-input line-entry system (SMILES),[35] which, in turn, is a system to represent components by simple text strings. SMARTS are used here for a rigorous definition of the distinguished structural groups in the NMR fingerprinting framework and enable a fully automated training workflow. SMARTS also provide great flexibility in defining the groups so that the approach can be straightforwardly tailored to a specific application.

Additionally, the NMR fingerprinting method was extended in the present work to optionally consider prior knowledge about the presence or absence of labile protons in the sample, that is, protons that show chemical exchange with other protons in the sample, if such information is available. Information about the presence of labile protons can, for example, be identified by their broad peak form in ^{1}H NMR spectroscopy or from heteronuclear single quantum coherence (HSQC) experiments.

The new method was trained on spectra of 2839 pure components from two data banks, namely, the

NMRShiftDB[36] and the Biological Magnetic Resonance Data Bank (BMRB),[37] and was tested using rigorous nested cross-validation (CV). Furthermore, a data augmentation technique was developed to substantially increase the training data set and address the fact that no comprehensive data bank for NMR spectra of mixtures is available. Finally, several test mixtures were analyzed with an 80 MHz benchtop NMR spectrometer, and the recorded spectra were used as input for testing the new approach in a practical setting.

Together with this paper, an interactive website (https://nmr-fingerprinting.de) was published, which enables testing and applying the method presented here through a graphical user interface (GUI) without the need for any program installation. The user supplies the spectral information and gets the corresponding NMR fingerprint.

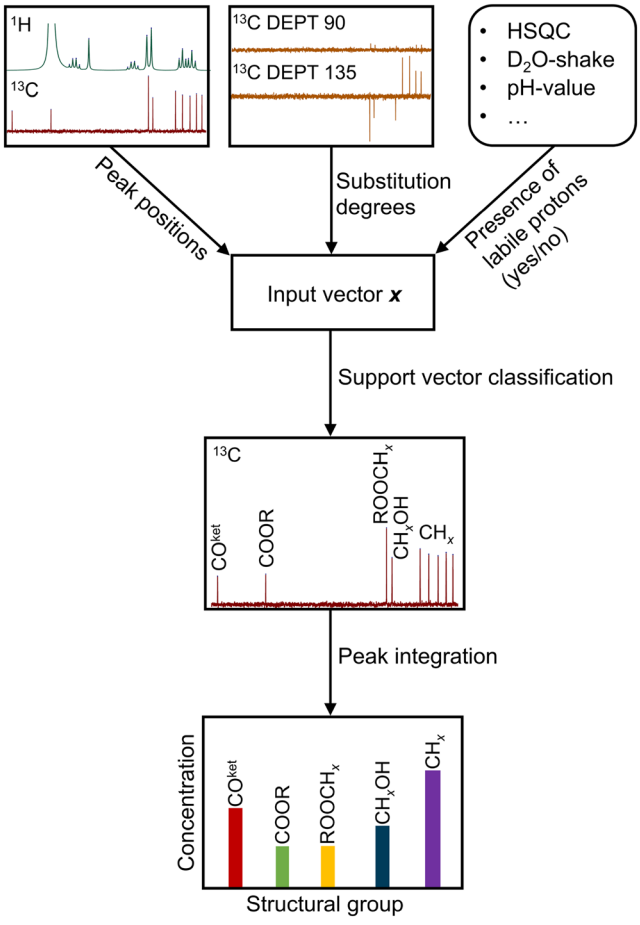

FIGURE 1 Workflow of the developed NMR fingerprinting method using an SVC for predicting structural groups based on spectral information from ^{1}H, ^{13}C, and ^{13}C DEPT NMR spectra as well as using binary information about the presence or absence of labile protons in the sample. In the Supporting Information, a variant of the method that does not require information about labile protons is presented.

2 | OVERVIEW OF THE WORKFLOW

Figure 1 visualizes the workflow of the method developed in this work, which can be used via the website (https://nmr-fingerprinting.de). The method's goal is identifying structural groups in an unknown sample and assigning them to peaks in the ^{13}C NMR spectrum of the sample. In the present version, the NMR fingerprinting method can differentiate thirteen structural groups, which are summarized in Table 1. The groups are the same as in our recent work,[33] but SMARTS strings for each group are also provided here.

For using the method, a ^{1}H NMR spectrum, a ^{13}C NMR spectrum, and ^{13}C DEPT 90/135 NMR spectra of the studied sample are required. Specifically, the chemical shift of all peaks in the ^{1}H NMR (except the peaks of labile protons) and in the ^{13}C NMR spectrum as well as the substitution degree of each carbon atom, which can be automatically determined considering the signs of the peaks in the DEPT 90/135 NMR spectra,[38] are needed for defining the input of the method. The preprocessing of the NMR spectra and peak picking is currently up to the user. Automatic workflows as implemented in commercial software like MNova can be (and were) used for this purpose in the present work. However, these tools normally still require human supervision, although there are attempts to fully automatize the peak picking task by ML methods,[18] which will be interesting to study in future work.

Furthermore, the method described here uses binary information about the presence or absence of labile protons in the sample. In the Supporting Information, an additional variant of the NMR fingerprinting method that does not require this additional information on labile protons is presented, which should be used in cases where such information is unavailable.

The spectral information, as well as the information on the presence or absence of labile protons, are used for defining the input vector $\boldsymbol{x}_i$ of a sample i for the SVC, whereby the NMR spectra are in general equidistantly binned and the peaks are assigned to the respective sections in the spectra (cf. Section 3.1 for details). The information about the presence or absence of labile protons in the sample can come from different sources. In many cases, one will know if, for example, carboxylic acids or alcohols are present in the sample. However, even if this information is not available a priori, there are multiple ways to its determination. For instance, labile protons can often be recognized from the ^{1}H NMR spectrum because they usually show characteristic, very broad peaks. Alternatively, HSQC NMR spectra can be used, where peaks of labile protons show no correlation with any carbon nucleus in the sample. Another established method is the

so-called D_2O-shake, that is, the addition of a small amount of deuterated water to the sample, which will cause the peaks stemming from labile protons to vanish or at least significantly change their position and/or amplitude if compared to a ^{1}H NMR spectrum before adding D_2O. Also, simple pH measurements can detect labile protons, specifically ones from carboxylic acids.

The SVC method was trained on labeled data for 2839 pure components to predict the structural groups summarized in Table 1 (cf. Section 3.3 and Supporting Information for details). Finally, by integrating the peaks in the ^{13}C NMR spectrum, the concentrations of all identified structural groups can be directly obtained.

3 | DATA AND METHODS

3.1 | Generation of input and output data for training

For training the method, only experimental data for pure components were used and retrieved from two NMR data banks (cf. Section 3.2). Furthermore, synthetic mixture data were generated to augment the training set by combining the processed experimental pure-component data in the nested CV (cf. below for details). In general, for training the method and evaluating it on the pure-component data, all peaks of labile protons were removed from the ^{1}H NMR spectra.

For generating the input data for the training, the substitution degree of each carbon atom in each pure component, that is, primary (P), secondary (S), tertiary (T), or quaternary (Q), was determined automatically based on the structure of the component using RDKit[39] (cf. Supporting Information for details). Furthermore, the ^{13}C and ^{1}H NMR spectra of all pure components from the data set were divided into equidistant discrete sections. For the ^{13}C NMR spectra, the chemical shift range from 0 to 210 ppm was considered and divided into $S^{^{13}C} = 21$ sections of 10 ppm width. For the ^{1}H NMR spectra, the chemical shift range from 0 to 10 ppm was considered and divided into $S^{^{1}H} = 20$ sections of 0.5 ppm width.

Following this binning procedure and taking into account the information on the substitution degree of the carbon atoms, the ^{13}C NMR spectrum of component i was translated into four bit vectors $\boldsymbol{x}_i^{^{13}C}$ (one for each substitution degree), each of the length $S^{^{13}C}$ in the following way: starting from 0 ppm, the bit vector's entry for a specific section s was set to 1 if at least one peak associated to a carbon atom with the respective substitution degree was observed in the respective section. The four vectors for the different substitution degrees were subsequently concatenated to a single vector of length $4 \cdot S^{^{13}C}$. Furthermore, the input vector $\boldsymbol{x}_i^{^{1}H}$ resulting from the binned ^{1}H NMR spectrum was generated analogously and appended to the carbon bit vector resulting in a single vector of length 104 ($4 \cdot 21 + 20$). If peaks were observed outside the above-defined entire ranges of the NMR spectra, they were assigned to the respective nearest (edge) sections. Finally, a single bit indicating whether labile protons are present ($\hat{=}1$) or absent ($\hat{=}0$) in the component was appended, resulting in the input vector $\boldsymbol{x}_i$ of length 105 for each component i. More details on the input data generation can be found in the Supporting Information.

The general goal of the developed method is to identify and assign the structural groups g from Table 1 to sections s in the ^{13}C NMR spectrum of a component i. Hence, as the output of the method, the matrix $\boldsymbol{Y}_i$ was

TABLE 1 Structural groups distinguished in the present work and the respective SMARTS strings for their representation.

Label	Group name	SMARTS representation
CH_3	Methyl	[CX4;D1;!$(C[!#6])]
CH_x	Alkyl; $x \in \{0,1,2\}$	[CX4;D2,D3,D4;!$(C[!#6]);!R]
$cyCH_x$	Cyclic alkyl; $x \in \{0,1,2\}$	[CX4;!$(C[!#6]);R]
CH_xOH	Alcohol; $x \in \{0,1,2,3\}$	[CX4;!$(C[OX2H0][CX3H1,CX3](=O))][OX2H]
CH_xO	Ether; $x \in \{0,1,2,3\}$	[CX4;$(C[OD2]);!$(C[OX2H0][CX3H1,CX3](=O));!$(C[OX2H])]
$CH_x=$	Aliphatic double bond; $x \in \{0,1,2\}$	[CX3;!$(C~[!#6])]
$CH_x^{ar}=$	Aromatic carbon; $x \in \{0,1\}$	[cX3;!$(c~[!#6])]
$RO\text{-}CH_x^{ar}=$	Aromatic carbon with oxygen substituent; $x \in \{0,1\}$	[cX3;!$(c=O);$(c~[#8X2])]
COOR	Ester/lactone/anhydride carbonyl	[CX3H1,#6X3](=O)[#8X2H0]
$ROOCH_x$	Alkyl next to ester/lactone oxygen; $x \in \{0,1,2,3\}$	[CX4;$(C[OX2H0;$(O(C(=O)))])]
COOH	Carboxylic acid	[CX3](=O)[OX2H1]
CO^{ald}	Aldehyde	[CX3H1;!$(C[!#6])](=O)
CO^{ket}	Ketone	[#6X3H0;!$([#6][!#6])](=O)

Note: Each group contains exactly one carbon atom, and x is the number of protons directly bonded to the carbon atom.

defined for each component i from the data set, which is a bit matrix of dimension $S^{^{13}C}$ x G, where $\boldsymbol{Y}_i(s,g) = 1$ indicates the presence of at least one peak induced by structural group g in section s and G is the total number of distinguished groups ($G = 13$; cf. Table 1). The output matrix for each component was generated automatically using the respective SMARTS strings (cf. Table 1) and the RDKit package[39] (cf. Supporting Information for details).

3.2 | Collection of pure-component NMR data

Raw NMR spectra of 2839 pure components were adopted from the BMRB[37] data bank and the NMRShiftDB[36] data bank and used for training and evaluation of the developed method, whereby spectra from the NMRShiftDB were preferred if data for a given component were available in both data banks. However, not all components and respective spectra reported in these NMR data banks were used. Therefore, besides removing erroneous spectra (cf. Supporting Information for details), the following criteria had to be met:

1. Both a ^{13}C *and* a ^{1}H spectrum of the component are available.
2. The component is composed only of carbon (C), hydrogen (H), and/or oxygen (O) and no other elements.
3. The component can unambiguously be segmented into structural groups from the list in Table 1.

The first restriction ensures that only components are considered for which the complete spectral information that is needed as input for the SVC method is available. This could be relaxed in future work by the augmentation of incomplete data sets by predicted NMR spectra, for example, using hierarchically ordered-spherical description of environment (HOSE) methods,[40] density-functional theory (DFT) calculations,[41] or ML approaches.[42] The second restriction is needed because the proposed structural groups are only made up of the elements C, H, and O in the current work (cf. Table 1). Considering additional elements will be interesting in future work but will require additional analytical data, for example, from further NMR experiments or other analytical techniques. The third restriction is needed to ensure that all components can be completely divided into our list of structural groups. However, the extension to further groups directly depends on the availability of NMR spectra of components containing the group of interest to ensure meaningful training of the method.

For all ^{13}C NMR spectra and ^{1}H NMR spectra from the data set, the chemical shifts of all peaks were extracted with only one exception: the peaks in the ^{1}H NMR spectra originating from protons directly bonded to oxygen. The reason for this is as follows: such protons are often labile due to exchange with other protons in the sample and, depending on the conditions during acquisition (like temperature and composition of the sample), their position in the NMR spectrum can vary strongly,[38] which makes their position less informative for our method. More details on the processing of the data are given in the Supporting Information.

Figure 2a gives an overview of the data set containing the input and output data of the considered 2839 pure components, which we call D^{pure} in the following. Figure 2a thereby indicates the frequency of the

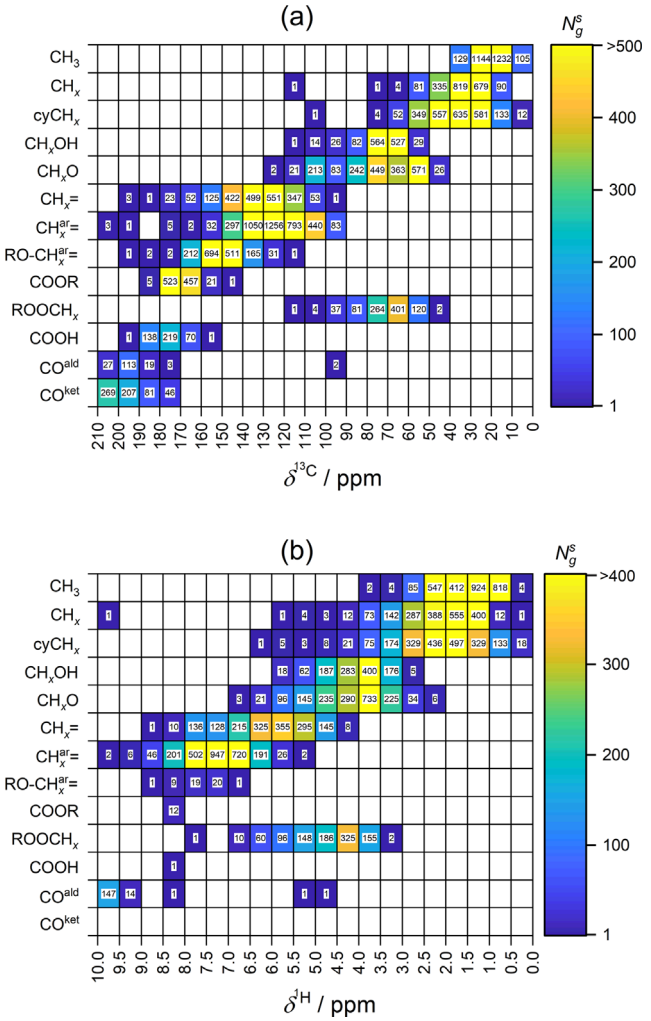

FIGURE 2 Positions of the peaks of the 2839 pure components from our data set (a) in the ^{13}C NMR spectrum and (b) in the ^{1}H NMR spectrum. The color code and the numbers inside the cells denote N_g^s, which is the number of components in the data set that contain the structural group g (row) that induces a peak in the section s of the spectrum (column). White cells refer to $N_g^s = 0$.

13 distinguished structural groups in the considered components and also in which sections of the ^{13}C NMR spectrum the respective peaks appear. The assignment of groups to sections in the spectra is not unique: peaks of at least two different structural groups are found in each section. Furthermore, there is a substantial imbalance regarding the frequency of structural groups in the data set and the frequency of peaks in the different sections of the NMR spectrum. Figure S.1 in the Supporting Information shows that by taking into account the information on the different substitution degrees, multiple assignments of groups to a single section are substantially reduced.

Figure 2b shows the respective information on the data set for the ^{1}H NMR spectrum. Similar to the ^{13}C NMR spectrum, the assignment of structural groups to the sections is not unique, so at least two different structural groups are found in each section. Furthermore, although there is a clear tendency for some groups to show peaks most frequently in some areas of the spectrum, all groups cover several chemical shift sections.

3.3 | Training of support vector classification

The core of the developed method is an SVC with a radial basis function (RBF) kernel implemented in scikit-learn 1.2.0,[43] which was trained on the pure-component data set D^{pure}.

During the development of the SVC method, D^{pure} was repeatedly divided into three subsets by a double-loop approach in the frame of a nested CV strategy[44]: a training set for fitting the model parameters, a validation set for optimizing the hyperparameters, and a test set for evaluating the predictive performance of the method. Details on how the data splits were performed are given in the Supporting Information. Nested CV was applied for each section of the ^{13}C NMR spectrum separately, whereby the data set for each section contained only those data points that induce a peak in the respective section of the ^{13}C NMR spectrum.

In the outer loop of the nested CV, 10% of the pure-component data were defined as test data, which was repeated 10 times so that each pure-component data point was part of the test set exactly once. In the inner loop, the remaining 90% of the data were divided into 80% training data and 20% validation data, which was repeated five times for each run of the outer loop so that each pure-component data point in the inner loop was in the validation set exactly once. Additionally, synthetic mixture data were generated and used in the inner loop. However, the synthetic data were *not* used in the outer loop (as test data) to ensure a fair evaluation of the method (cf. Supporting Information for details). The synthetic mixture data

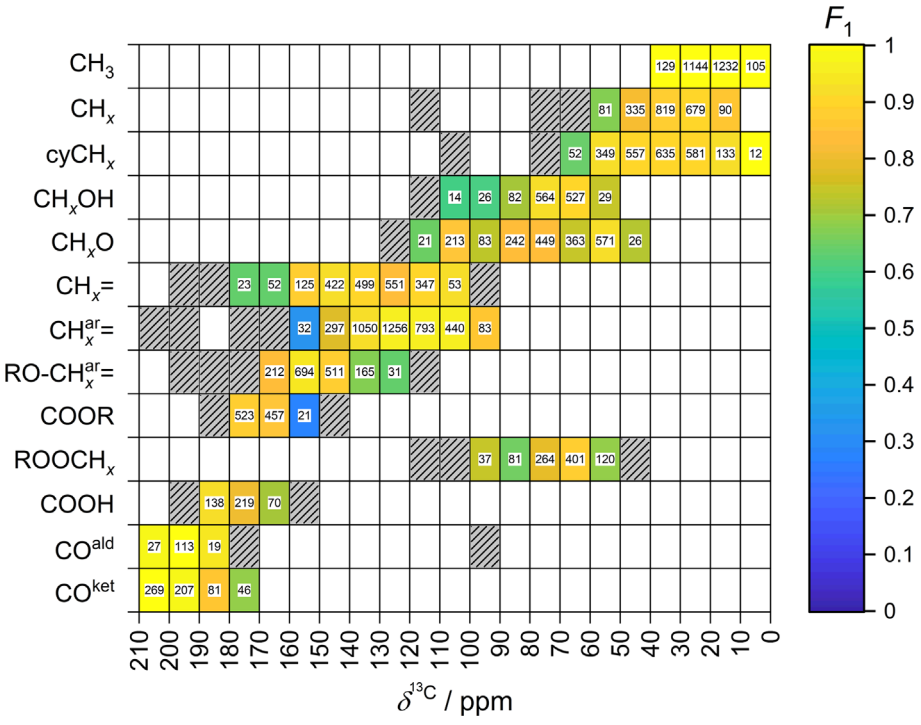

FIGURE 3 F_1 test scores (indicated by the color code) of the method for structural groups g and sections s of the ^{13}C NMR spectra of the pure components in our data set. The numbers inside the cells indicate the number of components N_g^s in the data set that contain the respective structural group g (row) inducing a peak in the respective section s of the ^{13}C NMR spectrum (column). White cells indicate group/section combination with $N_g^s = 0$, shaded cells with $N_g^s < 10$.

improves the variability of the data and, therefore, in particular, helps the method to identify the presence of combinations of different structural groups that rarely occur in a single pure component in the training set.

Optimizing the hyperparameters in the inner loop was carried out using a Bayesian optimization algorithm[45] to reduce computation time. Furthermore, only the scores for structural group/section combinations with at least 10 positive examples in the data set were considered for the optimization in the inner loop, which was done to reduce the influence of atypical data points (outliers) on the developed method (cf. Supporting Information for details).

Because an SVC is a priori only applicable to distinguish between two classes, that is, an SVC is a priori a binary classifier, but multiple classes need to be assigned to each data point here, the so-called one-vs-rest strategy[46] was employed in this work. For this purpose, multiple binary SVCs (called "units" in the following) were trained, one for predicting the presence or absence of each of the considered structural groups. The raw output of each binary SVC is its so-called decision function value, where the sign of the value indicates if a structural group is identified by the algorithm (positive value) or not (negative value), and the absolute value is proportional to the confidence of the method in the prediction.[47] Therefore, the decision that structural group g was assigned to a specific section s of the ^{13}C NMR spectrum was made by considering the values of the decision function d_g^s of all binary SVC units, whereby all groups with $d_g^s > 0$ were identified (cf. Supporting Information for details). For applying the final SVC method to NMR spectra of mixtures, this procedure was slightly adapted as described in detail in the Supporting Information.

TABLE 2 Overview of the test mixtures studied in this work.

Mixture	Components i	x_i (mol mol^{-1})
I	2-Butanone	0.0136
	Ethyl acetate	0.0145
	Water	0.9719
II	Cyclohexanone	0.0193
	Malic acid	0.0198
	1-Propanol	0.0197
	Water	0.9412
III	1-Octanol	0.9023
	tert-Butylhydroquinone	0.0977
IV	Acetone	0.1587
	Butanal	0.1313
	Oleic acid	0.7100

The predictive performance of the SVC was evaluated here using the so-called F_1 score $F_{1,g}$ for each structural group g:

$$F_{1,g} = \frac{2 \cdot TP_g}{2 \cdot TP_g + FP_g + FN_g}, \quad (1)$$

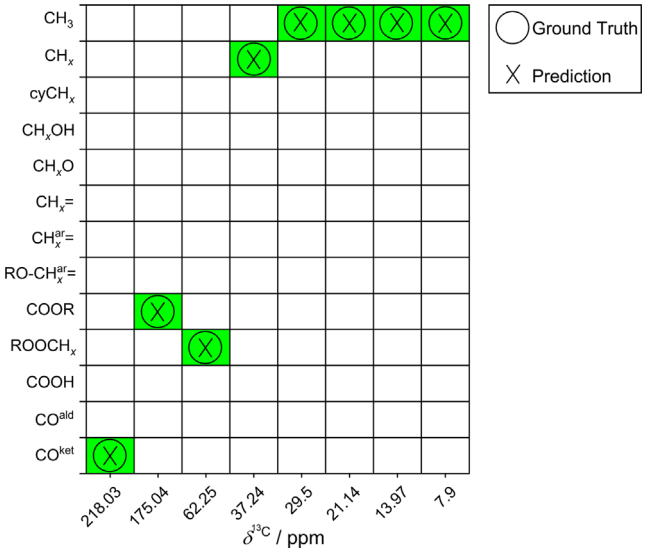

FIGURE 4 Results of the application of NMR fingerprinting to mixture I (cf. Table 2) for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

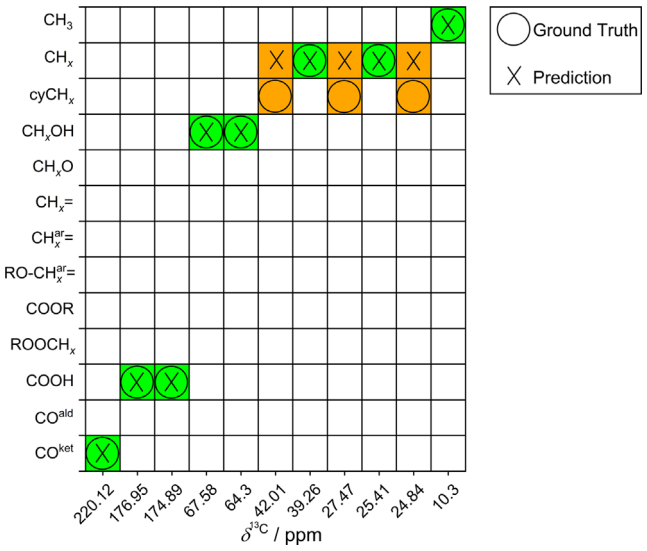

FIGURE 5 Results of the application of NMR fingerprinting to mixture II (cf. Table 2) for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions, and orange color indicates mistakes. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

where TP_g are the so-called true positives of group g, that is, the data points (pure components from our data set) that were correctly classified as containing group g; FP_g are the so-called false positives of group g, that is, the components that were incorrectly predicted to contain group g; and FN_g are the so-called false negatives of group g, that is, the components that contain group g but were falsely predicted not to contain group g. F_1 scores of 1 correspond to perfect predictions.

Additionally, a "final" method was trained following the same procedure but a CV with a split of the data set into 90% training and 10% validation data in each run. The final method was trained with the same procedure as the method that was used for testing the generalization performance by calculating the F_1 test scores. The only difference is that in the final method, no outer loop is needed where the test data are built. The final method was therefore not used to calculate the reported F_1 scores on the pure-component data but only applied to the experimentally studied mixtures in this work (cf. Supporting Information for details).

3.4 | Experimental methods

^{1}H NMR, ^{13}C NMR, and ^{13}C DEPT NMR spectra with pulse angles of 90° and 135° were recorded of four test mixtures on an 80 MHz (proton frequency) benchtop NMR spectrometer (Spinsolve 80 Carbon Ultra) from Magritek. The experimental time for recording the ^{1}H NMR spectra of one sample was approx. 0.25 h. The total experimental time for recording the ^{13}C NMR and ^{13}C DEPT NMR spectra for one sample was about 13 h. Note that the focus of the present work was not on the time efficiency of the measurements; furthermore, obtaining a sufficient signal-to-noise ratio depends strongly on the

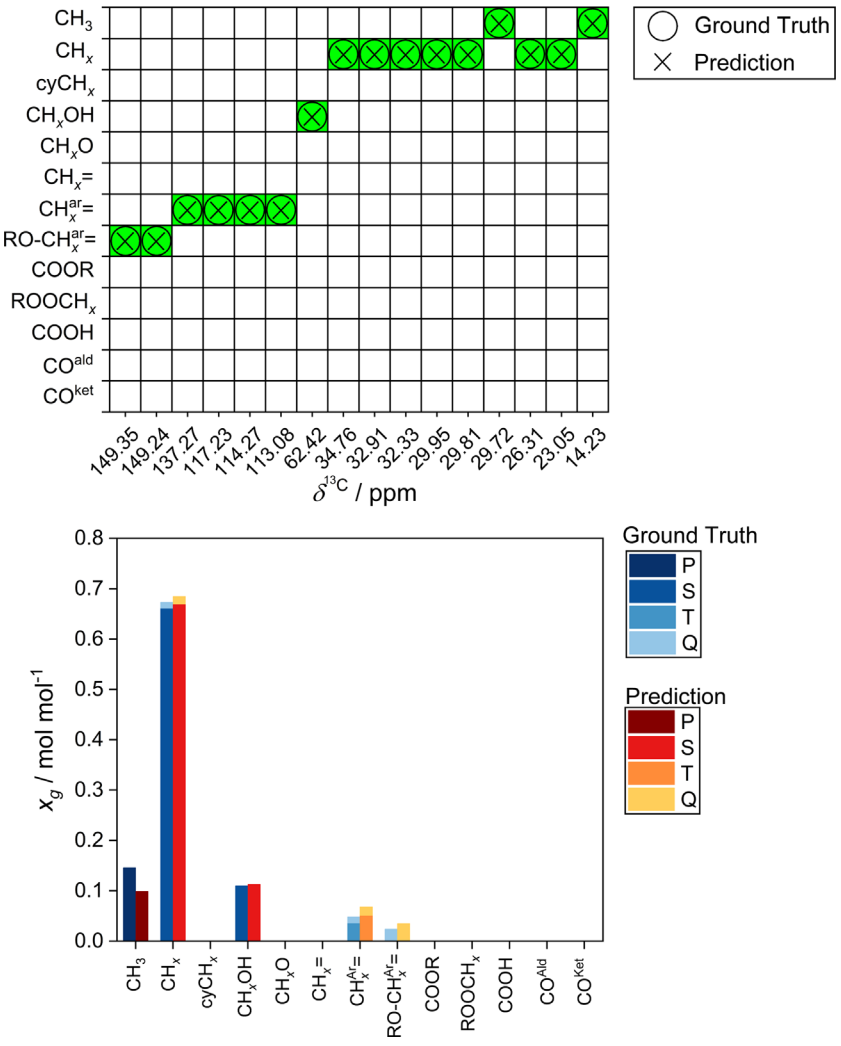

FIGURE 6 Results of the application of NMR fingerprinting to mixture III (cf. Table 2). (a) Prediction of structural groups and assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated. (b) Prediction of mole fractions of structural groups by integration of the peaks in the ^{13}C NMR spectrum; the results are shown individually for the different substitution degrees of the carbon atoms (P, S, T, and Q).

concentration of the samples. Quantitative information was obtained by integrating the respective peaks in the quantitative ^{13}C NMR spectrum, if applicable. In addition, the substitution degree of each carbon atom was determined from the ^{13}C DEPT NMR spectra of the respective mixture. Details are given in the Supporting Information.

4 | RESULTS AND DISCUSSION

4.1 | Prediction of structural groups from pure-component spectra

Figure 3 shows the results for the F_1 scores of the method for identifying structural groups and assigning the respective peaks to the different sections in the ^{13}C NMR spectrum of the pure-component data set of this work. White cells indicate group/section combinations with zero positive examples in the data set, and shaded cells indicate group/section combinations with one to nine positive examples in the data set. However, the sections with so little data were not considered in the evaluation for the reasons explained in the previous section.

Overall, high F_1 scores (>0.8) were obtained for all group/section combinations with generally higher F_1 scores for group/section combinations with larger numbers of positive examples in the data set. In the region 0–40 ppm, excellent results were obtained: The CH_3, CH_x, and $cyCH_x$ groups can be distinguished reliably. Excellent accuracy was also obtained in the region 190–210 ppm, where, for example, the CO^{ald} group can be differentiated from the CO^{ket} and COOH groups (cf. Figure 2). Some other groups, such as CH_xOH,

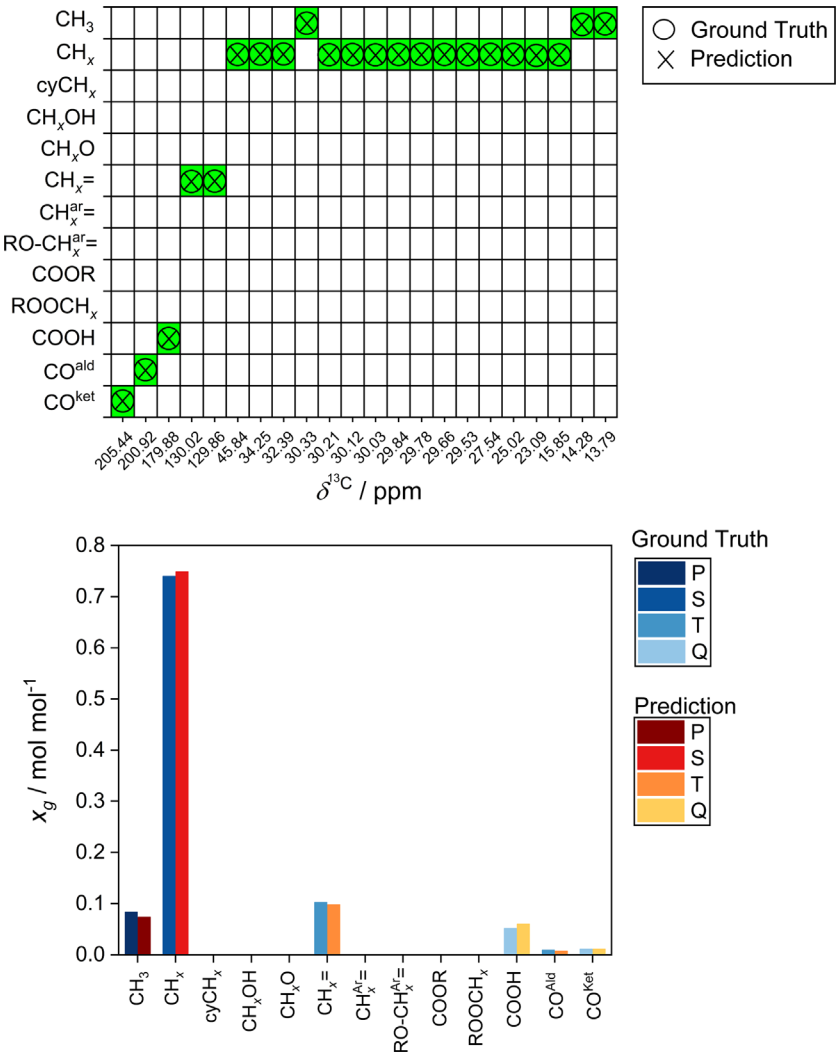

FIGURE 7 Results of the application of NMR fingerprinting to mixture IV (cf. Table 2). (a) Prediction of structural groups and assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the *x*-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated. (b) Prediction of mole fractions of structural groups in mixture IV by integration of the peaks in the ^{13}C NMR spectrum; the results are shown individually for the different substitution degrees of the carbon atoms (P, S, T, and Q).

CH_xO, and $ROOCH_x$, are harder to distinguish but are still reasonably predicted. In Figure S.11 in the Supporting Information, additional results for a variant of the method that does not use information about the presence or absence of labile protons are shown, which are slightly worse, particularly for the COOH and CH_xOH groups.

4.2 | Prediction of structural groups from mixture spectra

In the following, the results of applying the NMR fingerprinting method to four test mixtures are shown. Two aqueous and two organic mixtures were chosen for this purpose; the components used for preparing these mixtures were selected randomly but in a way that most of the structural groups (except CH_xO groups) covered by the developed method (cf. Table 1) were represented in at least one of the mixtures. Table 2 summarizes information on the mixtures studied as examples here.

4.2.1 | Results for aqueous mixtures

Because the signal-to-noise ratio of the aqueous mixtures was relatively low due to the high dilution (cf. Table 2) and the low magnetic field strength of the benchtop NMR spectrometer, a signal enhancement strategy, namely, based on the nuclear Overhauser effect (NOE), was used for obtaining the shown results; in consequence, no quantitative results were obtained here.

Figure 4 shows the results for applying the method to mixture I. All structural groups in the mixture were correctly identified and assigned to the respective peaks in the ^{13}C NMR spectrum. In Figure S.12 in the Supporting Information, results for the variant of the method without using prior information about the presence or absence of labile protons are presented; the same holds for the other studied mixtures discussed in the following.

Figure 5 shows the results of applying the method to mixture II. In this case, the $cyCH_x$ groups were misinterpreted as CH_x groups, which can be considered a minor error in many applications. All other groups are identified correctly.

4.2.2 | Results for organic mixtures

In Figure 6a, the results for identifying structural groups in mixture III are shown, which was accomplished correctly for all groups. Quantitative results, namely, the concentration of all identified structural groups in the form of group mole fractions x_g, are shown in Figure 6b. The differences between the predicted group mole fractions and the ground truth can mainly be attributed to the experimental error of the NMR spectra indicated by the signal-to-noise ratio.

Figure 7a shows the respective results for the identification of structural groups in mixture IV. Again, all structural groups were predicted correctly. Also, the agreement between the predicted mole fractions of the structural groups and the ground truth is excellent, as shown in Figure 7b.

5 | CONCLUSIONS

In this work, a method for the group-specific qualitative and quantitative analysis of unknown samples based on standard NMR experiments (^1H NMR, ^{13}C NMR, and ^{13}C DEPT NMR) is presented. The method is fully automated and requires no prior information on the samples and practically no expert knowledge, apart from that to carry out the experiments, which could also be automated. From the spectra, only the chemical shifts of the peaks, which can usually be picked in a semi-automatic way, are needed as input for the identification of the groups. If also peak areas are supplied, quantitative results on the group composition are provided. Furthermore, no expensive high-field NMR devices are needed; benchtop NMR devices are sufficient. In future work, it would be interesting to combine the NMR fingerprinting with an automated NMR acquisition: the experiments are automatically conducted, the spectra are automatically processed, and the interpretation of the data is presented.

The method is particularly interesting for applications in which information on the complete speciation of the sample is difficult to obtain, for example, in biotechnology or refinery technology. In such applications, the method opens up new routes for process monitoring and process and quality control. It could also be beneficial in the field of biofluid analysis in combination with machine-learning (ML) methods for clinical diagnosis. Furthermore, the results from the method also provide a basis for quantitative physical modeling of mixtures with group-contribution methods—without having to know the complete speciation. This paper describes the method and its background. However, we have also made it freely available for testing and application via an interactive website (https://nmr-fingerprinting.de) with a graphical user interface and a tutorial.

While the method already has an extensive range of applicability, it is still not more than a first version. It can presently identify 13 different structural groups containing only C, H, and O atoms. The assignment of the peaks

to these groups is accomplished by support vector classification, which was trained on a large set of NMR data of pure components from the literature. The initial tests carried out in the present work confirm the method's reliability, but more extensive testing would be desirable. Furthermore, the studied samples only contained components composed of the groups in the list. In practice, this may not be the case and would inevitably lead to misassignments. In cases where false groups were identified, they were similar to the actual groups. This is simply due to the nature of NMR spectroscopy and the origin of the chemical shift: we can expect a polar group to be falsely interpreted as another polar group, which may not result in significant deviations if the properties of the mixtures are predicted based on the fingerprint.

Still, the list of groups needs to be extended in future work, which should also involve groups containing atoms other than C, H, and O. The definition of the structural groups via SMARTS, as used in the present work, makes the method very flexible so that new groups can easily be included, simply by specifying a SMARTS pattern and subsequently retraining the method. However, while such an extension is straightforward, it requires including knowledge from other NMR experiments to ensure a meaningful differentiation of the groups, such as heteronuclear multiple bond correlation (HMBC) or diffusion-ordered spectroscopy (DOSY). Another issue is the presence of labile protons that are exchanging rapidly between components, such as the protons of hydroxyl groups. They lead to broad peaks in the ^{1}H NMR spectrum that may strongly shift if experimental conditions vary. The user must presently identify such peaks (which is generally an easy task) and which are excluded from the peak list.

A general problem in extending the method is the availability of suitable training data. However, this could be relaxed in future work by the augmentation of incomplete experimental data sets by predicted NMR spectra, for example, using hierachically ordered-spherical description of environment (HOSE) methods,[40] density-functional theory (DFT) calculations,[41] or ML approaches.[42]

ACKNOWLEDGMENTS

F.J. gratefully acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG) under Grant JI 401/1-1. The authors gratefully acknowledge financial support from the Carl-Zeiss-Stiftung. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Thomas Specht https://orcid.org/0000-0002-0462-8968
Justus Arweiler https://orcid.org/0009-0001-2222-0755
Johannes Stüber https://orcid.org/0009-0001-3789-5835
Kerstin Münnemann https://orcid.org/0000-0001-5247-8856
Hans Hasse https://orcid.org/0000-0003-4612-5995
Fabian Jirasek https://orcid.org/0000-0002-2502-5701

REFERENCES

[1] M. Elyashberg, K. Blinov, S. Molodtsov, Y. Smurnyy, A. J. Williams, T. Churanova, *J. Cheminformatics* **2009**, *1*(1), 1.
[2] D. C. Burns, E. P. Mazzola, W. F. Reynolds, *Nat. Prod. Rep.* **2019**, *36*(6), 919.
[3] M. Elyashberg, D. Argyropoulos, *Magn. Res. Chem.* **2020**, *59*(7), 669.
[4] Z. Huang, M. S. Chen, C. P. Woroch, T. E. Markland, M. W. Kanan, *Chem. Sci.* **2021**, *12*(46), 15329.
[5] M. Valli, H. M. Russo, A. C. Pilon, M. E. F. Pinto, N. B. Dias, R. T. Freire, I. Castro-Gamboa, V. da Silva Bolzani, *Phys. Sci. Rev.* **2019**, *4*(10), 20180108.
[6] R. Behrens, E. Kessler, K. Münnemann, H. Hasse, E. von Harbou, *Fluid Phase Equilib.* **2019**, *486*, 98.
[7] D. Bellaire, H. Kiepfer, K. Münnemann, H. Hasse, *J. Chem. Eng. Data* **2020**, *65*(2), 793.
[8] J.-N. Dumez, *Chem. Commun.* **2022**, *58*, 13855.
[9] Y. Lee, Y. Matviychuk, B. Bogun, C. S. Johnson, D. J. Holland, *J. Magn. Res.* **2022**, *335*, 107138.
[10] M. Lin, M. J. Shapiro, *Anal. Chem.* **1997**, *69*(22), 4731.
[11] Y. Lu, F. Hu, T. Miyakawa, M. Tanokura, *Metabolites* **2016**, *6*(2), 19.
[12] Y. Matviychuk, E. Steimers, E. von Harbou, D. J. Holland, *J. Magn. Res.* **2020**, *319*, 106814.
[13] Y. Matviychuk, E. Steimers, E. von Harbou, D. J. Holland, *Magn. Res.* **2020**, *1*(2), 141.
[14] Y. Matviychuk, E. von Harbou, D. J. Holland, *J. Magn. Res.* **2017**, *285*, 86.
[15] M. I. Osorio-Garcia, D. M. Sima, F. U. Nielsen, U. Himmelreich, S. V. Huffel, *J. Chemom.* **2011**, *25*(4), 183.
[16] A. A. Smith, *J. Biomol. NMR* **2017**, *67*(2), 77.
[17] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka, P. Giraudeau, *J. Magn. Res.* **2019**, *298*, 91.
[18] N. Schmid, S. Bruderer, F. Paruzzo, G. Fischetti, G. Toscano, D. Graf, M. Fey, A. Henrici, V. Ziebart, B. Heitmann, H. Grabner, J. D. Wegner, R. K. O. Sigel, D. Wilhelm, *J. Magn. Res.* **2023**, *347*, 107357.
[19] A. F. Tawfike, C. Viegelmann, R. Edrada-Ebel, *in Metabolomics Tools for Natural Product Discovery*, Methods in Molecular Biology, Humana Press, Totowa, NJ **2013**, 227–244. https://doi.org/10.1007/978-1-62703-577-4_17
[20] A. Bakiri, J. Hubert, R. Reynaud, S. Lanthony, D. Harakat, J.-H. Renault, J.-M. Nuzillard, *J. Nat. Prod.* **2017**, *80*(5), 1387.
[21] A. Bruguière, S. Derbré, J. Dietsch, J. Leguy, V. Rahier, Q. Pottier, D. Bréard, S. Suor-Cherer, G. Viault, A.-M. L. Ray, F. Saubion, P. Richomme, *Anal. Chem.* **2020**, *92*(13), 8793.
[22] M. Elyashberg, *TrAC Trends Anal. Chem.* **2015**, *69*, 88.
[23] F. Jirasek, J. Burger, H. Hasse, *Ind. Eng. Chem. Res.* **2018**, *57*(21), 7310.
[24] F. Jirasek, J. Burger, H. Hasse, *Chem. Eng. Sci.* **2019**, *208*, 115161.
[25] F. Jirasek, J. Burger, H. Hasse, *Ind. Eng. Chem. Res.* **2019**, *58*(21), 9155.

[26] F. Jirasek, J. Burger, H. Hasse, *AIChE J.* **2020**, *66*(2), e16826.
[27] T. Specht, K. Muennemann, F. Jirasek, H. Hasse, *Fluid Phase Equilib.* **2020**, *516*, 112604.
[28] T. Specht, K. Münnemann, H. Hasse, F. Jirasek, *Phys. Chem. Chem. Phys.* **2023**, *25*, 10288.
[29] M. Edgar, B. C. Percival, M. Gibson, F. Jafari, M. Grootveld, *Diabetes Res. Clin. Pract.* **2021**, *171*, 108554.
[30] J. Leenders, M. Grootveld, B. Percival, M. Gibson, F. Casanova, P. B. Wilson, *Metabolites* **2020**, *10*(4), 155.
[31] B. C. Percival, M. Grootveld, M. Gibson, Y. Osman, M. Molinari, F. Jafari, T. Sahota, M. Martin, F. Casanova, M. L. Mather, M. Edgar, J. Masania, P. B. Wilson, *High-Throughput* **2019**, *8*(1), 2.
[32] K. P. C. Vollhardt, N. E. Schore, *Organic Chemistry: Structure and Function*, 8th ed., Macmillan Learning, New York **2018**.
[33] T. Specht, K. Münnemann, H. Hasse, F. Jirasek, *J. Chem. Inf. Model.* **2021**, *61*(1), 143.
[34] *Daylight Theory Manual, Version 4.9*, Daylight Chemical Information Systems, Inc., Aliso Viejo, CA. https://www.daylight.com/dayhtml/doc/theory/index.html (Last accessed: 12.12.2022).
[35] D. Weininger, *J. Chem. Inf. Model.* **1988**, *28*(1), 31.
[36] S. Kuhn, N. E. Schlörer, *Magn. Res. Chem.* **2015**, *53*(8), 582.
[37] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, J. L. Markley, *Nucleic Acids Res.* **2007**, *36*, D402.
[38] T. D. W. Claridge, *High-Resolution NMR Techniques in Organic Chemistry*, Elsevier, Amsterdam, Netherlands **2016**. https://doi.org/10.1016/c2015-0-04654-8
[39] RDKit: Open-Source Cheminformatics, https://www.rdkit.org (Last accessed: 13.12.2022).
[40] W. Bremser, *Anal. Chim. Acta* **1978**, *103*(4), 355.
[41] A. Bagno, F. Rastrelli, G. Saielli, *Chem. - A European J.* **2006**, *12*(21), 5514.
[42] Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. S. John, R. S. Paton, *Chem. Sci.* **2021**, *12*(36), 12012.
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[44] M. Stone, *J. R. Stat. Soc.: Ser. B (Methodol.)* **1974**, *36*(2), 111.
[45] T. Head, M. Kumar, H. Nahrstaedt, G. Louppe, I. Shcherbatyi, scikit-optimize v0.9.0, https://scikit-optimize.github.io (Last accessed: 20.04.2023).
[46] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, A. Bahamonde, *Progr. Artif. Intell.* **2012**, *1*(4), 303.
[47] C. Bishop, *Pattern Recognition and Machine Learning*, 1st ed., Springer, New York **2006**.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: T. Specht, J. Arweiler, J. Stüber, K. Münnemann, H. Hasse, F. Jirasek, *Magn Reson Chem* **2024**, *62*(4), 286. https://doi.org/10.1002/mrc.5381