Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: J. Wagner, K. Munneman, T. Specht, H. Hasse and F. Jirasek, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00490J.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D5DD00490J

Deep Set Model for the Automated NMR Fingerprinting of Unknown Mixtures

Jens Wagner, Kerstin Münnemann, Thomas Specht, Hans Hasse, and Fabian Jirasek*

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

E-mail: fabian.jirasek@rptu.de

Phone: +49 (0)631 - 205 4685

Abstract

Elucidating unknown mixtures is a critical challenge in chemistry and chemical engineering. Nuclear magnetic resonance (NMR) spectroscopy is a powerful analytical technique generally suited for this purpose. However, component-wise elucidation with NMR is tedious for complex mixtures, requires expert knowledge, and often yields ambiguous results. In contrast, identifying and quantifying structural groups in a mixture from NMR spectra is much more straightforward. In prior work, we have introduced 'NMR fingerprinting' for the automated elucidation of carbon-, hydrogen-, and oxygen-containing structural groups in unknown mixtures based on standard NMR experiments and a support vector classification (SVC) from machine learning (ML). In the present work, we present a substantially advanced NMR fingerprinting method that employs a deep set model (DSM), addressing major shortcomings of the SVC, and integrates additional information from 2D NMR experiments. The DSM was trained on experimental NMR spectra of pure components from open-source databases, augmented with synthetic spectral data, and comprises invariant and equivariant network

View Article Online
DOI: 10.1039/D5DD00490J

structures to ensure predictions independent of the input order of the NMR signals. Tested on experimental pure-component test data, the DSM performs excellently, significantly outperforming our previous approaches. Furthermore, we demonstrate the applicability of the DSM to unknown mixtures by predicting the structural groups from NMR spectra of test mixtures measured using a benchtop NMR spectrometer. The predictions agree very well with the true mixture compositions, highlighting the method's potential for efficient automated mixture analysis and providing a reliable basis for downstream tasks, such as thermodynamic modeling using group-contribution methods.

Introduction

Complex mixtures of unknown compositions containing unknown components are ubiquitous in chemistry and chemical engineering, constituting a stiff challenge to process design and optimization. Nuclear magnetic resonance (NMR) spectroscopy is a powerful analytical technique well-suited for component elucidation, particularly if used with computer-assisted structure elucidation (CASE) programs,[1] which follow a set of predefined rules and incorporate predicted spectra to propose possible molecular structures. However, CASE programs often require some prior knowledge of the molecular formula of the component to be identified, typically obtained by high-resolution mass spectrometry,[2–5] adding experimental complexity to their application. Recently, machine-learning (ML) approaches that rely solely on NMR information[6–8] or incorporate additional spectroscopic input[9,10] have emerged as promising alternatives. Nonetheless, all these methods are restricted to identifying pure components, severely limiting their applicability in chemical engineering practice, where mixtures are usually present.

NMR spectroscopy has also successfully been applied for the qualitative and quantitative analysis of mixtures.[11–16] If the mixture components are known, a variety of automated quantification methods are available, even if signals in the NMR spectra overlap.[17–23] How-

View Article Online
DOI: 10.1039/D5DD00490J

ever, elucidating unknown components in mixtures remains a significant challenge, whose solution often depends on expert knowledge, which becomes infeasible if complex mixtures are studied. In cases where mixtures contain unknown components with signal overlap, already the first step of separating the relevant signals, the so-called deconvolution of the NMR spectrum, becomes inherently ambiguous, though some ML approaches for automated deconvolution of NMR spectra have been introduced.[24–26]

An alternative approach to elucidating components in complex mixtures that avoids the ambiguities of assigning the signals to the unknown components is dereplication,[27–32] which identifies individual components by comparing the NMR spectrum of the mixtures to those of pure compounds retrieved from reference databases. However, the limited coverage of these databases confines dereplication to those molecules already represented within them.[33] Moreover, methods relying solely on spectral comparisons remain sensitive to experimental conditions due to inherent biases in the reference data.[4] Consequently, no broadly applicable solution currently exists for the automated elucidation of unknown components in mixtures by NMR spectroscopy.

While, for the reasons discussed above, elucidating *components* in unknown mixtures by NMR spectroscopy still poses a significant challenge, identifying the *structural groups* that constitute these components is considerably more straightforward. This group-based task, which we call 'NMR fingerprinting', is based on the fact that in an NMR spectrum, the chemical shift of an analyzed nucleus reflects its electronic environment, thereby revealing the structural group containing it. Traditionally, chemical shift tables that outline characteristic ranges in NMR spectra have been used to assign structural groups to NMR signals.[34] However, overlapping characteristic ranges in chemical shift tables lead to ambiguity in assigning structural groups based solely on them. Also, the "static" nature of these tables leads to problems in practice.

From an ML perspective, assigning the correct structural group to signals in an NMR spectrum represents a classification problem. Therefore, we have recently developed a sup-

View Article Online
DOI: 10.1039/D5DD00490J

port vector classification (SVC) for the automated NMR fingerprinting of carbon-, hydrogen-, and oxygen-containing structural groups in unknown mixtures based on standard NMR experiments.[35,36] Trained on thousands of pure-component spectra from the open-source databases Biological Magnetic Resonance Data Bank (BMRB)[37] and NMRShiftDB,[38] the SVC automatically assigns structural groups to signals in ^{13}C NMR spectra, leveraging additional information from ^{1}H and ^{13}C DEPT (distortionless enhancement by polarization transfer) NMR spectroscopy. Utilizing SMARTS[39] strings as a machine-readable representation of the respective structural groups during model training enables straightforward modification and extension of the considered structural group list. Applied to test mixtures, the predictions by the SVC achieved good agreement with the true mixture compositions, making it a reliable method for the structural group elucidation of unknown mixtures. The results of NMR fingerprinting can subsequently be used for the rational definition of pseudo-components[40] and thermodynamic modeling using group-contribution methods,[41–45] enabling the conceptual design of fluid separation processes.[46,47]

However, due to the characteristics of NMR data, SVC-based NMR fingerprinting has significant limitations in its application. Specifically, the signals in the NMR spectrum classified into structural groups can vary substantially in number, depending on the complexity and number of different components in the mixture of interest. This poses a challenge for developing SVCs for NMR fingerprinting, as an SVC requires inputs of constant length, which we have solved by binning the NMR spectra in multiple regions of defined chemical shift width. However, binning leads to the problem that signals with very similar chemical shifts, which in consequence are assigned to the same bin, cannot be distinguished, leading to classification errors. Furthermore, while there are natural choices for ordering the NMR signals in the input of the ML models, particularly with increasing chemical shift, it is not guaranteed that all data sets consistently comply with this ordering. Similarly, there is no inherent physical order of the different NMR spectra, e.g., ^{1}H, ^{13}C, ^{13}C DEPT. Since SVCs are not permutation-invariant, i.e., their results depend on the input order, this poses

View Article Online
DOI: 10.1039/D5DD00490J

another source of error for the NMR fingerprinting.

Within the realm of ML, these properties suggest that NMR signals and their corresponding nuclei information are best modeled as elements of sets rather than as fixed-length data instances.[48]

In this work, we overcome these limitations by developing a classification model based on a deep-set architecture.[49] Deep set models (DSM) are a specialized neural network (NN) class within the field of geometric deep learning,[50] specifically designed to preserve the symmetries inherent in set-structured data while introducing only minimal additional model complexity.[48] Our DSM incorporates both invariant and equivariant network structures, ensuring that predictions are independent of input size and permutation, allowing the model to efficiently handle the unordered and variable-sized nature of NMR signals and nuclei. To fully capture the set-based characteristics of the NMR data, we extend our approach by incorporating information on the carbon–hydrogen correlations from ^{1}H $-^{13}$ C HSQC (heteronuclear single quantum coherence) NMR spectroscopy as the first 2D NMR experiment in our mixture analysis. In doing so, the HSQC information is not directly used as additional input to the DSM but instead serves to construct the set structure of the model input by linking the information gathered from the ^{1}H and ^{13}C NMR experiments.

Additionally, we address the challenge of limited and incomplete training data in the used open-source NMR databases by augmenting incomplete NMR spectra with information derived from magnetically identical nuclei and predicted spectra using the open-source tools RDkit[51] and NMRium.[52] In this way, we have obtained complete spectral information for 2767 pure components, which we have used to train the model and rigorously test its predictive performance exclusively on unseen experimental NMR spectra. Finally, we have applied the model to test mixtures whose spectra were measured using a 60 MHz benchtop NMR device, demonstrating the approach in practical low-field NMR applications.

View Article Online
DOI: 10.1039/D5DD00490J

Methods

Overview

Figure 1 provides an overview of the NMR fingerprinting method developed in this work to predict the structural groups and assign them to signals in the ^{13}C NMR spectra of unknown samples using additional information from ^{1}H, ^{13}C DEPT, and ^{1}H $-^{13}$ C HSQC NMR experiments. Central to our method is the DSM, which integrates invariant and equivariant network architectures to ensure predictions independent of the input order of the NMR signals and their associated nuclei information. The DSM was trained on the NMR spectra of 2767 pure components taken from the open-source databases BMRB[37] and NMRShiftDB.[38] To address the issue of incomplete spectral data within these databases, we employed augmentation techniques that utilize information from magnetically equivalent nuclei identified via RDKit[51] and synthetic spectra predicted using NMRium.[52] The details on the individual steps of our NMR fingerprinting are explained in the following subsections.

Currently, the NMR fingerprinting method distinguishes 13 structural main groups. The method can also distinguish between different substitution degrees, so that, in total, 30 different subgroups can be identified, which are the same as in our previous works[36,44] and summarized in Table 1. The quantification of the identified structural groups is finally achieved through signal integration in the ^{13}C NMR spectra.[36,44]

View Article Online
DOI: 10.1039/D5DD00490J

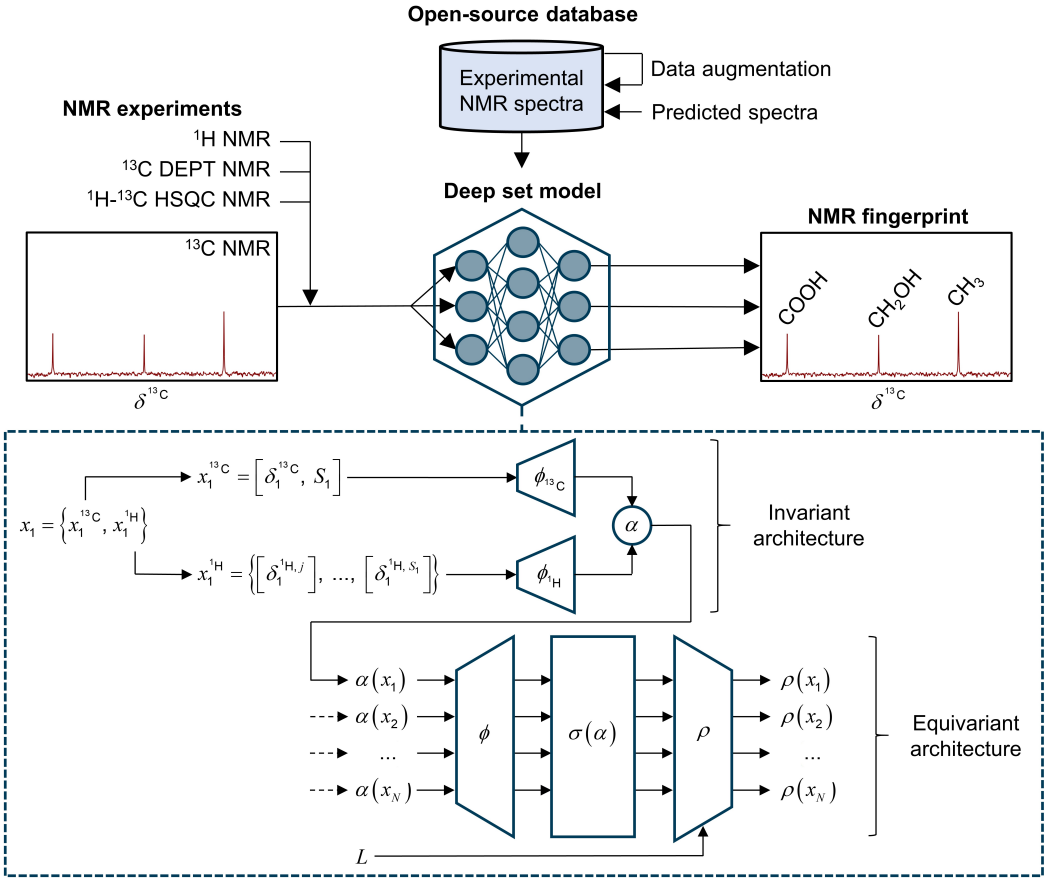

Figure 1: Overview on the NMR fingerprinting method based on a deep set model (DSM)[49] architecture for predicting the structural groups in unknown mixtures from NMR spectra and assigning them to signals in the ^{13}C NMR spectrum. The DSM was trained on pure-component NMR spectra from the open-source databases BMRB[37] and NMRShiftDB,[38] with missing information augmented from magnetically identical nuclei and predicted spectra using the open-source tools RDkit[51] and NMRium.[52] The architecture of the DSM and its input, which is obtained from NMR experiments, are described in Section Deep-set Architecture.

View Article Online
DOI: 10.1039/D5DD00490J

Table 1: Structural groups distinguished by the DSM developed in the present work, with SMARTS[39] strings for their machine-readable representation. Each structural group contains exactly one carbon atom. x determines the substitution degrees that the DSM can distinguish. The SMARTS strings are the same as in prior work of our group.[36]

Label	Structural group	SMARTS representation
CH_3	Methyl	[CX4;D1;!$(C[!#6])]
CH_x	Alkyl; $x \in \{0,1,2\}$	[CX4;D2,D3,D4;!$(C[!#6]);!R]
CH_x^{cy}	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_{ar}^x	Aromatic carbon; $x \in \{0,1\}$	[cX3;!$(c~[!#6])]
RO-CH_{ar}^x	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)

Deep-set Architecture

The input of the DSM consists of a set of x_i, where i denotes one of the N signals in the ^{13}C NMR spectrum of the studied sample, cf. Figure 1. Each x_i contains NMR-spectroscopic information on the respective ^{13}C nucleus associated with that signal ($x_i^{^{13}\mathrm{C}}$) and on the ^{1}H nuclei directly bonded to it ($x_i^{^1\mathrm{H}}$). Specifically, $x_i^{^{13}\mathrm{C}}$ contains the respective chemical shift $\delta_i^{^{13}\mathrm{C}}$ determined from the ^{13}C NMR spectrum and the substitution degree S_i derived from the intensities in the ^{13} DEPT 90/135 NMR spectra.[36,53] $x_i^{^1\mathrm{H}}$ comprises the chemical shifts $\delta_i^{^1\mathrm{H}}$ of the $j \ldots S_i$ ^{1}H nuclei directly bonded to the respective ^{13}C nucleus, obtained from the ^{1}H NMR spectrum of the studied sample. The ^{1}H nuclei are thereby assigned to their corresponding ^{13}C nuclei by the cross-signals observed in the ^{1}H $-^{13}$ C HSQC spectrum. Additionally, the DSM uses the boolean input L indicating the presence of labile protons in the sample, which is also obtained from the ^{1}H $-^{13}$ C HSQC spectrum.[36]

The DSM developed in this work combines invariant and equivariant network structures. In the first step, the input information $x_i^{^{13}\mathrm{C}}$ for the ^{13}C nuclei and $x_i^{^1\mathrm{H}}$ for the ^{1}H nuclei is independently processed by dedicated embedding networks $\phi_{^{13}\mathrm{C}}$ and $\phi_{^1\mathrm{H}}$, respectively. Unlike classical neural networks, these embeddings are computed in parallel rather than jointly,[49]

View Article Online
DOI: 10.1039/D5DD00490J

ensuring that the set-based nature of the nuclei data is respected. Subsequently, the nuclei embeddings are aggregated using the summation as a permutation-invariant function α, leading to the intermediate prediction $\alpha(x_i)$ for each structural group based only on NMR-spectroscopic information on the respective nuclei. By employing parallel embeddings and the permutation-invariant function α, the DSM ensures an invariant prediction independent of the input order of the nuclei information.

In the second step, the intermediate predictions $\alpha(x_i)$ are refined within the context of *all* structural groups in the studied sample to account for mutual influences on their respective NMR signals. Therefore, the intermediate predictions $\alpha(x_i)$ for each signal in the ^{13}C NMR spectrum are processed in parallel by the main embedding network ϕ, directing them to the equivariant layer $\sigma(\alpha)$. The equivariant layer $\sigma(\alpha)$ is a specialized NN layer that combines a standard per-element feed-forward layer σ with summation-based aggregation α,[54] allowing the interaction of the embedded structural group predictions $\alpha(x_i)$ while maintaining the relation between input and output.[48] This summation-based aggregation captures inter-signal relationships in the context of all signals, which is the simplest form of contextualization and does not explicitly encode pairwise interactions between individual signals, as employed, for example, in self-attention-based architectures.[48] Finally, through parallel processing by the prediction network ρ, which uses the additional input regarding the presence of labile protons L, the prediction $\rho(x_i)$ for each ^{13}C NMR signal is obtained, independent of the input order of the signals.

The DSM does not provide absolute predictions for structural groups; instead, it assigns a probability to each group in Table 1 for every ^{13}C NMR signal, with many groups receiving a probability of zero. This probability is interpreted as the model's confidence in the corresponding group assignment. The structural group with the highest probability (i.e., highest model confidence) is selected as the absolute prediction.

View Article Online
DOI: 10.1039/D5DD00490J

Augmentation of Pure-component NMR Data

Collecting and processing pure-component NMR data from BMRB and NMRShiftDB was conducted analogous to our previous work.[36] Only pure components for which the following conditions are fulfilled were considered: composed exclusively of carbon, hydrogen, and oxygen; can be unambiguously segmented into the structural groups presented in Table 1; and for which both an experimental ^{13}C and ^{1}H NMR spectrum are available.

However, some of the NMR spectra from these databases are incomplete, lacking assignments of chemical shifts to the respective nuclei. Upon closer examination, these omissions generally fall into two categories. Sometimes, only one of multiple magnetically equivalent nuclei has an assigned chemical shift. This partial assignment is likely attributable to non-standardized data structures within the databases, which manage redundant information inconsistently. In other cases, none of the magnetically equivalent nuclei have assigned chemical shifts, suggesting that the missing assignment is probably the result of human error during NMR spectra recording or evaluation. Since the DSM classifies ^{13}C signals in the context of all structural groups in the sample rather than individually, it is essential to provide complete spectral information of the components as input. To address the issue of missing spectral data, we implemented a two-step augmentation process:

1. Missing chemical shifts were supplemented by automatically identifying magnetically equivalent nuclei within each component using RDKit and adopting their corresponding spectral information from magnetically equivalent nuclei for which information was available.

2. Any remaining gaps in the spectra were filled using data from synthetic spectra predicted for each pure component with NMRium. In this step, no completely synthetic spectra were used; only existing but incomplete experimental spectra were augmented.

Through these augmentation steps, the number of pure components with complete spectral information increased from 839 to 2767, substantially extending the data set available

View Article Online
DOI: 10.1039/D5DD00490J

for model training. Additional details on the data augmentation from synthetic spectra are provided in the Supporting Information.

In Figure 2, the final augmented data set covering 2767 pure components and consisting of a total of 40838 structural groups is visualized considering the information from the ^{13}C NMR spectrum. The analogous presentation of the data set for the respective ^{1}H NMR-spectroscopic information is provided in Figure S.1 in the Supporting Information.

Figure 2a denotes the number of each of the 13 distinguished structural groups N_g (cf. Table 1) in the augmented pure-component data set, broken down to segments in the ^{13}C NMR spectrum where their respective signals occur. It is important to note that the segmentation of the spectrum used in Figure 2 is solely for visualization purposes but not used in the DSM, which is in contrast to our previous SVC-based approach, where spectral segmentation was required.[35,36] The structural groups exhibit significant overlap in their chemical shift distributions, i.e., they are not confined to specific regions but span a wide range of the ^{13}C NMR spectrum.

Figure 2b gives an overview of the proportion $P_{\text{syn}} = N_g^{\text{syn}}/N_g$ of the number of structural groups N_g^{syn} that incorporate synthetic data for either ^{13}C, ^{1}H, or both. Separate visualizations showing the distribution of P_{syn} for structural groups containing synthetic data exclusively for ^{13}C or ^{1}H are provided in Figure S.2 in the Supporting Information. Overall, structural groups containing synthetic spectral data account for 11.80 % of the entire data set, with 0.84 % containing synthetic data for ^{13}C and 11.14 % for ^{1}H. Most structural groups with synthetic data are concentrated in the regions below 80 ppm and between 110 and 140 ppm in the ^{13}C NMR spectrum. This distribution likely results from the high density of various structural groups, i.e., aliphatic, cyclic, and double bound carbon groups, in these regions for organic molecules, which complicates signal differentiation and accurate assignment of chemical shifts $\delta^{1\text{H}}$ in the crowded ^{1}H NMR spectrum (cf. Supporting Information for details). Furthermore, augmentations with synthetic data are necessary for carbonyl ketones with signals exceeding 220 ppm, as experimental spectra do not extend to

View Article Online
DOI: 10.1039/D5DD00490J

these elevated chemical shifts δ^{13C} by default.

In the Supporting Information, we provide a detailed analysis of the influence of synthetic NMR data on the training and predictive performance of the DSM, thereby demonstrating the robustness of the model with respect to the composition of the training data.

View Article Online
DOI: 10.1039/D5DD00490J

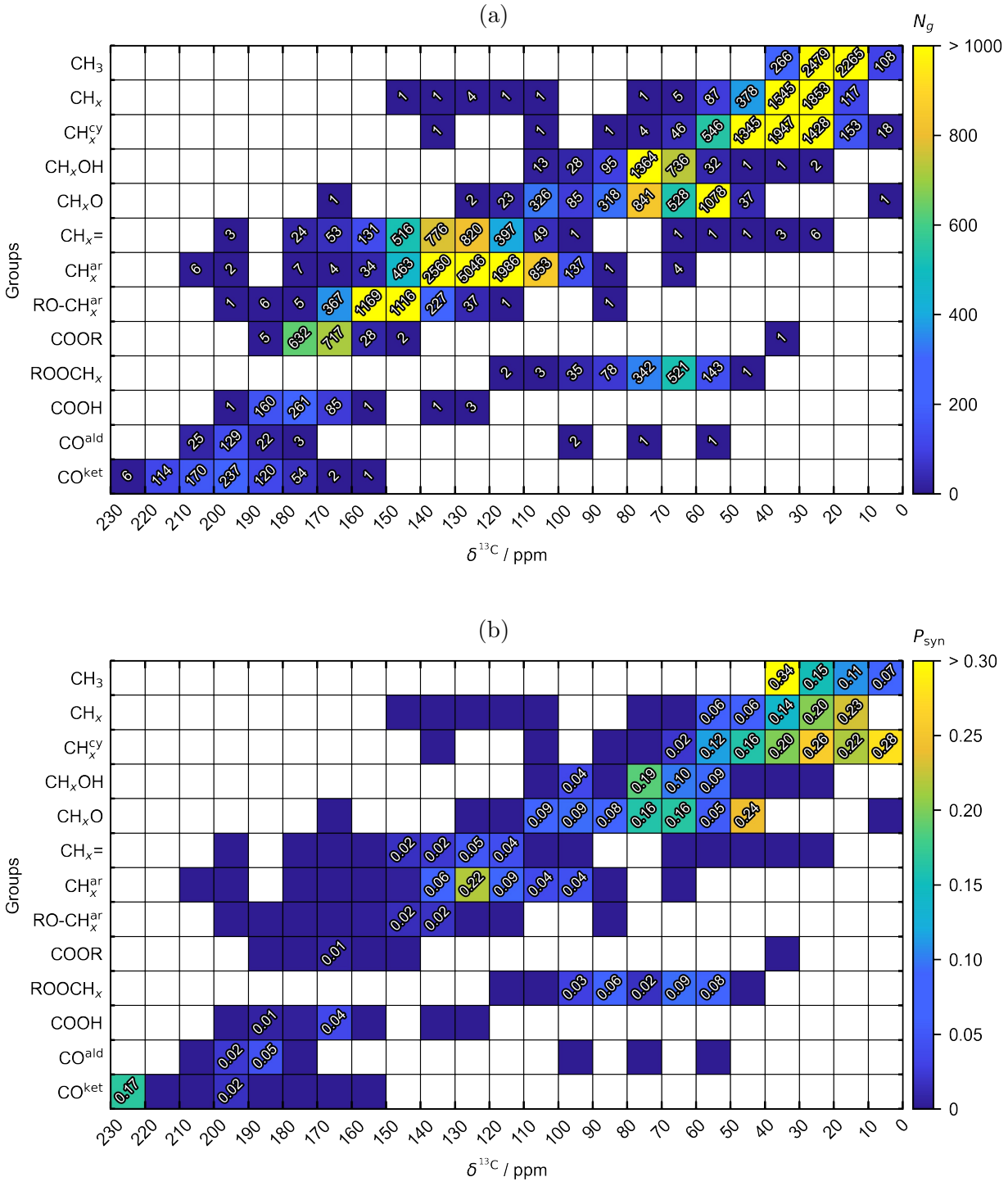

Figure 2: Distribution of the augmented pure-component data set in the ^{13}C NMR spectrum, with (a) specified and color-coded number of structural groups N_g (b) specified (greater than zero) and color-coded proportion $P_{\text{syn}} = N_g^{\text{syn}}/N_g$ of structural groups N_g^{syn} incorporating synthetic data for ^{13}C, ^{1}H, or both. The segmentation of the ^{13}C NMR spectrum is solely for visualization purposes and no requirement for the application of the DSM.

View Article Online
DOI: 10.1039/D5DD00490J

Generation of Input and Output Data

The chemical shifts for the ^{13}C and ^{1}H nuclei, the substitution degree of the ^{13}C nuclei, the boolean variable L denoting the presence or absence of labile protons, and the correct structural groups for each pure component in our data set were automatically obtained from the pure-component NMR data using RDKit, as described in our previous work.[36] The input data for each pure component and its structural groups were organized into a set-based structure, as illustrated in Figure 1 and detailed in Section Deep-set Architecture. This set-structured input data was automatically generated by determining the carbon-hydrogen connections based on the pure-component structures using RDKit. The associated output data was generated by one-hot encoding the structural groups contained in the respective pure component, as defined in Table 1.

Training and Evaluation of the Deep Set Model

We trained and evaluated the DSM using the generated pure-component data set to assess predictive performance and robustness. The measured test mixture data are used solely to demonstrate the practical applicability of the method.

The generated data set was randomly split into a training, a validation, and a test set, comprising 80 %, 10 %, and 10 % of the pure components from our data set, respectively. The test set was constrained to include only pure components with entirely experimental spectral data, i.e., not including synthetic data, to demonstrate the model's performance in the most realistic scenario. Furthermore, the training set was augmented by synthetic binary and ternary mixture data obtained by simply "mixing" the spectra of the respective pure components in the training set. As a result, each pure component present in the training set appeared three times in the final training set: once with its pure component spectra, once with the spectra of a binary mixture with a randomly chosen other component from the training set, and once with the spectra of a ternary mixture with two randomly chosen other components. In Table S.1 in the Supporting Information, we provide an analysis

View Article Online
DOI: 10.1039/D5DD00490J

demonstrating the robustness of the DSM to different random splits of the data set.

All models and scripts for training and evaluation were implemented in Python 3.6.8 using PyTorch 2.2.1.[55] Training was performed on an A40 GPU using the CrossEntropyLoss function with default PyTorch settings. The Adam optimizer was employed for weight optimization, and a learning rate scheduler with a decay factor of 0.1 and a patience of 20 epochs based on validation loss was utilized. Training was terminated early if the validation loss did not improve for 30 consecutive epochs, and the model achieving the lowest validation loss was selected. Typical training times ranged between one and two hours, while typical inference times were between four and six milliseconds.

Hyperparameter optimization, including the weight decay λ of the Adam optimizer, the initial learning rate, the batch size, and the number of layers and nodes in each network, was performed using a grid search based on validation loss. In the Supporting Information, we discuss the sensitivity of the model to the varied hyperparameters and present the validation loss results. The following hyperparameters were selected as final settings: a weight decay of $\lambda = 5 \cdot 10^{-4}$, an initial learning rate of $1 \cdot 10^{-4}$, and a batch size of one. The network architectures were defined with three layers containing eight nodes each for $\phi_{^{13}\mathrm{C}}$ and $\phi_{^{1}\mathrm{H}}$ and two layers containing 256 nodes each for ϕ and ρ. In all networks, the Sigmoid Linear Unit (SiLU) activation function with default PyTorch settings was applied. In all cases, the number of nodes for the equivariant layer $\sigma(\alpha)$ was chosen to match those of the networks ϕ and ρ. The input dimensions of $\phi_{^{13}\mathrm{C}}$ and $\phi_{^{1}\mathrm{H}}$ were set to five according to the input data dimension, while the network ρ included an additional node, to account for the boolean variable L indicating the presence of labile protons, and had an output dimension of 13, corresponding to the number of distinct structural groups.

The predictive performance of the DSM on unseen test data was evaluated using the F_1 score F_1, g for each structural group g:

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

View Article Online
DOI: 10.1039/D5DD00490J

where TP_g (true positive) represents the number of instances where structural group g was correctly identified by the model, FP_g (false positive) denotes the number of instances where the model incorrectly predicted the presence of structural group g when it was not present, and FN_g (false negatives) signifies the number of instances where structural group g was present but was not detected by the model. Consequently, $F_1, g = 1$ corresponds to a perfect prediction.

For comparison, we have also retrained and evaluated the SVC from our previous work[36] using this work's data set and the same partitioning of the data into training, validation, and test sets, as employed for the DSM. Further information on the training and evaluation of the SVC is provided in the Supporting Information.

Furthermore, a final version of the DSM was trained by randomly using the data for 90 % of the pure components from our data set for training and the remaining 10 % for validation. Unlike the primary evaluation approach described above, this model was not evaluated on a separate pure-component test set. Instead, it was directly applied to experimentally studied test mixtures to demonstrate its practical applicability in predicting the structural groups in real mixtures, as detailed below.

Experimental Methods

The compositions of the test mixtures studied in this work, as determined gravimetrically and used a ground truth here, are given in Table 2. Details on the chemicals and protocols used for mixture preparation are given in the Supporting Information. The test mixtures were selected so that each structural group considered in the developed NMR fingerprinting method, cf. Table 1, is represented in at least one of the mixtures.

View Article Online
DOI: 10.1039/D5DD00490J

Table 2: Test mixtures studied in this work.

Mixture	Components i	x_i mol mol^{-1}
I	Water	0.9266
	Acetone	0.0244
	Tartaric acid	0.0246
	1,4-Butanediol	0.0244
II	Diethyl ether	0.7005
	Butanal	0.1494
	Butyl acetate	0.1501
III	Cyclohexane	0.5001
	Hexene	0.300
	Diglyme	0.1999
IV	Anisole	0.7998
	1-Octanol	0.1198
	3-Methylbutan-2-one	0.0804

^{1}H NMR, ^{1}H $-^{13}$ C HSQC NMR, ^{13}C NMR, and ^{13}C DEPT NMR spectra with pulse angles of 90° and 135° were recorded for each test mixture using a 60 MHz benchtop NMR spectrometer (Spinsolve 60 Ultra, Magritek). The settings of the NMR experiments, spectral processing procedures, and extraction of spectral information are reported in the Supporting Information.

In ^{1}H $-^{13}$ C HSQC NMR spectra of mixtures, significant signal overlap is a common challenge, obscuring the cross-signals between ^{1}H and ^{13}C nuclei at low concentrations, especially when using benchtop NMR spectrometers with limited sensitivity and resolution. In our experiments, we encountered this exact problem: despite clear evidence of ^{1}H bonded to ^{13}C nuclei as determined by the substitution degree via ^{13}C DEPT NMR, in some cases, the absence of observable cross-signals in the ^{1}H $-^{13}$ C HSQC spectra prevented the determination of the chemical shifts $\delta_i^{^1\mathrm{H}}$ necessary for applying the DSM. To address this challenge, we have developed a model for the relationship between the chemical shifts $\delta_i^{^{13}\mathrm{C}}$ of ^{13}C nuclei and the chemical shifts $\delta_i^{^1\mathrm{H}}$ of their connected ^{1}H nuclei using linear regression, which we have fitted to our comprehensive data set for pure components. This regression model en-

View Article Online
DOI: 10.1039/D5DD00490J

ables the determination of the most likely value of $\delta_i^{^1\mathrm{H}}$ of the connected ^{1}H nuclei for a given $\delta_i^{^{13}\mathrm{C}}$. In cases where no cross signals for a ^{13}C signal in the $^1\mathrm{H}-^{13}\mathrm{C}$ HSQC were identified in the studied mixture but the ^{13}C DEPT results indicated that there should be a cross-signal, we have supplemented the missing experimental spectral information by estimating the $\delta_i^{^1\mathrm{H}}$ based on the respective $\delta_i^{^{13}\mathrm{C}}$ using the regression model. Further details on the regression model are provided in the Supporting Information.

The processed spectral information was then fed as input to the DSM to identify the structural groups in the test mixtures. The task here is to assign a group from Table 1 to each signal in the ^{13}C NMR spectrum. The resolution, even of the benchtop NMR spectrometer, is generally high enough to avoid that two different groups produce signals that cannot be distinguished. Even though this case cannot be strictly excluded, we do not consider it here. While the identification and assignment of structural groups is fully automated, the subsequent quantification step is currently performed manually. Quantitative information on the identified structural groups was finally obtained by manually integrating their signals in the ^{13}C NMR spectrum and calculating the group mole fractions x_g from the signal areas A_g, see Eq. 2:

$$x_g = \frac{A_g}{\sum_{g=1}^{N} A_g} \tag{2}$$

Results and Discussion

Prediction of Structural Groups from Pure-component Spectra

Figure 3 presents the $F_{1,g}$ scores of the DSM in predicting the structural groups of the pure components from the test set. The model generally achieves high $F_{1,g}$ scores, indicating high prediction accuracy, across all structural groups. Decreases in the $F_{1,g}$ score are observed only at the boundaries of the characteristic ranges in the ^{13}C NMR spectrum for each structural group. Specifically, $F_{1,g}$ scores below 0.5 are found exclusively for $\mathrm{CH}_x^{\mathrm{ar}}$, RO-$\mathrm{CH}_x^{\mathrm{ar}}$, and

View Article Online
DOI: 10.1039/D5DD00490J

ROOCH$_x$ groups, with chemical shifts δ_i^{13C} outside their characteristic ranges (cf. Figure 2). Overall, the developed DSM achieves an average F_1 score of 0.92 across all structural groups, demonstrating excellent predictive performance, while significantly outperforming our previous SVC model, attaining a F_1 score of 0.85 (cf. Figure S.3 in the Supporting Information for details).

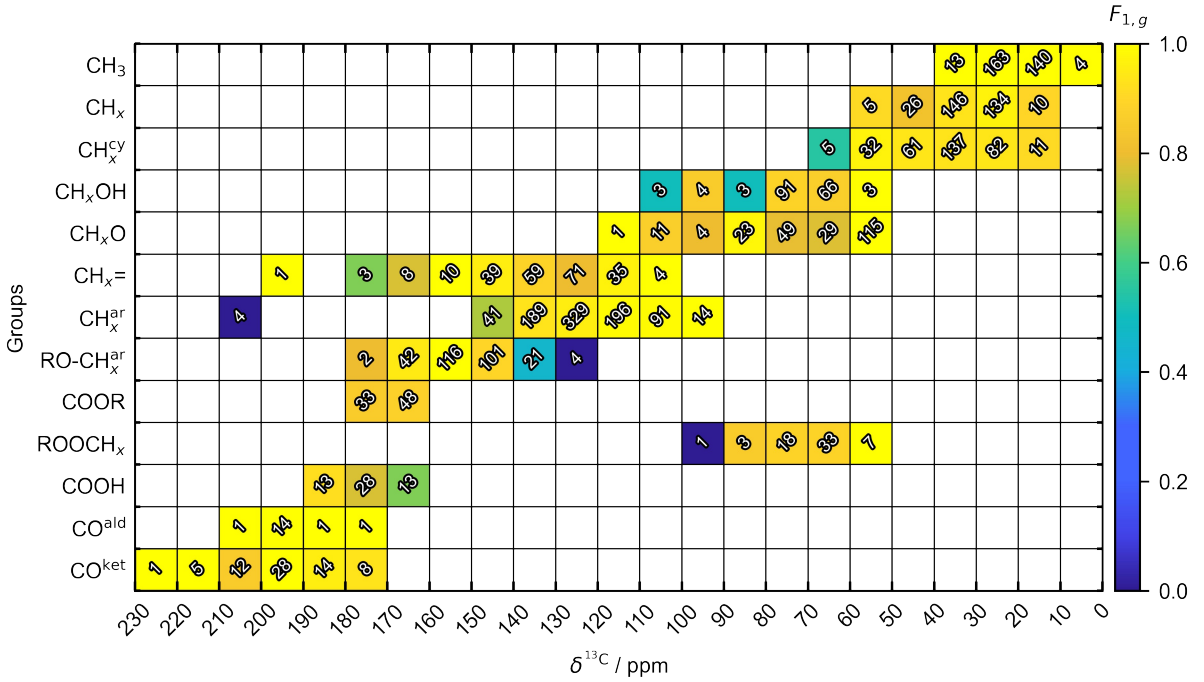

Figure 3: $F_{1,g}$ scores (indicated by color code) of the DSM in predicting the structural groups of the pure components in the test set based on NMR spectra. The numbers in the cells indicate the number of structural groups N_g per segment of the ^{13}C NMR spectrum.

Prediction of Structural Groups from Mixture Spectra

In the following, the results from the analysis of the studied test mixtures (cf. Table 2) with the DSM are presented and discussed.

View Article Online
DOI: 10.1039/D5DD00490J

Mixture I

Figure 4 shows the results for Mixture I. All structural groups in the mixture were correctly predicted and assigned to the respective signals in the ^{13}C NMR spectrum. Figure 4 also gives the DSM's confidence in each group assignment. The confidence of the assigned groups are always close to 1, indicating a high confidence of the model in its predictions. Somewhat lower numbers for the confidence are only found for the assignment of the CH_2 group at lowest chemical shift 30.77 ppm, most likely caused by the high number of possible structural groups in this range of the ^{13}C NMR spectrum (cf. Figure 2a).

View Article Online
DOI: 10.1039/D5DD00490J

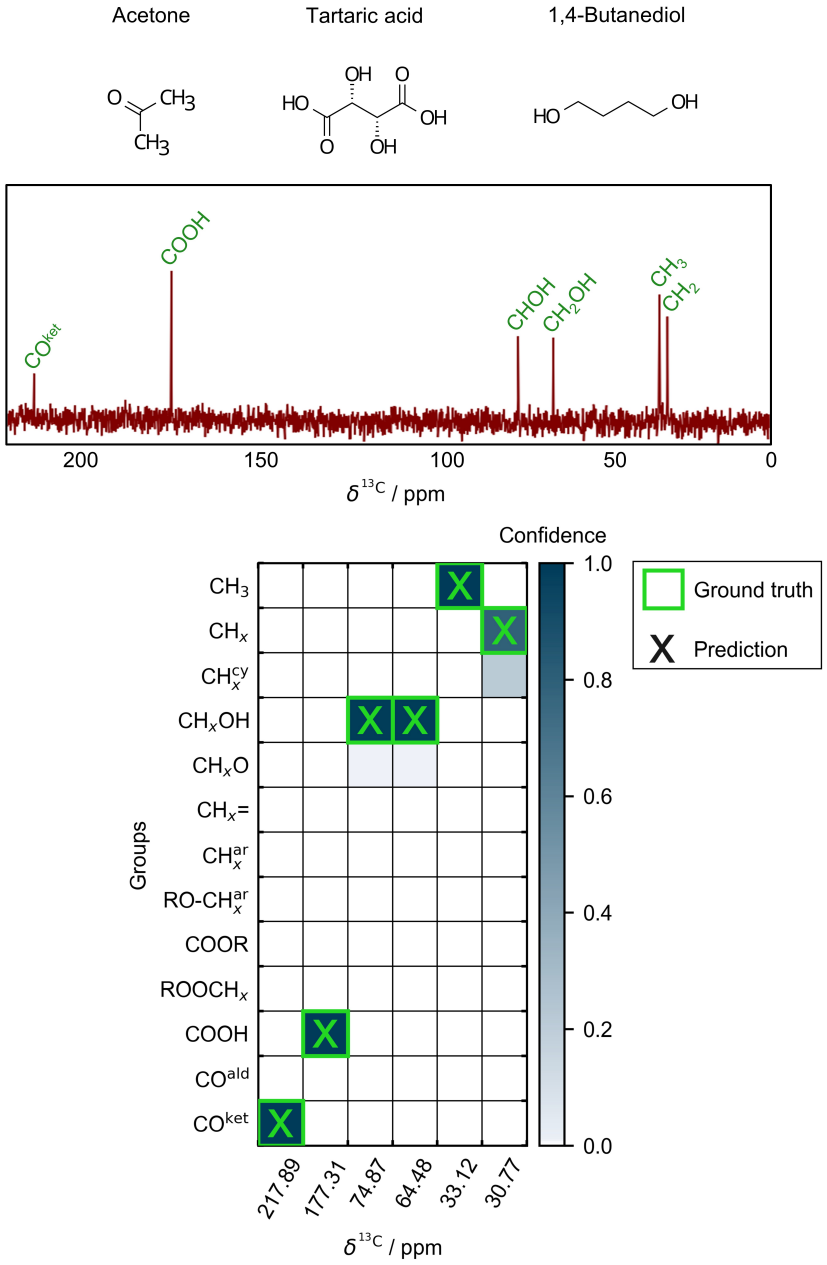

Figure 4: Results from applying the DSM to Mixture I. Top: Structural formulas of the true mixture components and assignment of the predicted structural groups, including their substitution degrees, to the corresponding ^{13}C NMR signals. Bottom: Comparison of the predicted structural groups to the ground truth. Green color indicates correct predictions, and the model's confidence is color-coded in blue.

View Article Online
DOI: 10.1039/D5DD00490J

Mixture II

Figure 5 shows the results for Mixture II. The DSM correctly predicts all structural groups in the mixture. The model's confidence in the assignment is generally high, with decreased values observed only for the two CH_3 groups at lowest chemical shifts.

View Article Online
DOI: 10.1039/D5DD00490J

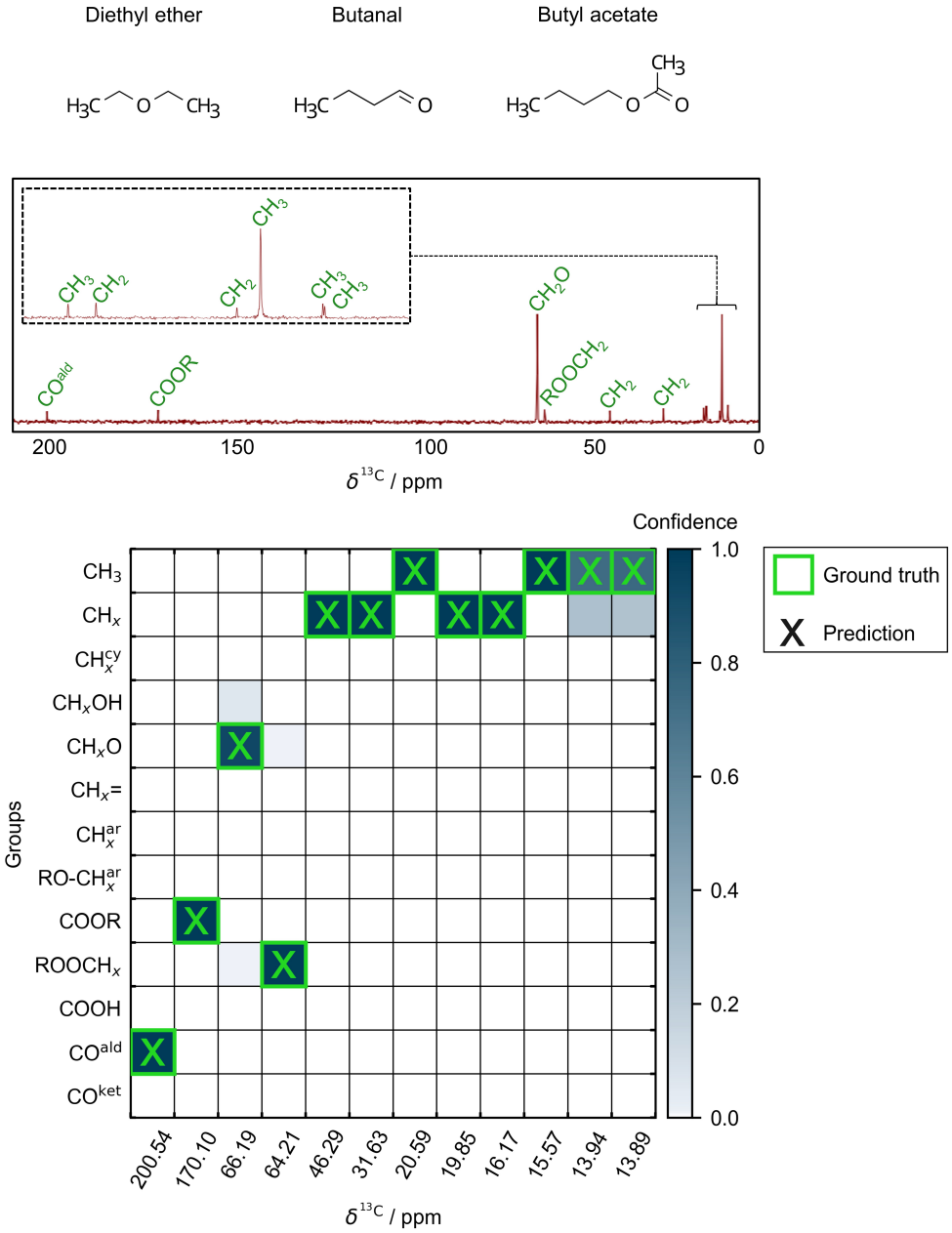

Figure 5: Results from applying the DSM to Mixture II. Top: Structural formulas of the true mixture components and assignment of the predicted structural groups, including their substitution degrees, to the corresponding ^{13}C NMR signals. Bottom: Comparison of the predicted structural groups to the ground truth. Green color indicates correct predictions, and the model's confidence is color-coded in blue.

View Article Online
DOI: 10.1039/D5DD00490J

Mixture III

Figure 6 shows the results for Mixture III. Although the DSM exhibits absolute confidence in the prediction of all structural groups, the CH_2^{cy} group at 27.56 ppm is mispredicted as a CH_2 group. The misprediction of aliphatic instead of cyclic groups is likely due to their overlapping characteristic ranges in the ^{13}C NMR spectrum, cf. Figure 2. In applying the results in group contribution methods such a misinterpretation would only have minor consequences in many cases. All other structural groups are correctly identified.

View Article Online
DOI: 10.1039/D5DD00490J

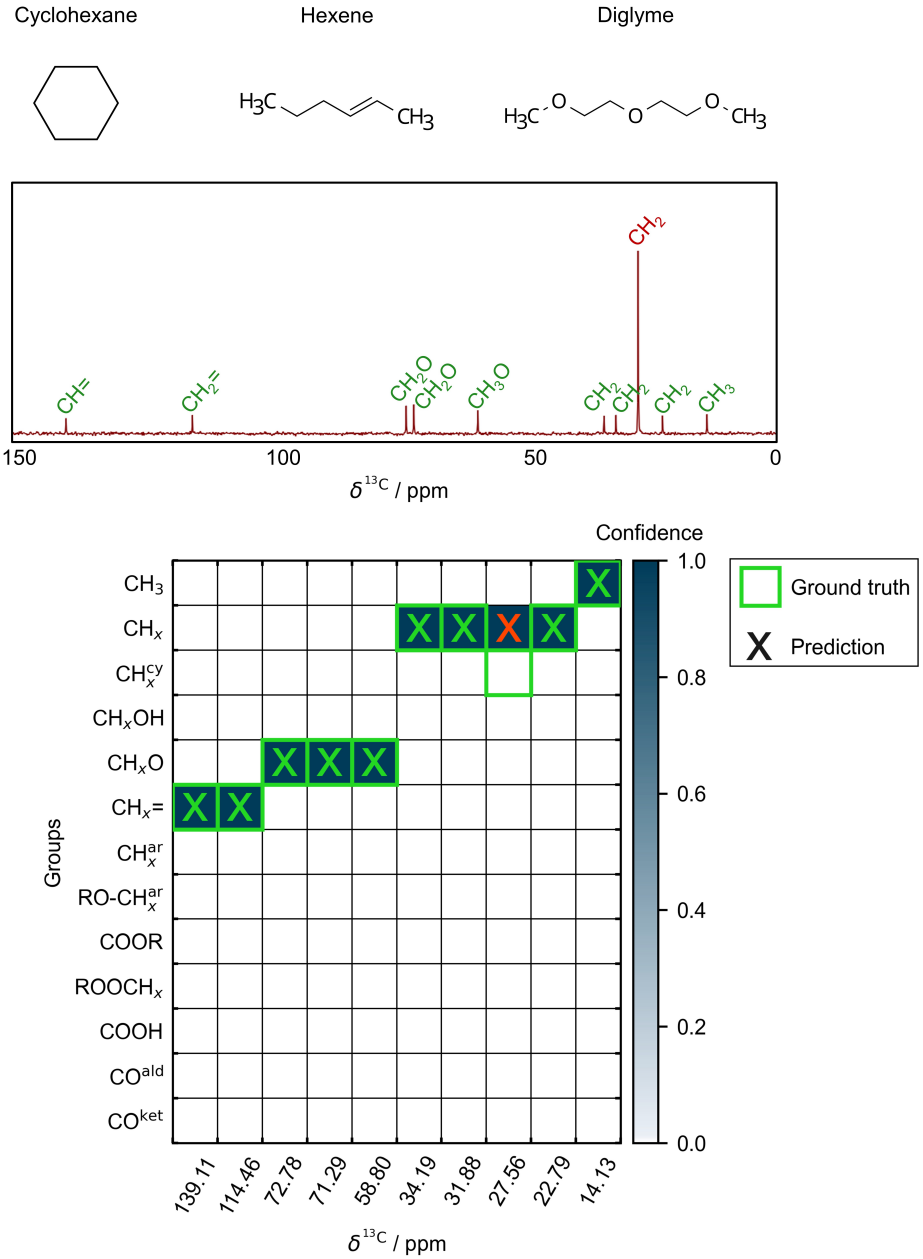

Figure 6: Results from applying the DSM to Mixture III. Top: Structural formulas of the true mixture components and assignment of the predicted structural groups, including their substitution degrees, to the corresponding ^{13}C NMR signals. Bottom: Comparison of the predicted structural groups to the ground truth. Green color indicates correct predictions, red color indicates false predictions, and the model's confidence is color-coded in blue.

View Article Online
DOI: 10.1039/D5DD00490J

Mixture IV

Figure 7 presents the results for Mixture IV. The model accurately predicts all structural groups in the mixture, except the RO-C^{ar} group at 160.21 ppm, which is mispredicted as COOR group. However, the model was not sure about this decision, with a model confidence of 0.3 for RO–C^{ar} and 0.7 for COOR. The confidence of the assignments is generally lower for the groups of Mixture IV, which is likely due to the mixture's increased number of structural groups and complexity compared to the other test mixtures.

View Article Online
DOI: 10.1039/D5DD00490J

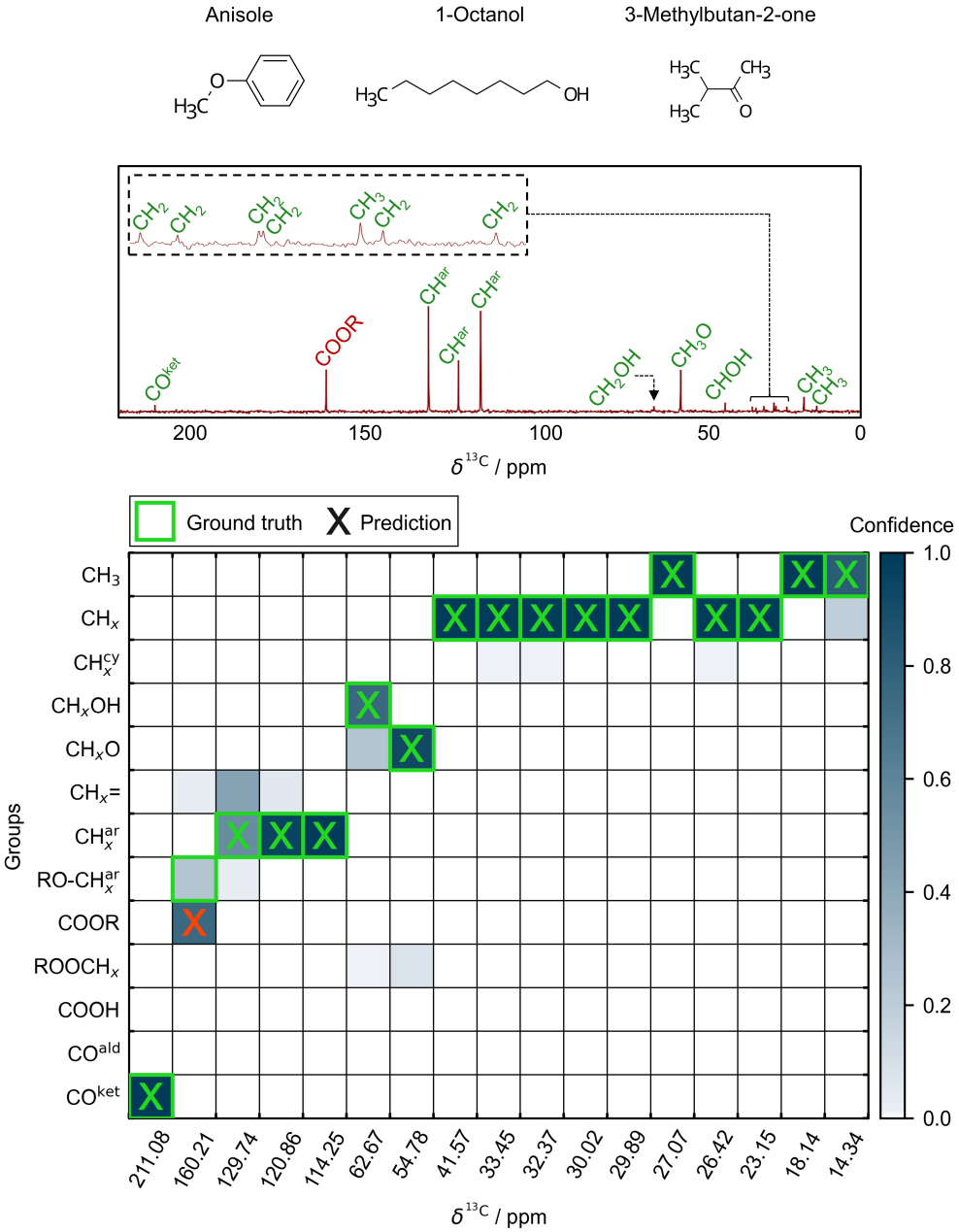

Figure 7: Results from applying the DSM to Mixture IV. Top: Structural formulas of the true mixture components and assignment of the predicted structural groups, including their substitution degrees, to the corresponding ^{13}C NMR signals. Bottom: Comparison of the predicted structural groups to the ground truth. Green color indicates correct predictions, red color indicates false predictions, and the model's confidence is color-coded in blue.

View Article Online
DOI: 10.1039/D5DD00490J

Quantitative Predictions for Test Mixtures

Figure 8 presents the quantitative results for the four test mixtures in terms of the predicted mole fractions of the structural groups x_g. For all mixtures, the predicted mole fractions x_g agree reasonably well with the ground truth. The observed discrepancies between the predicted and true mole fractions are primarily attributed to experimental errors in the NMR measurements, particularly those arising from low signal-to-noise ratios in the solvent water of Mixture I and at low concentrations.

For the test mixtures studied in this work, no overlapping ^{13}C NMR signals were observed that would affect peak integration and, consequently, the predicted mole fractions x_g. In the event of signal overlap in the analysis of a mixture, integration could be performed by integrating the observable peak envelopes according to their shapes. In cases of more pronounced overlap, recently proposed ML-based deconvolution methods[24–26] could be employed to facilitate signal separation prior to integration. If individual peaks cannot be reliably distinguished even after such treatment, the overlapping signals could be treated as a single contribution and assigned to all corresponding predicted structural groups.

View Article Online
DOI: 10.1039/D5DD00490J

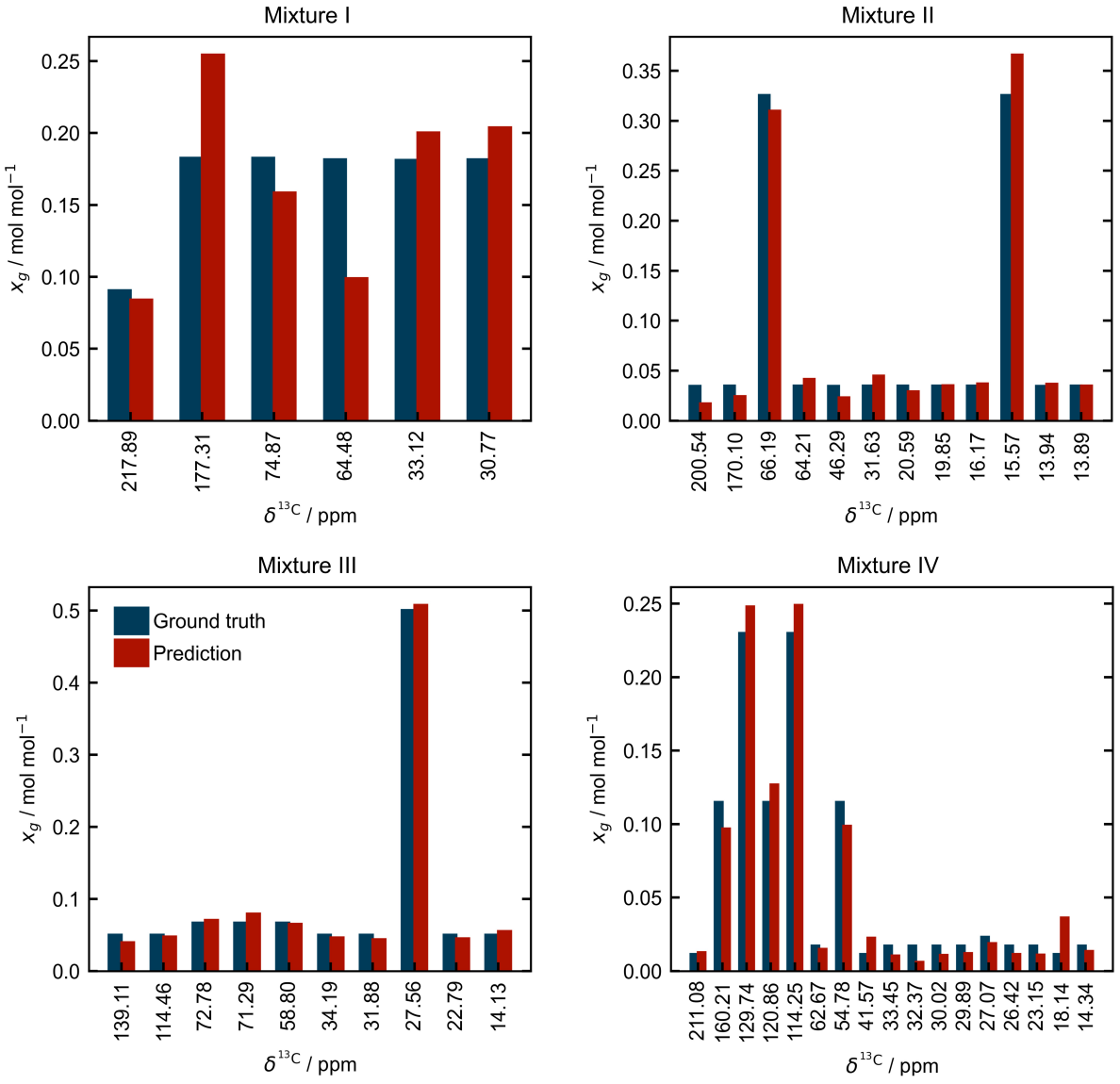

Figure 8: Comparison of the predicted mole fractions x_g for the structural groups corresponding to the signals at respective chemical shift $\delta_i^{^{13}\mathrm{C}}$ in the ^{13}C NMR spectrum of the test mixtures with the true compositions of Mixtures I - IV.

View Article Online
DOI: 10.1039/D5DD00490J

Conclusions

In this work, we have introduced a deep set model (DSM) to automatically elucidate the structural groups in unknown pure samples and mixtures using the spectral information from standard NMR experiments of the sample, a task we call NMR fingerprinting. The DSM was specifically engineered to process the characteristics of NMR data, e.g., to allow inputs of varying length as generally different numbers of NMR signals must be expected for different samples, and trained on experimental pure-component NMR spectra from open-source databases to predict the structural groups and assign them to the corresponding signals in the ^{13}C NMR spectrum of any given sample. To overcome problems with limited experimental training data, we have augmented the experimental data set with synthetic spectral data generated from predicted NMR spectra.

Furthermore, we have incorporated ^{1}H $-^{13}$ C HSQC NMR data as the first 2D NMR information into the NMR fingerprinting approach. This integration opens the possibility of mapping the structural groups identified in the ^{13}C spectrum to the corresponding signals in the ^{1}H spectrum, improving the deconvolution and interpretability of complex ^{1}H spectra of mixtures. However, achieving such mapping remains challenging due to extensive signal overlap and low resolution of ^{1}H spectra measured using benchtop NMR spectrometers. Therefore, realizing the mapping to ^{1}H spectra when applying the NMR fingerprinting method to high-field NMR spectrometers could be a goal of future work.

In scenarios where HSQC acquisition is impractical, e.g. when minimal measurement time is required, the DSM-based approach cannot be applied, as the HSQC information is essential to establish the set-based structure of the NMR input. In such cases, our previous SVC-based NMR fingerprinting approach,[36] which operates solely on 1D NMR data, can be used.

Evaluation on experimental test data for unseen pure components demonstrates the excellent performance of the DSM in predicting the structural groups of pure components, significantly exceeding our previous NMR fingerprinting approach, which was based on an

View Article Online
DOI: 10.1039/D5DD00490J

SVC. To demonstrate its applicability to unknown mixtures, we have applied the DSM to NMR spectra of test mixtures measured using a simple benchtop NMR device. The results show remarkable agreement with the true mixture compositions, demonstrating the potential of the DSM-based NMR fingerprinting method for efficient automated mixture analysis, including in low-field NMR settings, which provides the basis for thermodynamic modeling of unknown mixtures via group-contribution methods.[45,47]

Despite the high-quality predictions, the current method still has limitations. Most importantly, it is restricted to structural groups consisting of only carbon, oxygen, and hydrogen. Expanding the range of structural groups poses challenges due to the increasing overlap in their spectral ranges, which can in principle be mitigated by combining multiple spectroscopic techniques, as demonstrated for Fourier-transform infrared (FT-IR) spectroscopy.[56,57] However, the flexible architecture of the DSM allows for the incorporation of additional NMR data, e.g., from heteronuclear multiple bond correlation (HMBC) NMR, which can enhance distinguishability and enable the inclusion of further structural groups, such as those containing nitrogen, without requiring spectroscopic methods beyond NMR in future work. Furthermore, the successful integration of synthetic data in this work suggests that augmenting the training data with synthetic NMR spectra can further improve the model's predictive capabilities for new structural groups with limited available data.

View Article Online
DOI: 10.1039/D5DD00490J

Author Contributions

Jens Wagner: Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft.

Kerstin Münnemann: Funding acquisition, Resources, Writing - review & editing.

Thomas Specht: Conceptualization, Methodology, Writing - review & editing.

Hans Hasse: Funding acquisition, Resources, Supervision, Writing - review & editing.

Fabian Jirasek: Conceptualization, Funding acquisition, Resources, Supervision, Writing - review & editing.

Conflicts of Interest

There are no conflicts of interest to declare.

Data Availability

Data for this article, including the NMR spectra data, training scripts, and the final trained model, are available at Zenodo. The archived version corresponding to the manuscript is available at `https://doi.org/10.5281/zenodo.18310430`. The most recent version is available at `https://doi.org/10.5281/zenodo.17597708`.

Acknowledgement

We gratefully acknowledge financial support by the Carl Zeiss Foundation in the projects 'Process Engineering 4.0' and 'Halocycles', as well as by DFG in the Research Grant (project number 462456621), the Core Facility 'LASE-MR' (project number 537627671), and the research training group 'WERA' (project number 503479768).

View Article Online
DOI: 10.1039/D5DD00490J

Supporting Information Available

Generation of synthetic NMR data; data distribution in the ^{1}H and ^{13}C spectrum; sensitivity studies; comparison of DSM with support vector classification; experimental methods; augmented data set, training scripts, and final model.

References

(1) Elyashberg, M.; Blinov, K.; Molodtsov, S.; Smurnyy, Y.; Williams, A. J.; Churanova, T. Computer-assisted methods for molecular structure elucidation: realizing a spectroscopist's dream. *Journal of Cheminformatics* **2009**, *1*.

(2) Burns, D. C.; Mazzola, E. P.; Reynolds, W. F. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. *Natural Product Reports* **2019**, *36*, 919–933.

(3) Elyashberg, M.; Argyropoulos, D. Computer Assisted Structure Elucidation (CASE): Current and future perspectives. *Magnetic Resonance in Chemistry* **2020**, *59*, 669–690.

(4) Huang, Z.; Chen, M. S.; Woroch, C. P.; Markland, T. E.; Kanan, M. W. A framework for automated structure elucidation from routine NMR spectra. *Chemical Science* **2021**, *12*, 15329–15338.

(5) Valli, M.; Russo, H. M.; Pilon, A. C.; Pinto, M. E. F.; Dias, N. B.; Freire, R. T.; Castro-Gamboa, I.; Bolzani, V. d. S. Computational methods for NMR and MS for structure elucidation I: software for basic NMR. *Physical Sciences Reviews* **2019**, *4*.

(6) Sridharan, B.; Mehta, S.; Pathak, Y.; Priyakumar, U. D. Deep Reinforcement Learning for Molecular Inverse Problem of Nuclear Magnetic Resonance Spectra to Molecular Structure. *The Journal of Physical Chemistry Letters* **2022**, *13*, 4924–4933.

View Article Online
DOI: 10.1039/D5DD00490J

(7) Alberts, M.; Zipoli, F.; Vaucher, A. C. Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models. *ChemRxiv* **2023**, ver. 1.

(8) Hu, F.; Chen, M. S.; Rotskoff, G. M.; Kanan, M. W.; Markland, T. E. Accurate and Efficient Structure Elucidation from Routine One-Dimensional NMR Spectra Using Multitask Machine Learning. *ACS Central Science* **2024**, *10*, 2162–2170.

(9) Tan, X. A transformer based generative chemical language AI model for structural elucidation of organic compounds. *Journal of Cheminformatics* **2025**, *17*.

(10) Alberts, M.; Hartrampf, N.; Laino, T. Automated Structure Elucidation at Human-Level Accuracy via a Multimodal Multitask Language Model. 2025; `http://dx.doi.org/10.26434/chemrxiv-2025-q80r9`.

(11) Behrens, R.; Kessler, E.; Münnemann, K.; Hasse, H.; von Harbou, E. Monoalkylcarbonate formation in the system monoethanolamine–water–carbon dioxide. *Fluid Phase Equilibria* **2019**, *486*, 98–105.

(12) Bellaire, D.; Kiepfer, H.; Münnemann, K.; Hasse, H. PFG-NMR and MD Simulation Study of Self-Diffusion Coefficients of Binary and Ternary Mixtures Containing Cyclohexane, Ethanol, Acetone, and Toluene. *Journal of Chemical & Engineering Data* **2020**, *65*, 793–803.

(13) Dumez, J.-N. NMR methods for the analysis of mixtures. *Chemical Communications* **2022**, *58*, 13855–13872.

(14) Lee, Y.; Matviychuk, Y.; Bogun, B.; Johnson, C. S.; Holland, D. J. Quantification of mixtures of analogues of illicit substances by benchtop NMR spectroscopy. *Journal of Magnetic Resonance* **2022**, *335*, 107138.

(15) Lin, M.; Shapiro, M. J. Mixture Analysis by NMR Spectroscopy. *Analytical Chemistry* **1997**, *69*, 4731–4733.

View Article Online
DOI: 10.1039/D5DD00490J

(16) Lu, Y.; Hu, F.; Miyakawa, T.; Tanokura, M. Complex Mixture Analysis of Organic Compounds in Yogurt by NMR Spectroscopy. *Metabolites* **2016**, *6*, 19.

(17) Matviychuk, Y.; Steimers, E.; von Harbou, E.; Holland, D. J. Bayesian approach for automated quantitative analysis of benchtop NMR data. *Journal of Magnetic Resonance* **2020**, *319*, 106814.

(18) Matviychuk, Y.; Steimers, E.; von Harbou, E.; Holland, D. J. Improving the accuracy of model-based quantitative nuclear magnetic resonance. *Magnetic Resonance* **2020**, *1*, 141–153.

(19) Matviychuk, Y.; von Harbou, E.; Holland, D. J. An experimental validation of a Bayesian model for quantification in NMR spectroscopy. *Journal of Magnetic Resonance* **2017**, *285*, 86–100.

(20) Osorio-Garcia, M. I.; Sima, D. M.; Nielsen, F. U.; Himmelreich, U.; Van Huffel, S. Quantification of magnetic resonance spectroscopy signals with lineshape estimation. *Journal of Chemometrics* **2011**, *25*, 183–192.

(21) Smith, A. A. INFOS: spectrum fitting software for NMR analysis. *Journal of Biomolecular NMR* **2017**, *67*, 77–94.

(22) Sokolenko, S.; Jézéquel, T.; Hajjar, G.; Farjon, J.; Akoka, S.; Giraudeau, P. Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain. *Journal of Magnetic Resonance* **2019**, *298*, 91–100.

(23) Zhou, Z.; Liao, X.; Qiu, X.; Zhang, Y.; Dong, J.; Qu, X.; Lin, D. NMRformer: A Transformer-Based Deep Learning Framework for Peak Assignment in 1D 1H NMR Spectroscopy. *Analytical Chemistry* **2025**, *97*, 904–911.

(24) Schmid, N.; Bruderer, S.; Paruzzo, F.; Fischetti, G.; Toscano, G.; Graf, D.; Fey, M.; Henrici, A.; Ziebart, V.; Heitmann, B.; Grabner, H.; Wegner, J.; Sigel, R.; Wilhelm, D.

View Article Online
DOI: 10.1039/D5DD00490J

Deconvolution of 1D NMR spectra: A deep learning-based approach. *Journal of Magnetic Resonance* **2023**, *347*, 107357.

(25) Fischetti, G.; Schmid, N.; Bruderer, S.; Caldarelli, G.; Scarso, A.; Henrici, A.; Wilhelm, D. Automatic classification of signal regions in 1H Nuclear Magnetic Resonance spectra. *Frontiers in Artificial Intelligence* **2023**, *5*.

(26) Schmid, N.; Wanner, M.; Fischetti, G.; Henrici, A.; Meshkian, M.; Bruderer, S.; Füchslin, R. M.; Heitmann, B.; Wegner, J. D.; Sigel, R. K. O.; Wilhelm, D. MolDeTr: A Chemistry-Informed Deep Learning Model for Next-Generation Automated Analysis of 1H NMR Spectra. 2026; `http://dx.doi.org/10.26434/chemrxiv.10001683/v1`.

(27) Tawfike, A. F.; Viegelmann, C.; Edrada-Ebel, R. *Metabolomics Tools for Natural Product Discovery*; Humana Press, 2013; p 227–244.

(28) Bakiri, A.; Hubert, J.; Reynaud, R.; Lanthony, S.; Harakat, D.; Renault, J.-H.; Nuzillard, J.-M. Computer-Aided 13C NMR Chemical Profiling of Crude Natural Extracts without Fractionation. *Journal of Natural Products* **2017**, *80*, 1387–1396.

(29) Bruguière, A.; Derbré, S.; Dietsch, J.; Leguy, J.; Rahier, V.; Pottier, Q.; Bréard, D.; Suor-Cherer, S.; Viault, G.; Le Ray, A.-M.; Saubion, F.; Richomme, P. MixONat, a Software for the Dereplication of Mixtures Based on 13C NMR Spectroscopy. *Analytical Chemistry* **2020**, *92*, 8793–8801.

(30) Elyashberg, M. Identification and structure elucidation by NMR spectroscopy. *TrAC Trends in Analytical Chemistry* **2015**, *69*, 88–97.

(31) Jin, Y.; Wang, J.-J.; Xu, F.; Ji, X.; Gao, Z.; Zhang, L.; Ke, G.; Zhu, R.; E, W. NMR-Solver: Automated Structure Elucidation via Large-Scale Spectral Matching and Physics-Guided Fragment Optimization. 2025; `https://arxiv.org/abs/2509.00640`.

View Article Online
DOI: 10.1039/D5DD00490J

(32) Yuan, B.; Zhang, C.; Ji, C.; Liu, G.; Li, X.; Gong, S.; Huang, X.; Shen, A.; Li, X.; Liu, Y. HSQCid: A Powerful Tool for Paving the Way to High-Throughput Structural Dereplication of Natural Products Based on Fast NMR Experiments. *Analytical Chemistry* **2025**, *97*, 3227–3235.

(33) Yao, L.; Yang, M.; Song, J.; Yang, Z.; Sun, H.; Shi, H.; Liu, X.; Ji, X.; Deng, Y.; Wang, X. Conditional Molecular Generation Net Enables Automated Structure Elucidation Based on 13C NMR Spectra and Prior Knowledge. *Analytical Chemistry* **2023**, *95*, 5393–5401.

(34) Vollhardt, K. P. C.; Schore, N. E. *Organic Chemistry: Structure and Function*, 8th ed.; Macmillan Learning, 2018.

(35) Specht, T.; Münnemann, K.; Hasse, H.; Jirasek, F. Automated Methods for Identification and Quantification of Structural Groups from Nuclear Magnetic Resonance Spectra Using Support Vector Classification. *Journal of Chemical Information and Modeling* **2021**, *61*, 143–155.

(36) Specht, T.; Arweiler, J.; Stüber, J.; Münnemann, K.; Hasse, H.; Jirasek, F. Automated nuclear magnetic resonance fingerprinting of mixtures. *Magnetic Resonance in Chemistry* **2023**, *62*, 286–297.

(37) Ulrich, E. L. et al. BioMagResBank. *Nucleic Acids Research* **2007**, *36*, D402–D408.

(38) Kuhn, S.; Schlörer, N. E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories. *Magnetic Resonance in Chemistry* **2015**, *53*, 582–589.

(39) 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: 13.12.2024.

View Article Online
DOI: 10.1039/D5DD00490J

(40) Specht, T.; Münnemann, K.; Hasse, H.; Jirasek, F. Rational method for defining and quantifying pseudo-components based on NMR spectroscopy. *Physical Chemistry Chemical Physics* **2023**, *25*, 10288–10300.

(41) Jirasek, F.; Burger, J.; Hasse, H. Method for Estimating Activity Coefficients of Target Components in Poorly Specified Mixtures. *Industrial & Engineering Chemistry Research* **2018**, *57*, 7310–7313.

(42) Jirasek, F.; Burger, J.; Hasse, H. NEAT—NMR Spectroscopy for the Estimation of Activity Coefficients of Target Components in Poorly Specified Mixtures. *Industrial & Engineering Chemistry Research* **2019**, *58*, 9155–9165.

(43) Jirasek, F.; Burger, J.; Hasse, H. Application of NEAT for determining the composition dependence of activity coefficients in poorly specified mixtures. *Chemical Engineering Science* **2019**, *208*, 115161.

(44) Specht, T.; Münnemann, K.; Jirasek, F.; Hasse, H. Estimating activity coefficients of target components in poorly specified mixtures with NMR spectroscopy and COSMO-RS. *Fluid Phase Equilibria* **2020**, *516*, 112604.

(45) Wagner, J.; Romero, Z.; Münnemann, K.; Specht, T.; Jirasek, F.; Hasse, H. Thermodynamic modeling of poorly specified mixtures using NMR fingerprinting and group-contribution equations of state. *Fluid Phase Equilibria* **2025**, *596*, 114446.

(46) Jirasek, F.; Burger, J.; Hasse, H. Application of NEAT for the simulation of liquid–liquid extraction processes with poorly specified feeds. *AIChE Journal* **2019**, *66*.

(47) Specht, T.; Hasse, H.; Jirasek, F. Predictive Thermodynamic Modeling of Poorly Specified Mixtures and Applications in Conceptual Fluid Separation Process Design. *Industrial & Engineering Chemistry Research* **2023**, *62*, 10657–10667.

View Article Online
DOI: 10.1039/D5DD00490J

(48) Wagstaff, E.; Fuchs, F. B.; Engelcke, M.; Osborne, M. A.; Posner, I. Universal Approximation of Functions on Sets. 2021; https://arxiv.org/abs/2107.01959.

(49) Zaheer, M.; Kottur, S.; Ravanbakhsh, S.; Poczos, B.; Salakhutdinov, R.; Smola, A. Deep Sets. 2017; https://arxiv.org/abs/1703.06114.

(50) Bronstein, M. M.; Bruna, J.; Cohen, T.; Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2021; https://arxiv.org/abs/2104.13478.

(51) RDKit: Open-Source Cheminformatics. https://www.rdkit.org, Last accessed: 13.12.2024.

(52) Patiny, L.; Musallam, H.; Bolaños, A.; Zasso, M.; Wist, J.; Karayilan, M.; Ziegler, E.; Liermann, J. C.; Schlörer, N. E. NMRium: Teaching nuclear magnetic resonance spectra interpretation in an online platform. *Beilstein Journal of Organic Chemistry* **2024**, *20*, 25–31.

(53) Claridge, T. D. W. *High-Resolution NMR Techniques in Organic Chemistry*; Elsevier, Amsterdam, Netherlands, 2016.

(54) Soelch, M.; Akhundov, A.; van der Smagt, P.; Bayer, J. *Artificial Neural Networks and Machine Learning – ICANN 2019: Theoretical Neural Computation*; Springer International Publishing, 2019; p 444–457.

(55) Paszke, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019; https://arxiv.org/abs/1912.01703.

(56) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical Science* **2020**, *11*, 4618–4630.

View Article Online
DOI: 10.1039/D5DD00490J

(57) Lee, G.; Shim, H.; Cho, J.; Choi, S.-I. Machine-Learning Approach to Identify Organic Functional Groups from FT-IR and NMR Spectral Data. *ACS Omega* **2025**, *10*, 12717–12723.

Data Availability Statement

View Article Online
DOI: 10.1039/D5DD00490J

Data for this article, including the NMR spectra data, training scripts, and the final trained model, are available at Zenodo. The archived version corresponding to the manuscript is available at https://doi.org/10.5281/zenodo.18310430. The most recent version is available at https://doi.org/10.5281/zenodo.17597708.