RATIONAL METHOD FOR DEFINING AND QUANTIFYING PSEUDO-COMPONENTS BASED ON NMR SPECTROSCOPY *

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

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

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

Fabian Jirasek
Laboratory of Engineering Thermodynamics (LTD)
RPTU Kaiserslautern
Kaiserslautern, Germany
fabian.jirasek@rptu.de

ABSTRACT

Poorly specified mixtures, whose composition is unknown, are ubiquitous in chemical and biochemical engineering. In the present work, we propose a rational method for defining and quantifying pseudo-components in such mixtures that is free of ad-hoc assumptions. The new method requires only standard nuclear magnetic resonance (NMR) experiments and can be fully automated. In the first step, the method analyzes the composition of the poorly specified mixture in terms of structural groups, which is much easier than obtaining the component speciation. The structural groups are then clustered into pseudo-components based on information on the self-diffusion coefficients measured by pulsed-field gradient (PFG) NMR spectroscopy. We demonstrate the performance of the new method on several aqueous mixtures. The method is broadly applicable and provides a sound basis for modeling and simulation of processes with poorly specified mixtures, without the need for tedious and expensive structure elucidation. It is also attractive for process monitoring.

1 Introduction

Not being able to specify the molecular composition of a material raises many fundamental questions, such as: How can the properties of the material be estimated? How can processes with such materials be modeled, given that models usually require complete knowledge of the composition? How should the material be characterized?

Suppose an analytical elucidation of the speciation of the material is infeasible. In that case, these questions are usually targeted by one of two strategies: by specifying the way the material was obtained, which can, however, be very tedious, or, in the domain of modeling, by introducing pseudo-components, which is generally done based on ad-hoc assumptions. We deal with the issue of poorly specified liquid mixtures in the present work and propose a new, rational method to characterize them. The new method enables a meaningful quantitative characterization of the poorly specified mixtures without relying on ad-hoc assumptions.

A poor mixture specification can have different origins and manifestations: let us start from the ideal picture that all components are known both regarding their nature and their concentration, which is, unfortunately, rarely the case in practice. A typical deviation from the ideal picture is that only a small fraction of the mixture is not elucidated and that guarantees are given that the amount of these "impurities" does not exceed a certain threshold so that the properties of the mixture are not substantially influenced. In this case, the mixture can be described and modeled based on the known

* *Citation*: T. Specht, K. Münnemann, H. Hasse, F. Jirasek: Rational method for defining and quantifying pseudo-components based on NMR spectroscopy, Physical Chemistry Chemical Physics 25 (2023) 10288-10300. DOI: 10.1039/D3CP00509G.

part alone. However, there are also many cases in which less is known about the composition of the mixture and in which the unknown components influence its properties substantially; this is the case we are addressing here.

Well-known examples of such mixtures are petroleum oil fractions[1, 2, 3], polymerization products[4, 5], and fermentation broths[6]. The mixtures from these fields are generally far too complex to fully elucidate and quantify all constituent components in practice. To still describe these mixtures, pseudo-components have been used[7, 8, 9, 10, 11, 12]. There are two general ways of doing this: either a discrete set of pseudo-components is chosen, or a continuous distribution of the pseudo-components is assumed. The latter approach belongs to the field of "continuous thermodynamics"[13, 14]. The extraordinary relevance of introducing pseudo-components becomes clear by considering that all physical models of mixtures require some knowledge of the composition. Without information on the composition, the properties of poorly specified mixtures can only be correlated empirically.

The most prominent field in which poorly specified mixtures are modeled by defining pseudo-components is petrochemical engineering[8, 2, 3, 15, 16, 17, 12]. Here, a mixture, e.g., crude oil, is often divided into fractions based on their boiling points, and each fraction is then modeled as a pseudo-component[18]; such a procedure requires the physical separation of the mixture (e.g., by distillation), which is time-consuming and expensive. Choosing the pseudo-components and assigning thermodynamic properties to them is thereby generally based on ad-hoc assumptions, e.g., regarding the number and nature of the pseudo-components.

In contrast, the method we propose here is generic and allows the definition of pseudo-components in a consistent and automated way without requiring physical separation of the mixture and without relying on any ad-hoc assumptions on the number and nature of the pseudo-components. Our method may provide suggestions for pseudo-components that correspond to actual components, which, however, is not a prerequisite for successfully applying the method.

The method we propose is based on the elucidation and quantification of *structural groups* in the mixtures by nuclear magnetic resonance (NMR) spectroscopy. This is a much simpler task than elucidating and quantifying chemical components and can be accomplished swiftly also for complex mixtures. The output of the method, namely, the composition of an a priori unknown mixture in terms of its groups and their assignment to pseudo-components, can be used in group-contribution methods for predicting the mixture properties. Such methods are available for many thermodynamic properties, most notably for activity coefficients[19]. If the groups identified in the NMR analysis are the same as those used in the thermodynamic group-contribution method, the application of the results from the NMR spectroscopy is straightforward; in other cases, a mapping is needed, which can usually be found. Furthermore, there is some flexibility in the choice of the groups identified by the NMR spectroscopy, which can be used for an adaption to the thermodynamic task. The prediction of the thermodynamic properties of the mixture by group-contribution methods provides a basis for quantitative process modeling and simulation.

We demonstrate the applicability of our method by considering several complex aqueous mixtures as test cases, but the approach can also be applied to non-aqueous mixtures. The composition of all test mixtures was known from sample preparation, but this information was not used for the predictions - it was only used for evaluating the results. The new method paves the way for thermodynamic modeling of poorly specified mixtures without requiring the elaborate analytical elucidation of the composition.

2 Overview of the Method

The proposed method can be divided into two general steps, starting with identifying and quantifying structural groups in the poorly specified mixture and ending with defining and quantifying the pseudo-components. Based on this, in a subsequent step, which is not considered in the present work, predictive thermodynamic modeling of the properties of the mixture can be carried out. Figure 1 visualizes the workflow.

In the first step, which we call "NMR fingerprinting", the poorly specified mixture is analyzed by quantitative ^{13}C NMR spectroscopy and ^{13}C distortionless enhancement by polarization transfer (DEPT) NMR spectroscopy yielding a quantitative group-specific characterization, cf. Figure 1 (upper panel). In principle, also other experiments could be (additionally) used in this step, e.g., by NMR spectroscopy with other nuclei like ^{1}H or infrared (IR) spectroscopy. However, using ^{13}C NMR has the great advantage of high shift dispersion leading to only few overlapping signals also in spectra of complex mixtures.

For identifying different structural groups in this step, we make use of the fact that the position of a signal in an NMR spectrum is characteristic for the chemical environment of the nucleus that is observed, i.e., for the structural group in which it is located. The simplest way for assigning signals to structural groups is using chemical shift tables[22], namely assigning distinct structural groups to fixed regions of chemical shift of the NMR spectrum. In the present work, such a simple approach was applied, based on the shift table of the ^{13}C NMR spectrum used in our previous works[23, 24, 25, 26], which was, however, refined here as described in the following.

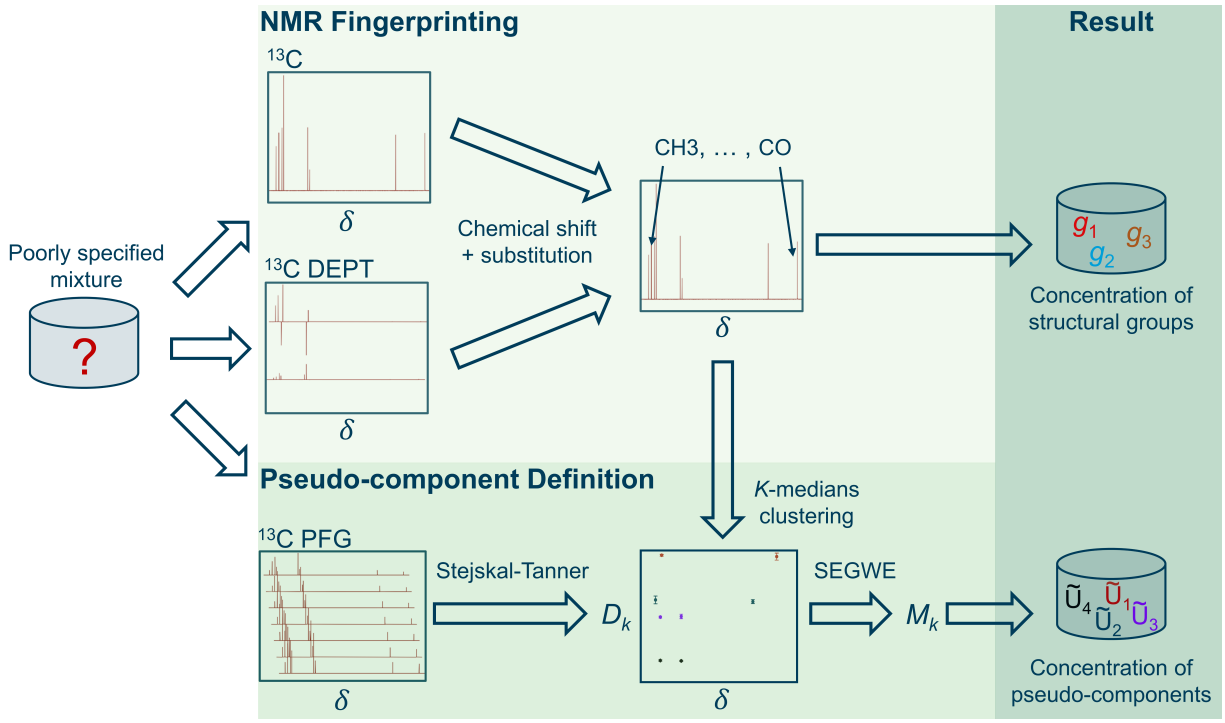

Figure 1: Scheme of the proposed method for the rational definition of pseudo-components. δ denotes the chemical shift. D_k and M_k denote the self-diffusion coefficient and the molar mass, as estimated with the SEGWE model[20, 21] of pseudo-component k, respectively.

In contrast to our previous work, the method in this work also relies on information from ^{13}C DEPT NMR spectra, which allows determining the substitution degree of structural groups, e.g., to differentiate between a primary (e.g., 'CH3') and a secondary (e.g., 'CH2') group. We only note that also more sophisticated approaches like, e.g., the machine-learning model from our previous work[27], could be used for this purpose, which will be interesting to combine with the method of this work in the future; this, however, will require extending the approach from Ref.[27] to also include information from DEPT NMR spectra.

The concentrations of all identified structural groups were then determined by integration of the corresponding peaks in the quantitative ^{13}C NMR spectrum. The obtained quantitative characterization with respect to structural groups can already be of great practical interest, e.g., for process and reaction monitoring. However, for other applications, e.g., the thermodynamic modeling of phase equilibria with poorly specified mixtures, group-specific information is not sufficient but pseudo-components need to be defined in a rational way.

We achieve this with the second step of our method, where the structural groups are clustered to multiple pseudo-components, cf. Figure 1 (lower panel) based on pulsed-field gradient (PFG) NMR spectroscopy. PFG NMR spectroscopy is a routine technique for measuring self-diffusion coefficients and was shown to yield accurate results for pure components and mixtures[28, 29, 30, 31]. PFG NMR spectroscopy can, of course, support the elucidation of components, and has been applied for this purpose, see, e.g., Refs. [32, 33, 34, 35, 36, 37]. However, to our knowledge, it has not been applied for defining pseudo-components yet.

For clustering the structural groups to pseudo-components, information on the self-diffusion coefficients of the groups are used, which are determined by ^{13}C PFG NMR experiments of a poorly specified mixture in this work. In ^{13}C PFG NMR spectroscopy, the overlap of signals is significantly reduced compared to ^{1}H PFG NMR, which is particularly relevant if complex mixtures are studied.[38, 34, 35, 39]However, ^{13}C PFG NMR spectroscopy has lower sensitivity and longer relaxation times than ^{1}H PFG NMR, which results in longer experimental times needed for a sound analysis. The named disadvantages of ^{13}C PFG NMR can, in principle, be partially compensated by using polarization transfer techniques, namely, DEPT, as demonstrated in Ref.[39]; however, since the signals of quaternary carbons would thereby be suppressed, we have not used this approach here.

The ratio behind using PFG NMR spectroscopy in this work is that groups *on the same molecule* inevitably have *the same self-diffusion coefficient*; hence, groups with similar self-diffusion coefficients can be clustered to pseudo-components.

Note that the inverse is not necessarily valid, as different components may have similar self-diffusion coefficients. Still, a clustering based on self-diffusion coefficients seems a natural choice, as, in the worst case, components of a similar size and nature are lumped together.

One challenge in the clustering task is that the results of the PFG NMR experiments are subject to uncertainties and there is no guarantee that the structural groups cluster unambiguously into a certain number of pseudo-components. We have, therefore, decided to use an unsupervised machine-learning technique for this purpose.

Specifically, we propose to use K-medians clustering, which is a variant of the K-means clustering algorithm[40], relying on both the values of the self-diffusion coefficients and of their uncertainties as inputs.

A second challenge in the clustering task is that the *number* of clusters, i.e., pseudo-components in our case, is a priori unknown. This is solved by using the so-called silhouette score[41], which is an unsupervised measure for the quality of a clustering, to predict a suitable number of clusters.

Based on the obtained clustering and together with quantitative information on the structural groups (cf. upper panel in Figure 1), this allows determining the *relative* amount of each structural group in a pseudo-component. This, however, still allows different solutions for the molar mass of each pseudo-component, as the ratio between groups in a pseudo-component is the same for arbitrary multiples of the molar mass. Fortunately, the measured self-diffusion coefficients also contain information on the molar mass of each pseudo-component. Corresponding relations are encoded in predictive models for diffusion coefficients, such as the Stokes-Einstein Gierer-Wirtz estimation (SEGWE) model[20, 21], which was used in the present work.

However, directly applying models such as SEGWE for the purpose of predicting the molar mass of pseudo-components from diffusion coefficients is hampered by two issues: firstly, the diffusion coefficient depends not only on the diffusing species (the pseudo-component here) but also on the solvent, which is a priori unknown as it is basically the poorly specified mixture here. However, in many practical applications, it will be valid to assume that the solvent predominantly contains only one component that is known, e.g., water. Then, the calculation can be made assuming that the solvent is just that main component. If this is not a valid assumption, there is always the option to carry out the PFG NMR measurements on a sample that has been strongly diluted with a known solvent, which can then be taken as the main solvent.

The second issue is that the diffusing component needs to be highly diluted, as basically all diffusion models, including the SEGWE model, predict diffusion coefficients only in the state of infinite dilution. Also this issue can in general be tackled by strongly diluting the sample prior to the diffusion measurement (at the cost of lower signal intensities of the remaining components). As an alternative to address this issue, we propose using the concept of relative diffusion coefficients[42, 43, 44, 45, 46], i.e., to relate the diffusion coefficient of a pseudo-component to a diffusion coefficient of a known component in the same mixture and for which the diffusion coefficient at infinite dilution is experimentally known (or can be calculated using predictive models). If such a component is not present in the mixture, it can always be added.

To summarize, the results after the two steps of our proposed method are:

1. The definition of a set of pseudo-components in terms of their group-composition and molar mass.
2. The concentration of the pseudo-components.

Of course, the defined pseudo-components may, in the best case, be identical with true components, but they may also be made up from several true components. If knowledge on one or more true components is a priori available (which will be the case in many practical problems), the procedure described above can be modified to accommodate that information, which is highly welcome.

After having accomplished the two steps of our method, the situation for the poorly specified mixture is technically the same as for any fully specified mixture: the constituents and their concentrations are known. Hence, predictive thermodynamic models can be applied for calculating the properties of the poorly specified mixture, namely using thermodynamic group-contribution methods. Such applications are not in the scope of the present paper and will be discussed in future work.

2.1 NMR Experiments

To characterize the poorly specified mixture, three types of standard NMR experiments are carried out for the proposed method:

1. ^{13}C NMR spectroscopy.

2. ^{13}C DEPT NMR spectroscopy.
3. ^{13}C PFG NMR spectroscopy.

In this work, the poorly specified mixtures were analyzed without any pretreatment. In principle, also an internal standard can be added to facilitate quantification, or the solution can be diluted in a known solvent, as mentioned in Section 2. In the ESI, Figures S.1-S.3, ^{13}C NMR spectra of the mixtures are shown. Details are given in the ESI in Section "*Experimental Methods*".

The experimental time for recording the quantitative ^{13}C NMR and ^{13}C DEPT NMR spectra was in total below 11 h in all cases, whereas the time for carrying out the ^{13}C PFG NMR experiments was below 41 h in all cases. We note that, in the present work, we did not focus on time efficiency, which will, however, be addressed in future work, e.g., by using fewer gradient steps in PFG NMR[45] or by exploiting polarization transfer techniques like polarization enhancement nurtured during attached nucleus testing (PENDANT)[47] in combination with PFG NMR. Moreover, a further reduction of measurement time can be achieved by the addition of T_1 relaxation agents, which significantly shorten the relaxation time of carbon nuclei and therefore enables faster accumulation of signal.[48, 49] Furthermore, an extensive analysis by PFG NMR might only be necessary once, e.g., for the feed prior to or at the beginning of a process.

2.2 Identification and Quantification of Structural Groups

In most practical situations that involve poorly specified mixtures, at least some information on the composition is available. Therefore, the decision on which structural groups to consider can be based on this prior knowledge as well as on the intended application, and can be tailored to the specific situation.

Given our background in thermodynamics, we have chosen a set of organic structural groups as it is used in a widely applied thermodynamic group-contribution method, the UNIFAC-method[19], for the regions in the chemical shift tables. This choice can be considered as an example for the application of our method. Furthermore, we only consider groups containing C, H, and O atoms in the present work. Extensions and adaptions of our method to other sets of chemical groups are straightforward.

UNIFAC distinguishes between 'main-groups' and associated 'sub-groups', where the sub-groups that belong to the same main-group usually only differ in the substitution degree of the carbon in the group. Information on the substitution degree can be obtained from DEPT NMR spectra, such that the DEPT NMR spectra are in particular helpful for distinguishing different sub-groups. More details can be found in the ESI.

Table 1 summarizes the structural groups considered in the present examples and their assignment to regions in the ^{13}C NMR spectrum. Note that some groups, e.g., the 'OH' group, show no signals in ^{13}C NMR spectroscopy but still can be determined by the characteristic shift of the neighboring alkyl group.

Table 1: Assignment of structural groups from the UNIFAC[19] table to regions of chemical shift in the ^{13}C NMR spectrum. A distinction between structural groups in the same chemical shift region was made by classifying each carbon as primary (P), secondary (S), tertiary (T), or quaternary (Q) by considering a quantitative ^{13}C NMR spectrum and ^{13}C DEPT NMR spectra, cf. ESI.

^{13}C NMR chemical shift region	Carbon	UNIFAC label	Group
0-60 ppm	P	CH3	alkyl
	S	CH2	
	T	CH	
	Q	C	
60-90 ppm	S	CH2 + OH	alcohol
	T	CH + OH	
	Q	C + OH	
90-150 ppm	S/T	CH=CH	alkenyl
	Q	C=C	
150-180 ppm	Q	COOH	carboxylic acid
>180 ppm	T	CHO	aldehyde
	Q	CH3CO/CH2CO	(alkyl + ketone)

The limitations imposed by strictly assigning a single structural group (combination) to each region can be relaxed in future work by using ML techniques[27]. After the identification of the structural groups, they can be quantified, which

was done here based on group mole fractions x_g:

$$x_g = \frac{\frac{A_g}{z_g}}{\sum_{g=1}^{G} \frac{A_g}{z_g}} \tag{1}$$

where A_g is the total area of all peaks associated to structural group g in the mixture, z_g is the number of NMR-active nuclei in the respective group g in the same chemical shift region, and G is the total number of distinguished structural groups, which is $G = 11$ here, cf. Table 1. After appropriate processing of the spectra, manual peak integration was conducted to obtain the areas A_g, which was sufficient in all cases here. In more complex cases, advanced peak fitting techniques can be employed[50, 51, 52]. This, however, will usually not be necessary in ^{13}C NMR spectroscopy. More technical details about the identification and quantification of structural groups can be found in the ESI.

2.3 Clustering of Structural Groups to Pseudo-components

2.3.1 Determination of Self-diffusion Coefficients of Structural Groups

The diffusion coefficient for each peak in the NMR spectrum (thereby for each assigned structural group) was determined from the results of the ^{13}C PFG NMR experiments (using a stimulated echo sequence with bipolar pulsed gradients), using Eq. (2), which is a modified version of the Stejskal-Tanner equation[30, 53]:

$$\ln\left(\frac{I_p}{I_{0,p}}\right) = -\sum_{n=1}^{2} c_n \left(D_p \gamma^2 \delta^2 \left(\Delta - \frac{\delta}{3} - \frac{\tau}{2}\right) G^2\right)^n \tag{2}$$

where I_p is the measured peak height, $I_{0,p}$ is the peak height in the absence of diffusion, γ is the gyromagnetic ratio of the observed nucleus, δ is the duration of the gradient pulse, Δ is the diffusion time, τ is the correction constant due to the usage of the bipolar gradients, G is the gradient strength, and D_p is the self-diffusion coefficient for peak p to be determined. In prior work of our group, it was empirically found that it is sufficient to consider the first two terms of the series in Eq. (2); the probe-specific fitting parameters c_1 and c_2, which account for weak non-linearities in the gradient profile of the used probe were adopted here from the prior work.[30] From the attenuation of the peak heights I_p with increasing G, a self-diffusion coefficient D_p for each peak p was obtained by a least square fit of Eq.(2) to the experimental data using MATLAB 2021 b[54], where $I_{0,p}$ was fitted simultaneously[55, 56].

The results are presented in so-called diffusion-ordered spectroscopy (DOSY) maps, in which the self-diffusion coefficients of the peaks are plotted over the chemical shift. Besides the self-diffusion coefficients D_p, also their uncertainties were retrieved, which was done by using the MATLAB function 'nlparci' assuming a *t*-distribution for the error of each self-diffusion coefficient.

2.3.2 Clustering Algorithm

The clustering of the identified structural groups to pseudo-components was done based on both the self-diffusion coefficient D_p for each peak p as measured by PFG NMR as well as the respective experimental uncertainty of D_p specified by the 95 % confidence intervals of a *t*-distribution, i.e., $(D_p - e_{p,95\%}, D_p + e_{p,95\%})$. As input for the clustering, only the vector $\boldsymbol{x}_p = (D_p - e_{p,95\%}, D_p + e_{p,95\%})$ for each peak p, was used, which is sufficient as it implicitly contains the information on D_p.

Formally, the goal of the clustering is to partition the set $\mathcal{X}^{\mathrm{mix}} = \{(\boldsymbol{x}_p)\}_{p=1}^{P}$, whereby P is the total number of peaks in the ^{13}C NMR spectrum (and the associated structural groups) of a mixture into *K* clusters. In the present work, we have used K-medians clustering, which is a variant of the K-means algorithm but which is more robust towards outliers[57, 58]. Each cluster thereby represents a distinct pseudo-component. For a given number of clusters K, the K-medians algorithm seeks to minimize the L_1 distance, i.e., the sum of the absolute distances in the individual coordinates of all (input) data points to their assigned cluster centers; the center of each cluster is thereby calculated as the component-wise median of all assigned data points[54]. Intuitively, K-medians assign those structural groups to the same pseudo-component that show similar diffusion behavior both with regard to the value of the diffusion coefficient as well as of the respective uncertainty.

The number of clusters K, i.e., the number of pseudo-components to be distinguished in a mixture, is a priori unknown. It was chosen here based on the overall silhouette score[41] $\overline{s}(K)$, which is a common metric for automatically selecting the most suitable number of clusters for a given clustering problem. Intuitively, $\overline{s}(K)$ measures in our case how consistent the definition of the K pseudo-components is, i.e., how similar the diffusion behavior of the structural groups inside each cluster (pseudo-component) is in average over all pseudo-components for the chosen number K. The value of $\overline{s}(K)$ generally lies between -1 and 1, where higher values indicate more consistent solutions. The clustering was

performed here with different values of K, and the number of clusters K with the highest $\overline{s}(K)$ was adopted. A detailed description of the K-medians algorithm and the silhouette score is given in the ESI.

Based on the clustering, the relative number of the structural groups in the pseudo-components can be determined in analogy to Eq.(1); this was again done here based on group mole fractions $x_{g,k}$, now for each pseudo-component k.

2.4 Prediction of Molar Masses

For obtaining the absolute number of structural groups in a pseudo-component, additional information on the molar mass of the pseudo-component is required, which was obtained here also based on the self-diffusion coefficients measured by PFG NMR spectroscopy. The self-diffusion coefficient of each pseudo-component was thereby calculated by taking the arithmetic mean of the self-diffusion coefficients of all peaks (structural groups) in the respective cluster.

For taking into account that available models relating self-diffusion coefficients to molar masses are restricted to infinitely diluted diffusing species and that, in general, the pseudo-components are not present at infinite dilution in a mixture of interest, the concept of *relative* diffusion coefficients D_{rel}[42, 43, 44, 45, 46] was applied. The diffusion coefficient of a pseudo-component $D_{\tilde{\text{U}}}$ was thereby related to the diffusion coefficient of a known reference component D_{ref} in the same sample:

$$D_{\text{rel}} = \frac{D_{\tilde{\text{U}}}}{D_{\text{ref}}} \tag{3}$$

In the literature[45, 46, 59], it was shown that relative diffusion coefficients are only a weak function of temperature and composition if the concentrations of the diffusing species are not too high. For an in-depth discussion, we refer to the ESI, where we verify this observation based on our own experiments with aqueous solutions up to mass fractions of the diffusing species of 0.28 g/g. Hence, the following assumption was used:

$$\frac{D^{\infty}_{\tilde{\text{U}}}}{D^{\infty}_{\text{ref}}} = \frac{D_{\tilde{\text{U}}}}{D_{\text{ref}}} = \text{const.} \tag{4}$$

From Eq. (4), the number of $D^{\infty}_{\tilde{\text{U}}}$ can be calculated from the experimental data for $D_{\tilde{\text{U}}}$ and D_{ref} from the PFG NMR experiments if D^{∞}_{ref} is known. We recommend selecting the reference component in a way that D^{∞}_{ref} can be adopted from the literature. As an alternative, it can also be determined experimentally, which is, however, usually tedious[60], or estimated using a prediction method[61, 62, 63, 20, 21].

From $D^{\infty}_{\tilde{\text{U}}}$, in turn, the molar mass $M_{\tilde{\text{U}}}$ of component $\tilde{\text{U}}$ can be calculated using basically any predictive model for self-diffusion coefficients at infinite dilution. We have used the SEGWE model[20, 21] in the present work, which is a semi-empirical extension of the Stokes-Einstein equation[64] and was found to be the best available semi-empirical model for predicting self-diffusion coefficients in a recent study[60]:

$$D^{\infty}_{\tilde{\text{U}}} = \frac{k_{\text{B}} T \left(\frac{3\alpha}{2} + \frac{1}{1+\alpha} \right)}{6\pi\eta_{\text{S}} \sqrt[3]{\frac{3M_{\tilde{\text{U}}}}{4\pi\rho_{\text{eff}} N_{\text{A}}}}} \tag{5a}$$

$$\alpha = \sqrt[3]{\frac{M_{\text{S}}}{M_{\tilde{\text{U}}}}} \tag{5b}$$

where $D^{\infty}_{\tilde{\text{U}}}$ is the self-diffusion coefficient of pseudo-component $\tilde{\text{U}}$ at infinite dilution, $M_{\tilde{\text{U}}}$ is the molar mass of $\tilde{\text{U}}$, k_{B} is the Boltzmann constant, η_{S} and M_{S} are the dynamic viscosity and molar mass of the solvent, respectively, T is the temperature, and ρ_{eff} is a lumped parameter of the SEGWE model, called effective density, whose default value[21] $\rho_{\text{eff}} = 627$ kg m^{-3} was used here. Calculating the molar mass $M_{\tilde{\text{U}}}$ from Eq.(5) requires solving a cubic equation and choosing the appropriate solution[20, 21].

From the molar mass, the absolute number of structural groups in each pseudo-component can be calculated, which will, in general, not result in integer values. Specifically, integer values are only realistic, if a defined pseudo-component represents only a single true component of the mixture; still, non-integer values can thereby result from experimental uncertainties and model errors, cf. ESI for a brief discussion. In the other case, namely, if two or more true components are lumped into a single pseudo-component, in general non-integer values can be expected. This is, fortunately, usually not a problem in the application of group-contribution methods. From the absolute number of structural groups in each pseudo-component, the mole fractions of all pseudo-components can be predicted.

3 Overview of Applications

The applicability of the proposed method for NMR fingerprinting and the definition and quantification of pseudo-components is demonstrated in the following by applying it to three dilute aqueous test mixtures of different complexity, cf. Table 2. The true composition of the mixtures was known from the sample preparation in all cases, but, at no point, any information about the concentration of any component was used.

All NMR experiments were carried out at 298.15 K. The temperature is, however, only needed for the application of the SEGWE model. Furthermore, the following information on the dynamic viscosity of the solvent water was used: $\eta_{\mathrm{W}} = 890.02 \cdot 10^{-6}$ Pa · s as reported for $T = 298.15$ K in Ref.[65]. The fact that only rather dilute mixtures were investigated facilitates the estimation of the molar mass with the SEGWE model, cf. above, but it poses challenges for the NMR spectroscopy due to sensitivity issues. The chosen examples are therefore neither particularly favorable nor unfavorable. More examples will be studied in future work. Here, the focus is on demonstrating the principal feasibility and on generating first application examples – not on comprehensiveness.

Table 2: Overview of the test mixtures considered in this work. All mixtures additionally contain the solvent water.

Mixture	Components i	M_i / g mol^{-1}	x_i / mol mol^{-1}
I	2-propanol	60.10	0.033
	acetone	58.08	0.038
	1,4-butanediol	90.12	0.035
	acetic acid	60.05	0.033
II	1,4-dioxane	88.11	0.006
	cyclohexanone	98.15	0.012
	citric acid	192.12	0.025
	glucose[a]	180.16	0.015
III	acetonitrile	41.05	0.022
	acetone	58.08	0.016
	acetic acid	60.05	0.015
	1-propanol	60.10	0.015
	2-propanol	60.10	0.015
	cyclohexanone	98.15	0.009
	1,4-butanediol	90.12	0.010
	malic acid	134.09	0.008
	xylose[a]	150.13	0.007
	ascorbic acid	176.12	0.006

[a]Glucose and xylose are present in different anomeric forms in aqueous solution, which were not differentiated here.

4 Results and Discussion

4.1 Prediction of Structural Groups and Clustering to Pseudo-Components

In the following, the results of the application of the proposed method to the three test mixtures, cf. Table 2, are shown. For better clarity, we distinguish here between the group-specific characterization (NMR fingerprinting) together with the clustering step in Section 4.1, which yields relative amounts for the structural groups in each pseudo-component, and the definition of pseudo-components that also involves predicting the respective molar masses in Section 4.2, whereby absolute numbers for the structural groups in the pseudo-components are obtained. For the sake of completeness, the results of the group-specific NMR fingerprinting alone, without the clustering step, are included in the ESI.

4.1.1 Mixture I

Figures 2 and 3 show the results of the qualitative definition of pseudo-components for test mixture I. In the left part of Figure 2, the respective DOSY map is shown together with the result of a clustering into four pseudo-components. The number of clusters that are distinguished here was automatically selected by the algorithm based on the overall silhouette score $\overline{s}(K)$, which is shown in the right part of Figure 2 for different numbers of clusters K. The highest

overall silhouette score $\overline{s}(K)$, corresponding to the most consistent definition of pseudo-components according to this metric, was found for $K = 4$, which is labeled by the red symbol in Figure 2 (right). In this case, the algorithm found the true number of components in the mixture and correctly assigned the signals (and the respective groups) to the different components, as indicated in the legend of Figure 2. Note that water, which was assumed as known solvent and shows no signal in ^{13}C NMR spectroscopy, is not explicitly considered here and in the following.

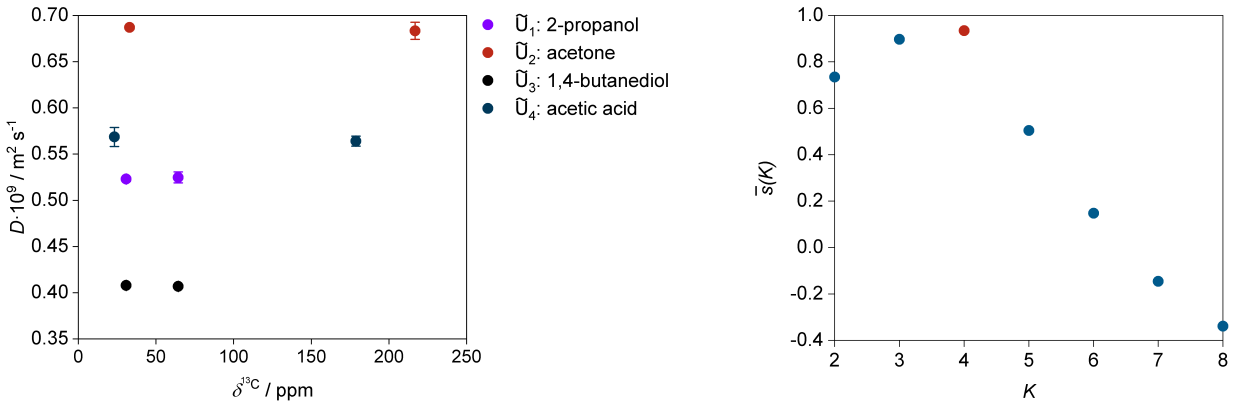

Figure 2: Left: DOSY map of mixture I showing the clustering of the structural groups into four pseudo-components $\tilde{\mathrm{U}}$ ($K = 4$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate the 95 % confidence intervals based on a *t*-distribution. Right: overall silhouette score $\overline{s}(K)$ for the clustering with the K-medians algorithm for different numbers of clusters K. The largest $\overline{s}(K)$, which was found for $K = 4$, is marked red.

Figure 3 shows the relative composition of the four pseudo-components $\tilde{\mathrm{U}}_1$-$\tilde{\mathrm{U}}_4$ in terms of group mole fractions defined by the algorithm for mixture I (bottom row) and compares them to the group mole fractions of the true components (top row). Note that the information on the true components is only used for comparison, but was not used for obtaining the predictions. The results show a very good agreement between the predicted compositions of the pseudo-components and those of the respective true components; the small deviations can be attributed to experimental errors in the NMR analysis.

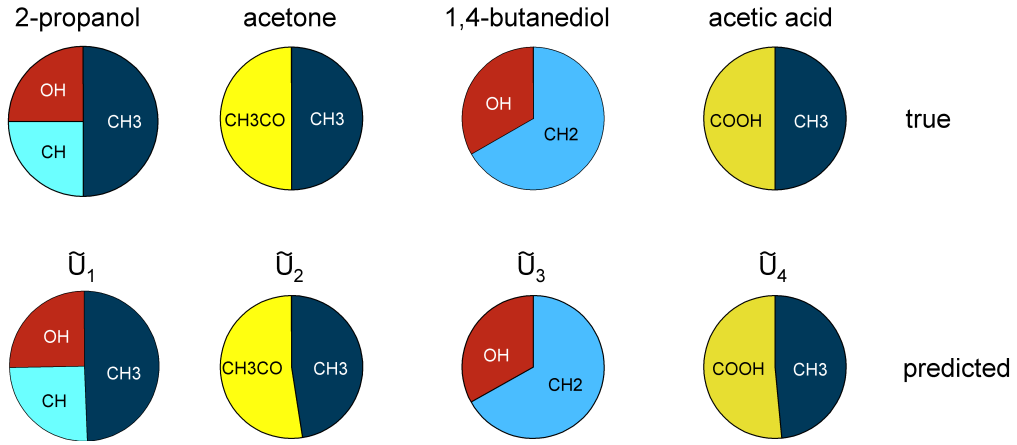

Figure 3: Relative composition of the true components (top row) and the predicted pseudo-components (bottom row) in mixture I in terms of group mole fractions $x_{g,k}$.

4.1.2 Mixture II

In Figures 4 and 5, the respective results for mixture II are shown. The algorithm distinguishes only three pseudo-components while there are four true components in this case; citric acid and glucose are lumped into one pseudo-component due to their similar self-diffusion coefficients, cf. Figure 4 left. In Figure S.7 in the ESI, we demonstrate that the clustering algorithm correctly assigns all peaks (structural groups) to their respective true components, including

the correct distinction between citric acid and glucose, if $K = 4$ is defined a priori, i.e., if four clusters are chosen. This demonstrates that the method can, in principle, be supported by prior knowledge, e.g., on the number of different components in a mixture here, whenever such information is available.

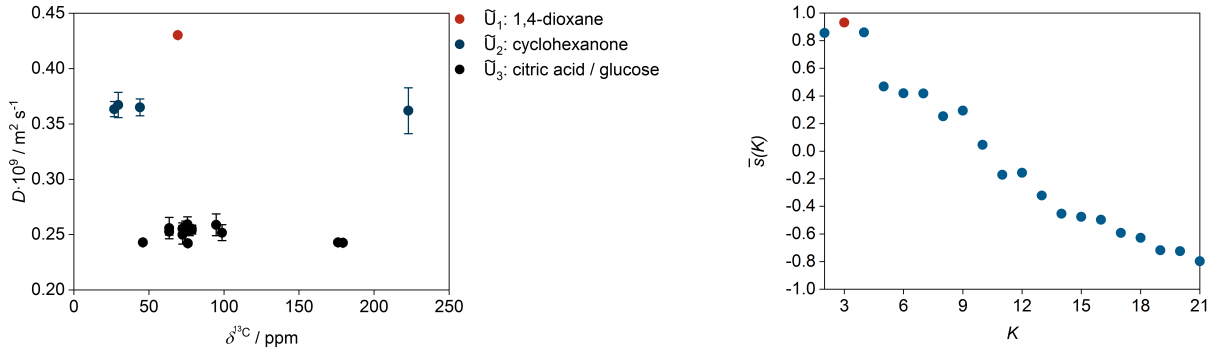

Figure 4: Left: DOSY map of mixture II showing the clustering of the structural groups into three pseudo-components $\tilde{U}$ ($K = 3$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate the 95 % confidence intervals based on a *t*-distribution. Right: overall silhouette score $\overline{s}(K)$ for the clustering with the K-medians algorithm for different numbers of clusters K. The largest $\overline{s}(K)$, which was found for $K = 3$, is marked red.

Figure 5 shows the comparison of the predicted and the true group mole fractions in mixture II. For the component 1,4-dioxane ($\tilde{U}_1$), 'OH' groups are incorrectly identified in contrast to cyclic ether groups, which we call cy-CH2O here and which are simply not included in the list of groups considered here, cf. Table 1. While such incorrect predictions of structural groups can naturally influence also the thermodynamic modeling of the mixture based on its predicted composition, the influence is small in many cases, as we have demonstrated in previous work.[24, 25] This can be attributed to the fact that our method identifies the structural groups in a physical way, namely based on information on the chemical shift of the respective peaks in NMR spectra, as well as on the substitution degree of the carbon nuclei. Since similar chemical shifts indicate a similarity of the structural groups, the falsely predicted structural groups will in many cases be similar to the true ones. For instance, we can expect a polar group to be falsely interpreted as another polar group, as it is also the case with the example studied here.

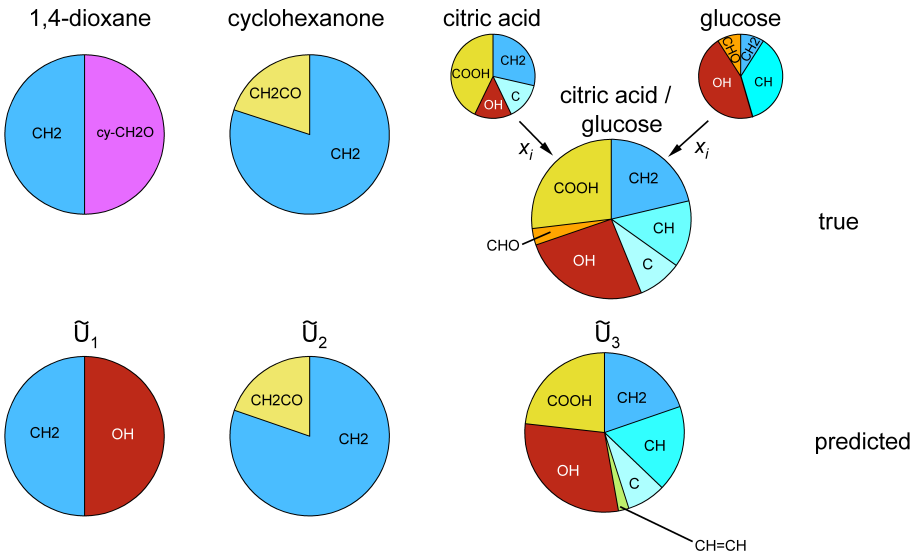

Figure 5: Relative composition of the true components (top row) and the predicted pseudo-components (bottom row) in mixture II in terms of group mole fractions $x_{g,k}$. Arrows indicate that a pseudo-component represents a 'mixture' of multiple true components.

The relative composition of pseudo-component $\tilde{U}_2$ shows an excellent agreement with that of the true component cyclohexanone, whereas the composition of pseudo-component $\tilde{U}_3$ is a combination of that of glucose and citric acid (weighted by their mole fractions in mixture II). For this pseudo-component, most structural groups are identified correctly; only a small number of 'CH=CH' groups are incorrectly assigned to pseudo-component $\tilde{U}_3$, which results from the peaks of glucose in the ^{13}C NMR spectrum appearing at > 90 ppm. Overall, the small deviations can mainly be attributed to shortcomings of the peak assignment using a simple chemical shift table, which could be refined in future work, e.g., by using ML approaches[27].

4.1.3 Mixture III

Figures 6 and 7 show the predictions for mixture III. In this case, the maximum overall silhouette score $\overline{s}(K)$ was found for $K = 8$ clusters, cf. Figure 6 right, while the true number of components in this mixture is ten. The algorithm fails to distinguish 1-propanol and 2-propanol, which exhibit very similar self-diffusion coefficients, as well as the strongly polar components malic acid and xylose, cf. Figure 6 left. Given the complexity of the ^{13}C NMR spectrum, cf. Figure S.3 in the ESI, and the DOSY map for this mixture, cf. Figure 6 left, the performance of the method is remarkable. Furthermore, again, if the true number of different components (ten) is set as prior information, the algorithm perfectly distinguishes all components and correctly assigns the structural groups to them, as we demonstrate in Figure S.8 in the ESI.

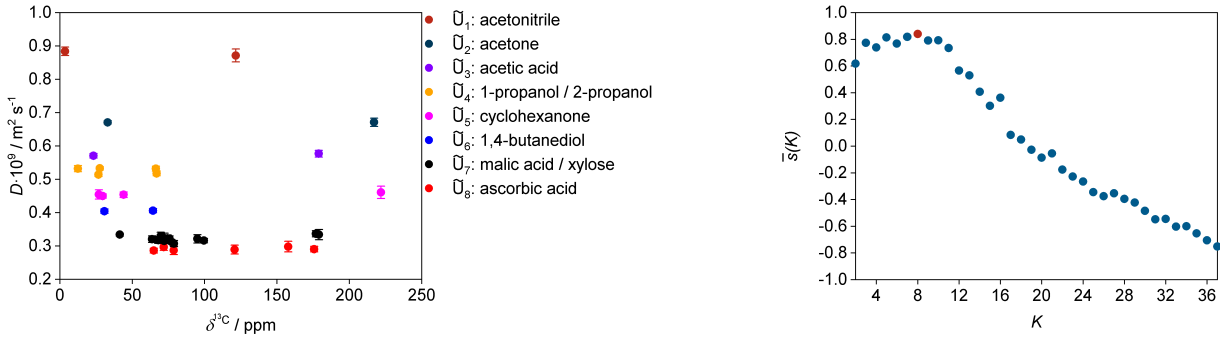

Figure 6: Left: DOSY map of mixture III showing the clustering of the structural groups into eight pseudo-components $\tilde{U}$ ($K = 8$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate the 95 % confidence intervals based on a *t*-distribution. Right: overall silhouette score $\overline{s}(K)$ for the clustering with the K-medians algorithm for different numbers of clusters K. The largest $\overline{s}(K)$, which was found for $K = 8$, is marked red.

In Figure 7, the group mole fractions of the true components (first and third row) of mixture III and those of the pseudo-components (second and fourth row) are compared. In most cases, the pseudo-components show an excellent agreement with the respective true components with respect to the group mole fraction (acetone, acetic acid, cyclohexanone, 1,4-butanediole) or represent a 'mixture' of the respective true components ($\tilde{U}_4$, $\tilde{U}_7$). Similar to the results observed for mixture II, cf. Figure 5, a small number of 'CH=CH' groups is incorrectly predicted for $\tilde{U}_7$. Furthermore, in pseudo-component $\tilde{U}_8$, the 'COO' group of ascorbic acid is misinterpreted as a 'COOH' group, which is simply due to the fact that there is no ester group in our list of groups, cf. Table 1. Acetonitrile ($\tilde{U}_1$) is made up of a single group in UNIFAC ('CH3CN') that is not included in our list and therefore misinterpreted as a combination of a 'C=C' and a 'CH3' group.

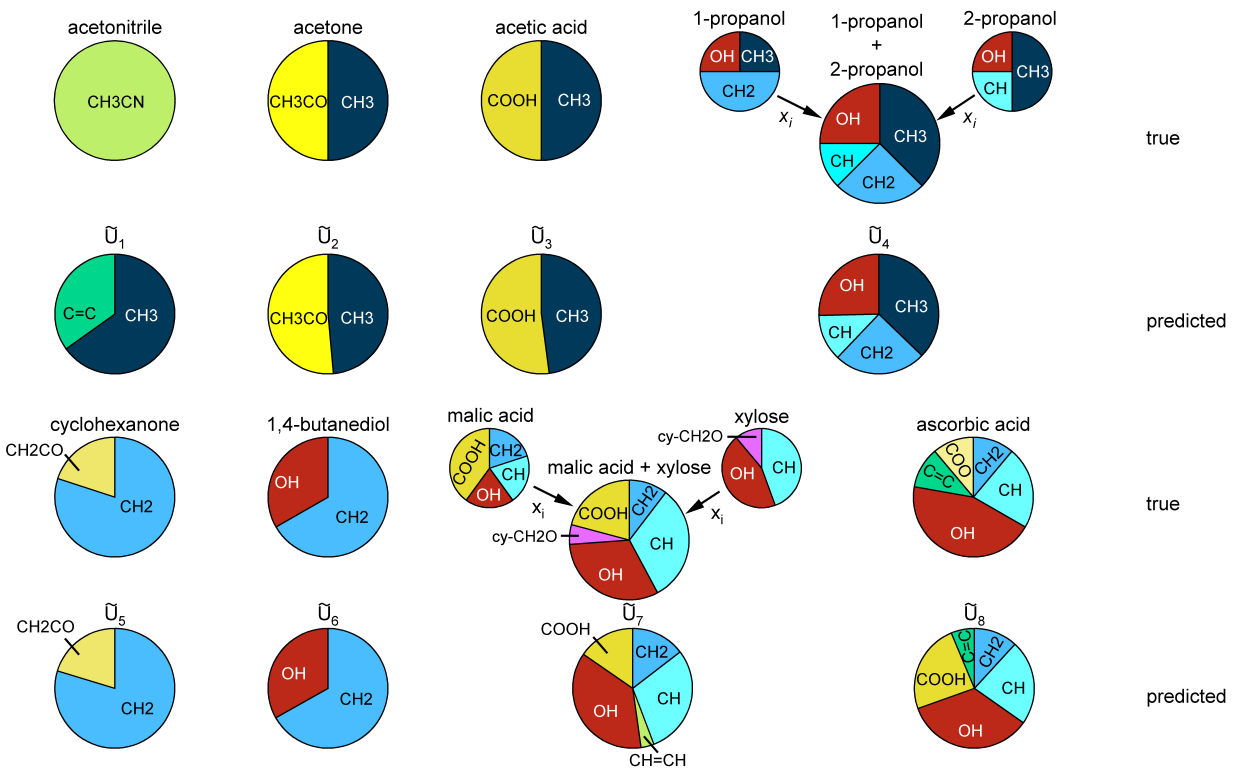

Figure 7: Relative composition of the true components (first and third row) and the predicted pseudo-components (second and fourth row) in mixture III in terms of group mole fractions $x_{g,k}$. Arrows indicate that a pseudo-component represents a 'mixture' of multiple true components.

4.2 Prediction of Molar Masses of Pseudo-components

The previous step yields the *relative* amounts of the structural groups in each pseudo-component in the form of group mole fractions, cf. Figures 3, 5, and 7. For determining the *absolute* number of structural groups in each pseudo-component, information on the molar mass of each pseudo-component is required. In our method, we propose to predict the molar mass based on the measured self-diffusion coefficients of each pseudo-component, and have used the SEGWE model[20, 21] for this purpose here. For applying the SEGWE model, the solvent has to be known, which was water in all studied mixtures here, an extrapolation of the measured self-diffusion coefficient to the state of infinite dilution has to be carried out, and a reference component with known self-diffusion coefficient at infinite dilution in the pure solvent is required in each mixture, cf. Section 2.4 for details. For convenience, we have simply designated one of the components of each mixture as reference component but note that also any other component could be added to the mixture. Note that no a priori information on the concentration of the reference component is required.

Figure 8 (left) shows the results for the prediction of the molar mass of the pseudo-components $\tilde{\mathrm{U}}$ of mixture I and compares them to the molar mass of each respective true component. As reference component, 2-propanol was chosen and its diffusion coefficient at infinite dilution in water was taken from Ref.[66] ($D^{\infty}_{\mathrm{ref}} = 0.99 \cdot 10^{-9}$ m^2 s^{-1} at $T = 298.15$ K). Fair agreement is obtained for acetone and acetic acid (and the respective pseudo-components), whereas the deviation is larger for 1,4-butanediol. Figure 8 (right) shows a comparison of the predicted composition of mixture I in terms of mole fraction in the water-free part of the mixture x^*_k and the respective true composition of the mixture. Overall, good predictions for the mole fractions are obtained. The deviations result mainly from poor estimates of the molar mass and are, hence, presumably related to shortcomings of the SEGWE model.

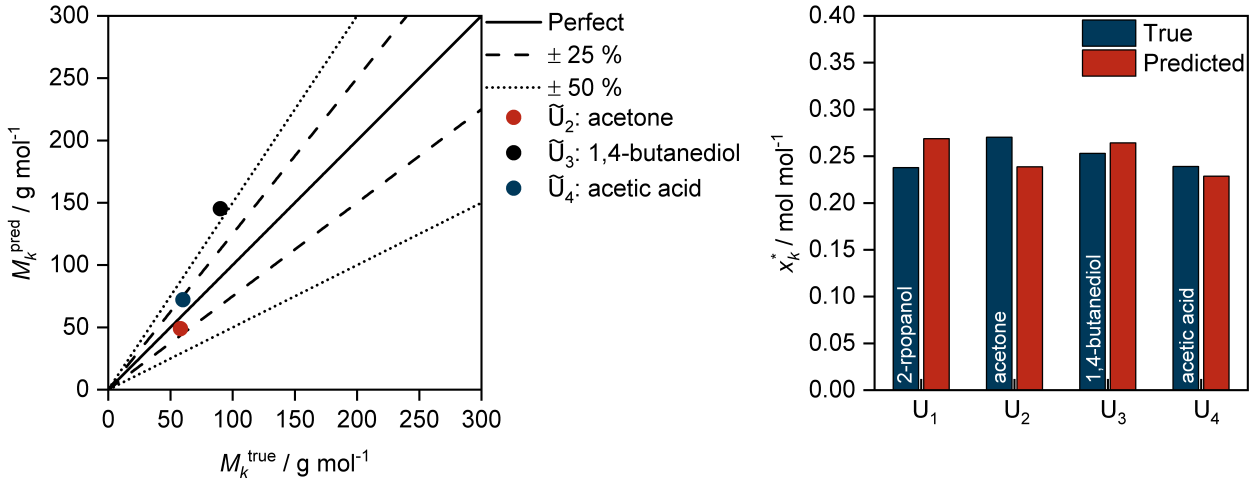

Figure 8: Left: prediction of the molar mass of the pseudo-components in mixture I based on the measured self-diffusion coefficients using the SEGWE model[20, 21]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture I and comparison to the true composition.

In Figure 9 (left), the prediction of the molar masses of the pseudo-components identified in mixture II are compared to the respective values for the true components. As reference component, 1,4-dioxane was chosen and its diffusion coefficient at infinite dilution in water was taken from Ref.[67] ($D_{\mathrm{ref}}^{\infty} = 1.110 \cdot 10^{-9}$ m^2 s^{-1} at $T = 298.15$ K). Overall, good predictions were obtained. Since citric acid and glucose were lumped into a single pseudo-component, $\tilde{\mathrm{U}}_3$, the predicted molar mass of $\tilde{\mathrm{U}}_3$ is depicted over the values for two true components (citric acid and glucose). In Figure 9 (right), the water-free composition x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture II is shown. Good predictions were obtained. Note that for the results of pseudo-component $\tilde{\mathrm{U}}_3$ the sum of the true components that are part of it (glucose and citric acid) are depicted.

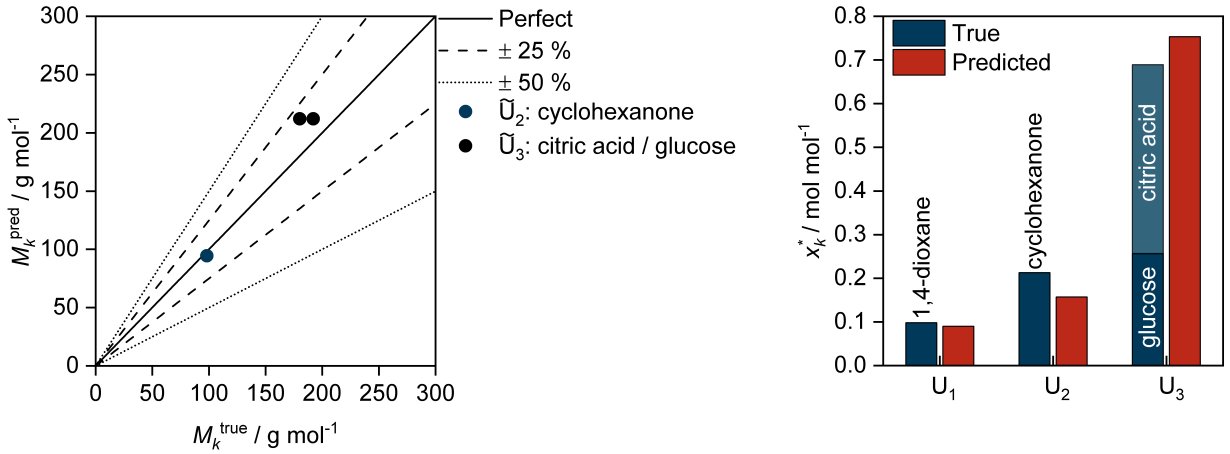

Figure 9: Left: prediction of the molar mass of the pseudo-components in mixture II based on the measured self-diffusion coefficients using the SEGWE model[20, 21]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture II and comparison to the true composition.

Figure 10 (left) shows the prediction of the molar masses of the pseudo-components identified in mixture III and compares them to the respective values for the true components. Acetonitrile was chosen as a reference component, thereby a diffusion coefficient at infinite dilution in water was taken from Ref.[68] ($D_{\mathrm{ref}}^{\infty} = 1.649 \cdot 10^{-9}$ m^2 s^{-1} at $T = 298.15$ K). In general, higher accuracies can be observed for the smaller components. By contrast, for large components, such as ascorbic acid and xylose, the molar mass was overpredicted. This presumably results from deficiencies of the SEGWE model in representing diffusion coefficients of large components with many polar groups. It is also interesting to note that when xylose is chosen as the reference component, the molar mass of the other highly

polar components (1,4-butanediol, malic acid, and ascorbic acid) is predicted with higher accuracy, but the prediction of the molar mass of the smaller and less polar components deteriorates, cf. Figure S.9 in the ESI. In future work, refined diffusion models that explicitly consider specific interactions, particularly between highly polar components, could be used instead of the SEGWE model. Significant improvements can thereby be expected as the SEGWE model does not use any information on the diffusing species except for its molar mass. The development of more sophisticated diffusion models would therefore be very valuable, in particular since information on the composition of the pseudo-components is automatically and reliably retrievable, as demonstrated above.

Figure 10 (right) shows a comparison of the water-free composition x_k^* of the pseudo-components and the true components in the mixture. Overall, a good estimate for the composition is obtained.

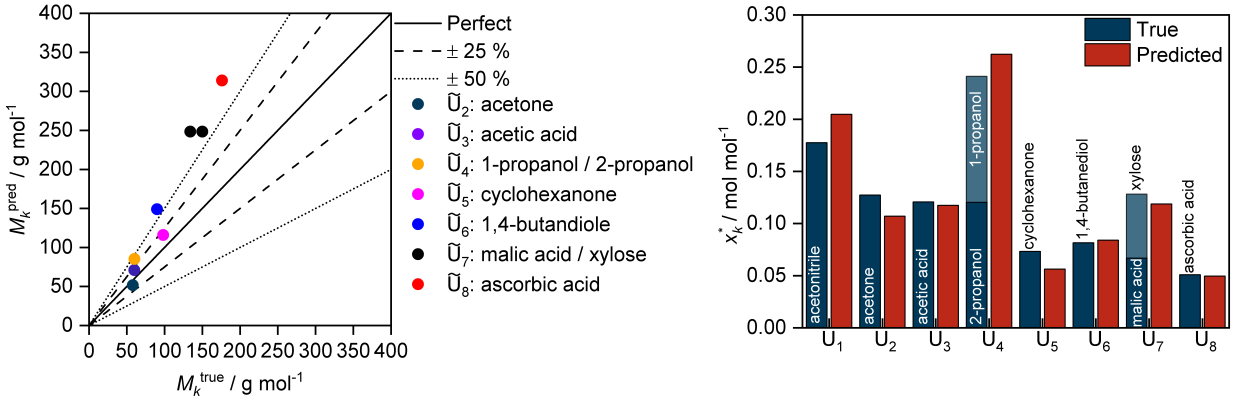

Figure 10: Left: prediction of the molar mass of the pseudo-components in mixture III based on the measured self-diffusion coefficients using the SEGWE model[20, 21]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture III and comparison to the true composition.

5 Conclusions

Poorly specified mixtures are common in chemical engineering. A common approach to model such mixtures is to introduce pseudo-components. In this work, a new method for the representation of poorly specified mixtures by pseudo-components is introduced. Standard nuclear magnetic resonance (NMR) experiments are carried out, yielding the basis for the *NMR fingerprinting*. We use this term for labeling the identification and quantification of structural groups in a mixture. Subsequently, the structural groups are clustered into pseudo-components based on measured self-diffusion coefficients. By using information about self-diffusion coefficients, the molar mass of the pseudo-components can also be estimated.

We have demonstrated the applicability of the method using three aqueous mixtures of different complexity as examples but an extension to non-aqueous mixtures is straightforward. The pseudo-components identified by the method were either identical with a true component or several true components were lumped into a pseudo-component of similar size and group composition as the true components. Good estimates for the composition were obtained. In cases where true components were lumped, the concentration of the pseudo-component was found to be close to the sum of the concentrations of the respective true components.

The method can be combined with thermodynamic group-contribution methods in a straightforward manner and thereby enables thermodynamic modeling of poorly specified mixtures without requiring cumbersome component elucidations. The method, therefore, paves the way for convenient process design and optimization with poorly specified mixtures.

The main source of error for the structural group composition of the pseudo-components is the uncertainty of the predicted molar mass, which emphasizes the importance of developing more accurate diffusion coefficient models. Deviations from integer values for the numbers of structural groups in a pseudo-component can indicate that a defined pseudo-component is not a true component, which could, ultimately, be used for refining the method. This requires, however, a quantitative error analysis. The topic was beyond the scope of the present work, but is worth being considered in a follow-up study. In future work, the NMR fingerprinting could be further improved by integrating knowledge from additional NMR experiments. In principle also results from other group-specific analytical methods could be incorporated. By combining suitable NMR techniques, the most common chemical groups can be identified and quantified. The list of groups that are actually identified depends on the chosen techniques and can be adapted to the

task at hand. In the present work, only groups containing C, H, and O were considered as examples. The list of groups can be adapted to that used in the thermodynamic group-contribution method, as long as a mapping is possible. The assignment of the NMR signals to groups was done here using simple chemical shift tables. This can be refined by using classification methods from machine learning such as support vector classification[27]. Furthermore, in the present work, the NMR fingerprinting was carried out based on results from cryogenic high-field NMR experiments. As an alternative, also NMR experiments with benchtop spectrometers should be considered, which are simpler and easier to handle and would facilitate the application of the new method. To compensate for the reduced sensitivity, new pulsed-field gradient (PFG) NMR methods, e.g., based on polarization enhancement nurtured during attached nucleus testing (PENDANT)[47] could be developed and applied, which overcome the disadvantage of suppressing the signals of quaternary carbons.

Conflicts of interest

There are no conflicts to declare.

Acknowledgments

The authors gratefully acknowledge financial support from Deutsche Forschungsgemeinschaft (DFG) under Grant No. JI 401/1-1.

References

[1] David T. Allen, Murray R. Gray, and Thuy T. Le. Structural characterization and thermodynamic property estimation for wood tars: a functional group approach. *Liquid Fuels Technology*, 2(3):327–353, 1984.

[2] Gregory L. Alexander, Barry J. Schwarz, and John M. Prausnitz. Phase equilibriums for high-boiling fossil fuel distillates. 2. correlation of equation-of-state constants with characterization data for phase equilibrium calculations. *Industrial & Engineering Chemistry Fundamentals*, 24(3):311–315, 1985.

[3] Bernardo Carreón-Calderón, Verónica Uribe-Vargas, Edgar Ramírez-Jaramillo, and Mario Ramírez-de Santiago. Thermodynamic characterization of undefined petroleum fractions using group contribution methods. *Industrial & Engineering Chemistry Research*, 51(43):14188–14198, 2012.

[4] Christian A Jackson and William J Simonsick. Application of mass spectrometry to the characterization of polymers. *Current Opinion in Solid State and Materials Science*, 2(6):661–667, 1997.

[5] Muhammad Malik, Jimmy Mays, and Muhammad Raza Shah, editors. *Molecular Characterization of Polymers*. Elsevier, 2021.

[6] Maria C Cuellar and Adrie JJ Straathof. Downstream of the bioreactor: Advancements in recovering fuels and commodity chemicals. *Current Opinion in Biotechnology*, 62:189–195, 2020.

[7] William J. Sim and Thomas E. Daubert. Prediction of vapor-liquid equilibria of undefined mixtures. *Industrial & Engineering Chemistry Process Design and Development*, 19(3):386–393, 1980.

[8] Gregory L. Alexander, A. Louise Creagh, and John M. Prausnitz. Phase equilibriums for high-boiling fossil fuel distillates. 1. characterization. *Industrial & Engineering Chemistry Fundamentals*, 24(3):301–310, 1985.

[9] Claude F. Leibovici. A consistent procedure for the estimation of properties associated to lumped systems. *Fluid Phase Equilibria*, 87(2):189–197, 1993.

[10] Mohamed A. Fahim and Amal S. Elkilani. Prediction of solubility of hydrogen in petroleum cuts using modified unifac. *The Canadian Journal of Chemical Engineering*, 70(2):335–340, 1992.

[11] Verônica J. Pereira, Victor B. Regueira, Gloria M. N. Costa, and Silvio A. B. Vieira de Melo. Modeling the saturation pressure of systems containing crude oils and co_2 using the srk equation of state. *Journal of Chemical & Engineering Data*, 64(5):2134–2142, 2019.

[12] Zhao Gao, Zhiming Xu, Suoqi Zhao, and Linzhou Zhang. Heavy petroleum supercritical fluid deasphalting process simulation based on the saturate, aromatic, resin, and asphaltene composition. *Energy & Fuels*, 36(16):8818–8827, 2022.

[13] Margit T. Rätzsch and Horst Kehlen. Continuous thermodynamics of polymer solutions: The effect of polydispersity on the liquid-liquid equilibrium. *Journal of Macromolecular Science: Part A - Chemistry*, 22(3):323–334, 1985.

[14] Margit T. Rätzsch. Continuous thermodynamics. *Pure and Applied Chemistry*, 61(6):1105–1114, 1989.

[15] Bernardo Carreón-Calderón, Verónica Uribe-Vargas, Mario Ramírez-de Santiago, and Edgar Ramírez-Jaramillo. Thermodynamic characterization of heavy petroleum fluids using group contribution methods. *Industrial & Engineering Chemistry Research*, 53(13):5598–5607, 2014.

[16] Mehrak Mahmudi and Mohammad Taghi Sadeghi. A novel three pseudo-component approach (ThPCA) for thermodynamic description of hydrocarbon-water systems. *Journal of Petroleum Exploration and Production Technology*, 4(3):281–289, 2013.

[17] Abdul Gani Abdul Jameel, Ayman M. Elbaz, Abdul-Hamid Emwas, William L. Roberts, and S. Mani Sarathy. Calculation of average molecular parameters, functional groups, and a surrogate molecule for heavy fuel oils using ^{1}h and ^{13}c nuclear magnetic resonance spectroscopy. *Energy & Fuels*, 30(5):3894–3905, 2016.

[18] Angelo Raimondi, Antonio Favela-Contreras, Francisco Beltrán-Carbajal, Alejandro Piñón-Rubio, and Jose Luis de la Peña-Elizondo. Design of an adaptive predictive control strategy for crude oil atmospheric distillation process. *Control Engineering Practice*, 34:39–48, 2015.

[19] Aage Fredenslund, Russell L. Jones, and John M. Prausnitz. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. *AIChE Journal*, 21(6):1086–1099, 1975.

[20] Robert Evans, Zhaoxia Deng, Alexandria K. Rogerson, Andy S. McLachlan, Jeff J. Richards, Mathias Nilsson, and Gareth A. Morris. Quantitative interpretation of diffusion-ordered nmr spectra: Can we rationalize small molecule diffusion coefficients? *Angewandte Chemie International Edition*, 52(11):3199–3202, 2013.

[21] Robert Evans, Guilherme Dal Poggetto, Mathias Nilsson, and Gareth A. Morris. Improving the interpretation of small molecule diffusion coefficients. *Analytical Chemistry*, 90(6):3987–3994, 2018.

[22] K. Peter C. Vollhardt and Neil Eric Schore. *Organic Chemistry: Structure and Function*. Macmillan Learning: New York, 8 edition, 2018.

[23] Fabian Jirasek, Jakob Burger, and Hans Hasse. Method for estimating activity coefficients of target components in poorly specified mixtures. *Industrial & Engineering Chemistry Research*, 57(21):7310–7313, 2018.

[24] Fabian Jirasek, Jakob Burger, and Hans Hasse. NEAT—NMR spectroscopy for the estimation of activity coefficients of target components in poorly specified mixtures. *Industrial & Engineering Chemistry Research*, 58(21):9155–9165, 2019.

[25] Fabian Jirasek, Jakob Burger, and Hans Hasse. Application of neat for the simulation of liquid–liquid extraction processes with poorly specified feeds. *AIChE Journal*, 66(2):e16826, 2020.

[26] Thomas Specht, Kerstin Münnemann, Fabian Jirasek, and Hans Hasse. Estimating activity coefficients of target components in poorly specified mixtures with NMR spectroscopy and COSMO-RS. *Fluid Phase Equilibria*, 516:112604, 2020.

[27] Thomas Specht, Kerstin Münnemann, Hans Hasse, and Fabian Jirasek. Automated methods for identification and quantification of structural groups from nuclear magnetic resonance spectra using support vector classification. *Journal of Chemical Information and Modeling*, 61(1):143–155, 2021. PMID: 33405926.

[28] C. D'Agostino, M.D. Mantle, L.F. Gladden, and G.D. Moggridge. Prediction of mutual diffusion coefficients in non-ideal mixtures from pulsed field gradient NMR data: Triethylamine–water near its consolute point. *Chemical Engineering Science*, 74:105–113, 2012.

[29] C. D'Agostino, J.A. Stephens, J.D. Parkinson, M.D. Mantle, L.F. Gladden, and G.D. Moggridge. Prediction of the mutual diffusivity in acetone–chloroform liquid mixtures from the tracer diffusion coefficients. *Chemical Engineering Science*, 95:43–47, 2013.

[30] Daniel Bellaire, Hendrik Kiepfer, Kerstin Münnemann, and Hans Hasse. 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*, 65(2):793–803, 2020.

[31] Daniel Bellaire, Oliver Großmann, Kerstin Münnemann, and Hans Hasse. Diffusion coefficients at infinite dilution of carbon dioxide and methane in water, ethanol, cyclohexane, toluene, methanol, and acetone: A PFG-NMR and MD simulation study. *The Journal of Chemical Thermodynamics*, 166:106691, 2022.

[32] Kevin F. Morris and Charles S. Johnson. Diffusion-ordered two-dimensional nuclear magnetic resonance spectroscopy. *Journal of the American Chemical Society*, 114(8):3139–3141, 1992.

[33] Kevin F. Morris, Peter. Stilbs, and Charles S. Johnson. Analysis of mixtures based on molecular size and hydrophobicity by means of diffusion-ordered 2d nmr. *Analytical Chemistry*, 66(2):211–215, 1994.

[34] G.S Kapur, M Findeisen, and S Berger. Analysis of hydrocarbon mixtures by diffusion-ordered nmr spectroscopy. *Fuel*, 79(11):1347–1351, 2000.

[35] Deyu Li, Russell Hopson, Weibin Li, Jia Liu, and Paul G. Williard. ^{13}c inept diffusion-ordered nmr spectroscopy (dosy) with internal references. *Organic Letters*, 10(5):909–911, 2008. PMID: 18251549.

[36] Stéphane Balayssac, Saleh Trefi, Véronique Gilard, Myriam Malet-Martino, Robert Martino, and Marc-André Delsuc. 2d and 3d dosy ^{1}h nmr, a useful tool for analysis of complex mixtures: Application to herbal drugs or dietary supplements for erectile dysfunction. *Journal of Pharmaceutical and Biomedical Analysis*, 50(4):602–612, 2009.

[37] G. Pagès, V. Gilard, R. Martino, and M. Malet-Martino. Pulsed-field gradient nuclear magnetic resonance measurements (PFG NMR) for diffusion ordered spectroscopy (DOSY) mapping. *The Analyst - The Analytical Journal of the Royal Society of Chemistry*, 142(20):3771–3796, 2017.

[38] Donghui Wu, Aidi Chen, and Charles S. Johnson, Jr. Heteronuclear-detected diffusion-ordered nmr spectroscopy through coherence transfer. *Journal of Magnetic Resonance, Series A*, 123(2):215–218, 1996.

[39] Adolfo Botana, Peter W.A. Howe, Valérie Caër, Gareth A. Morris, and Mathias Nilsson. High resolution ^{13}c dosy: The deptse experiment. *Journal of Magnetic Resonance*, 211(1):25–29, 2011.

[40] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, first edition, 2006.

[41] Peter J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[42] Jonathan A. Jones, Deborah K. Wilkins, Lorna J. Smith, and Christopher M. Dobson. Characterisation of protein unfolding by nmr diffusion measurements. *Journal of Biomolecular NMR*, 10:199–203, 1997.

[43] Shenggen Yao, Geoffrey J. Howlett, and Raymond S. Norton. Peptide self-association in aqueous trifluoroethanol monitored by pulsed field gradient nmr diffusion measurements. *Journal of Biomolecular NMR*, 16:109–119, 2000.

[44] Eurico J. Cabrita and Stefan Berger. Dosy studies of hydrogen bond association: tetramethylsilane as a reference compound for diffusion studies. *Magnetic Resonance in Chemistry*, 39(S1):S142–S148, 2001.

[45] Christopher A. Crutchfield and Douglas J. Harris. Molecular mass estimation by pfg nmr spectroscopy. *Journal of Magnetic Resonance*, 185(1):179–182, 2007.

[46] Emmanuelle Durand, Martin Clemancey, Jean-Marc Lancelin, Jan Verstraete, Didier Espinat, and Anne-Agathe Quoineaud. Aggregation states of asphaltenes: Evidence of two chemical behaviors by ^{1}h diffusion-ordered spectroscopy nuclear magnetic resonance. *The Journal of Physical Chemistry C*, 113(36):16266–16276, 2009.

[47] John Homer and Michael C. Perry. Enhancement of the NMR spectra of insensitive nuclei using PENDANT with long-range coupling constants. *Journal of the Chemical Society, Perkin Transactions 2*, (3):533, 1995.

[48] George C Levy and Joseph D Cargioli. Spin-lattice relaxation in solutions containing cr(III) paramagnetic relaxation agents. *Journal of Magnetic Resonance (1969)*, 10(2):231–234, 1973.

[49] Zhe Zhou, Yiyong He, Xiaohua Qiu, David Redwine, Janece Potter, Rongjuan Cong, and Matthew Miller. Optimum cr(acac)$_3$ concentration for NMR quantitative analysis of polyolefins. *Macromolecular Symposia*, 330(1):115–122, 2013.

[50] Albert A. Smith. INFOS: spectrum fitting software for NMR analysis. *Journal of Biomolecular NMR*, 67(2):77–94, 2017.

[51] Stanislav Sokolenko, Tangi Jézéquel, Ghina Hajjar, Jonathan Farjon, Serge Akoka, and Patrick Giraudeau. Robust 1d NMR lineshape fitting using real and imaginary data in the frequency domain. *Journal of Magnetic Resonance*, 298:91–100, 2019.

[52] Yevgen Matviychuk, Ellen Steimers, Erik von Harbou, and Daniel J. Holland. Improving the accuracy of model-based quantitative nuclear magnetic resonance. *Magnetic Resonance*, 1(2):141–153, 2020.

[53] Mark A. Connell, Paul J. Bowyer, P. Adam Bone, Adrian L. Davis, Alistair G. Swanson, Mathias Nilsson, and Gareth A. Morris. Improving the accuracy of pulsed field gradient nmr diffusion experiments: correction for gradient non-uniformity. *Journal of Magnetic Resonance*, 198(1):121–131, 2009.

[54] MATLAB. *version 9.11.0 (R2021b)*. The MathWorks Inc., Natick, Massachusetts, 2021.

[55] Gareth A. Morris. *Diffusion-Ordered Spectroscopy*. John Wiley & Sons, Ltd, 2009.

[56] Timothy D.W. Claridge, editor. *High-Resolution NMR Techniques in Organic Chemistry*. Elsevier, third edition, 2016.

[57] C. Whelan, G. Harrell, and J. Wang. Understanding the *K*-medians problem. In *Proceedings of the International Conference on Scientific Computing*, 2015.

[58] P. S. Bradley, O. L. Mangasarian, and W. N. Street. Clustering via concave minimization. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, NIPS'96, page 368–374, Cambridge, MA, USA, 1996. MIT Press.

[59] Roman Neufeld and Dietmar Stalke. Accurate molecular weight determination of small molecules via DOSY-NMR by using external calibration curves with normalized diffusion coefficients. *Chemical Science*, 6(6):3354–3364, 2015.

[60] Oliver Großmann, Daniel Bellaire, Nicolas Hayer, Fabian Jirasek, and Hans Hasse. Database for liquid phase diffusion coefficients at infinite dilution at 298 k and matrix completion methods for their prediction. *Digital Discovery*, 1(6):886–897, 2022.

[61] C. R. Wilke and Pin Chang. Correlation of diffusion coefficients in dilute solutions. *AIChE Journal*, 1(2):264–270, 1955.

[62] K. A. Reddy and L. K. Doraiswamy. Estimating liquid diffusivity. *Industrial & Engineering Chemistry Fundamentals*, 6(1):77–79, 1967.

[63] Myo T. Tyn and Waclaw F. Calus. Diffusion coefficients in dilute binary liquid mixtures. *Journal of Chemical & Engineering Data*, 20(1):106–109, 1975.

[64] A. Einstein. Über die von der molekularkinetischen theorie der waerme geforderte bewegung von in ruhenden fluessigkeiten suspendierten teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

[65] Germany VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen, Düsseldorf, editor. *VDI-Wärmeatlas*. Springer Berlin Heidelberg, 11 edition, 2013.

[66] K. C. Pratt, W. A. Wakeham, and Alfred Rene Jean Paul Ubbelohde. The mutual diffusion coefficient for binary mixtures of water and the isomers of propanol. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 342(1630):401–419, 1975.

[67] Derek G. Leaist, Kimberley MacEwan, Alexandra Stefan, and Muhannad Zamari. Binary mutual diffusion coefficients of aqueous cyclic ethers at 25 °c. tetrahydrofuran, 1,3-dioxolane, 1,4-dioxane, 1,3-dioxane, tetrahydropyran, and trioxane. *Journal of Chemical & Engineering Data*, 45(5):815–818, 2000.

[68] A. J. Easteal, L. A. Woolf, and R. Mills. Velocity cross-correlation coefficients for the system acetonitrile-water at 278 k and 298 k. *Zeitschrift für Physikalische Chemie*, 155(1-2):69–78, 1987.