Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

ESI for Rational Method for Defining and Quantifying Pseudo-components Based on NMR Spectroscopy

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

Laboratory of Engineering Thermodynamics (LTD), RPTU Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663 Kaiserslautern, Germany

E-mail: *fabian.jirasek@rptu.de

Experimental Methods

Chemicals

Deionized and purified water, which was used as solvent for all test mixtures studied in this work, was produced with a purification system of Merck Millipore (Elix Essential 5). In Table S.1, information on the other chemicals used for the preparation of these mixtures is summarized.

Table S.1: Suppliers and purities of the chemicals used in this work. Purities are indicated as specified by the suppliers.

Chemical	Formula	Supplier	Purity
acetone	C_3H_6O	Fisher Scientific	≥99.80%
acetic acid	$C_2H_4O_2$	Carl Roth	≥99.80%
acetonitrile	C_2H_3N	Fisher Scientific	≥99.90%
ascorbic acid	$C_6H_8O_6$	Carl Roth	≥99.00%
1,4-butanediol	$C_4H_{10}O_2$	Sigma Aldrich	≥99.00%
citric acid	$C_6H_8O_7$	Carl Roth	≥99.50%
cyclohexanone	$C_6H_{10}O$	Sigma Aldrich	≥99.80%
1,4-dioxane	$C_4H_8O_2$	Sigma Aldrich	≥99.80%
glucose	$C_6H_{12}O_6$	Carl Roth	≥99.50%
malic acid	$C_4H_6O_5$	Sigma Aldrich	≥99.00%
1-propanol	C_3H_8O	Honeywell	≥99.50%
2-propanol	C_3H_8O	Merck	≥99.90%
TMSP-d4	$NaC_6H_9D_4O_2Si$	Sigma Aldrich	≥98.00%
xylose	$C_5H_{10}O_5$	Alfa Aesar	≥98.00%

NMR Analysis

Sample Preparation and NMR Spectroscopy

Samples of test mixtures (>20 g) were prepared gravimetrically in glass vessels using a balance of Mettler Toledo with an accuracy of ±0.001 g. Approximately 1 ml of each sample was transferred to a 5 mm NMR tube. All NMR experiments were carried out at 25°C with a 400 MHz Avance NMR spectrometer from Bruker with a Double Resonance Broad Band CryoProbe. The temperature control of the spectrometer was calibrated against a platinum resistance thermometer. The absolute uncertainty of the temperature is estimated to be lower than 1 K for the NMR experiments.

Quantitative inverse gated 1D ^{13}C NMR spectra, with a flip angle of 90°, a relaxation delay of 185-200 s, 64-128 scans, a maximum acquisition time of 15.33 s, and a maximum bandwidth of 250 ppm were recorded. Inverse gated ^{13}C distortionless enhancement by polarization transfer (DEPT) 90/135 NMR spectra were recorded with a relaxation delay of

60-200 s, 4-32 scans, a maximum acquisition time of 15.33 s, and a maximum bandwidth of 250 ppm. An additional quantitative inverse gated 1D ^{13}C NMR spectrum with the same number of scans was recorded. A one-bond proton-carbon coupling constant $^{1}J_{CH}$ of 145 Hz that determines the specific delay in the DEPT experiment was chosen (for further details, see, e.g., Ref.[1]). All chemical shifts are referenced to the shift of sodium 3-(trimethylsilyl)tetradeuteriopropionate (TMSP-d4) by recording an additional ^{13}C NMR spectrum with a small amount of TMSP-d4 after all other NMR experiments were carried out. Automatic baseline and phase correction was applied with MestReNova before the manual peak integration was done. In most cases, the relative error compared to the true composition was smaller than 5%.

PFG NMR Spectroscopy

Self-diffusion coefficients in the studied mixtures were measured at 25°C with the same instrument that was used for the acquisition of the 1D NMR spectra as described above. For recording the ^{13}C pulsed-field gradient (PFG) NMR spectra, a stimulated echo sequence with bipolar pulsed gradients similar to the one used in recent work of our group[2,3] was applied. In contrast to our prior work,[2,3] the decoupler was additionally turned on for a maximum of 7 s prior to the stimulated echo sequence here to obtain an enhancement of the ^{13}C peaks based on the nuclear overhauser effect (NOE), that does not sacrifice the peaks of quaternary carbons.[4] For each mixture, seven PFG measurements with varying gradient strength G ranging from 2.55 to 48.46 G cm^{-1} (in equal steps of G^2) were performed; the diffusion of the components thereby causes an attenuation of the peaks, from which the self-diffusion coefficient can be calculated,[5] cf. Eq.(2) in the manuscript. The diffusion time Δ was chosen as 50 ms for all measurements, and τ was 218.4 µs. The gradient pulse duration δ was adjusted to the respective sample and was between 5.4 and 7.0 ms. A relaxation delay of 100-158 s, 40-120 scans, a maximum acquisition time of 18.42 s, and a maximum bandwidth of 250 ppm was chosen. Automatic baseline and phase correction, peak alignment,

and exponential line broadening of 1 Hz were applied with MestReNova. The peak heights needed in Eq.(2) in the manuscript were also evaluated by MestReNova.

Distinction between Substitution Degrees with DEPT NMR

By using different pulse angles, the DEPT experiments enable the differentiation of basically all substitution degrees of carbon nuclei, i.e., primary, secondary, tertiary, and quaternary ones, because they show, depending on the combination of pulse angle and substitution degree, either positive or negative enhancements of their peaks, or are (almost) completely suppressed from the spectrum.[1]

The distinction between primary, secondary, tertiary, and quaternary carbons was made as follows: quaternary carbons could easily be identified as they, in theory, do not show any peaks in conventional DEPT NMR spectra but only in the quantitative ^{13}C NMR spectrum. However, since, in practice, a small residual peak of quaternary carbons is usually detected also in DEPT spectra, we used a quantitative ^{13}C NMR spectrum with the same number of scans as the respective DEPT spectra as a reference for deciding whether a peak is 'present' or 'absent' in the DEPT spectra. In all cases here, the area of the residual peak of a quaternary carbon in the DEPT spectra was negligible compared to the area of the respective peak in the quantitative ^{13}C NMR spectrum.

Subsequently, the other types of carbons could also be distinguished in a straightforward manner: secondary carbons are the only type that shows negative peaks in DEPT 135 spectra. Primary and tertiary carbons can then be distinguished based on DEPT 90 spectra, where primary carbons should, in theory, show no peak; however, due to the appearance of residual peaks in practice, we considered the ratio of the areas of the respective peaks in the DEPT 90 and the DEPT 135 spectrum, which was well below unity in all cases of a primary carbon.

We used the peaks of the designated reference component to phase the DEPT 135 NMR spectrum. This was done for the sake of simplicity but, in practice, DEPT can also be used for the classification of peaks without prior knowledge of any component. In that situation, the phase correction could be carried out based on a peak of any reference component that is added to the mixture prior to the NMR analysis.

PFG NMR Spectra and Assignment of Peaks

Figures S.1-S.3 show ^{13}C PFG NMR spectra of mixtures I-III for a gradient strength G = 2:55 G cm^{-1}. Based on these spectra, it was decided which peaks were to be distinguished. We note that, especially for completely unknown mixtures, this procedure can be ambiguous, e.g., due to small distortions that can lead to a "splitting" of a peak. We, therefore, first carried out an exponential line broadening of 1 Hz, which is a standard processing step of NMR spectra. The great majority of peaks in the ^{13}C PFG NMR spectra recorded in this work did not show an overlapping with other peaks. Additionally, we used the peak heights to determine the self-diffusion coefficients in all cases to mitigate the effects of overlapping peaks on the evaluation of the diffusion coefficient.

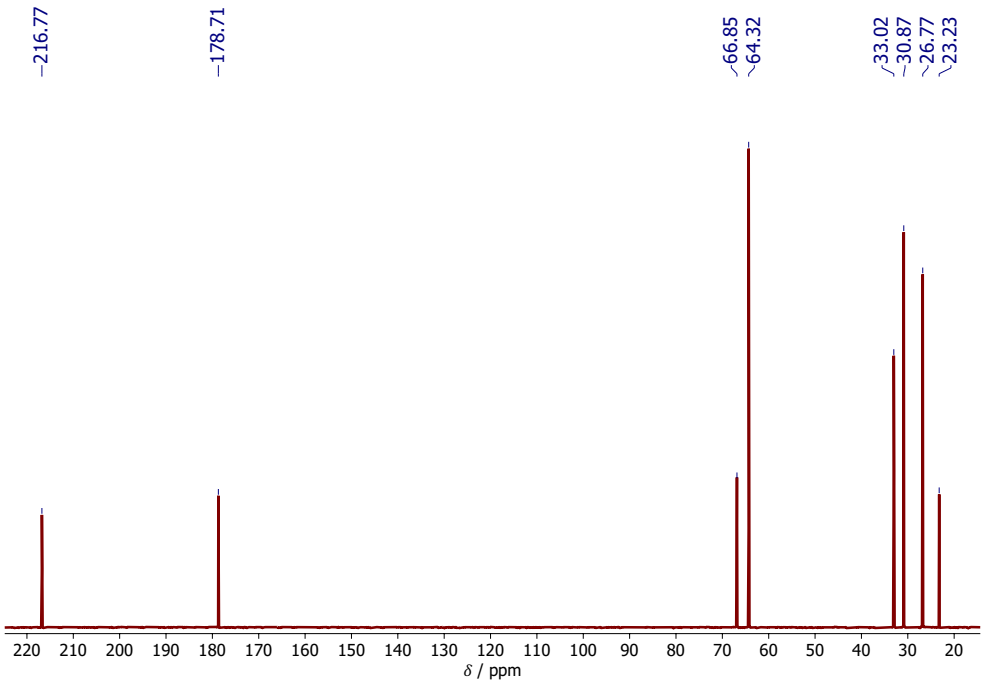

Figure S.1: ^{13}C PFG NMR spectrum of mixture I with gradient strength $G = 2{:}55$ G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

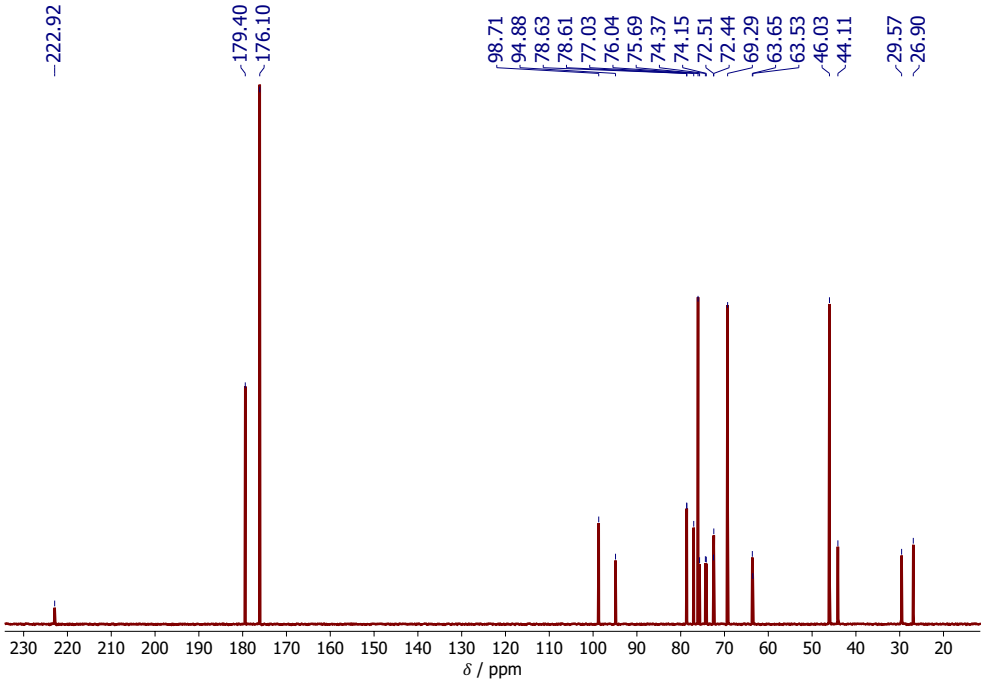

Figure S.2: ^{13}C PFG NMR spectrum of mixture II with gradient strength $G = 2{:}55$ G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

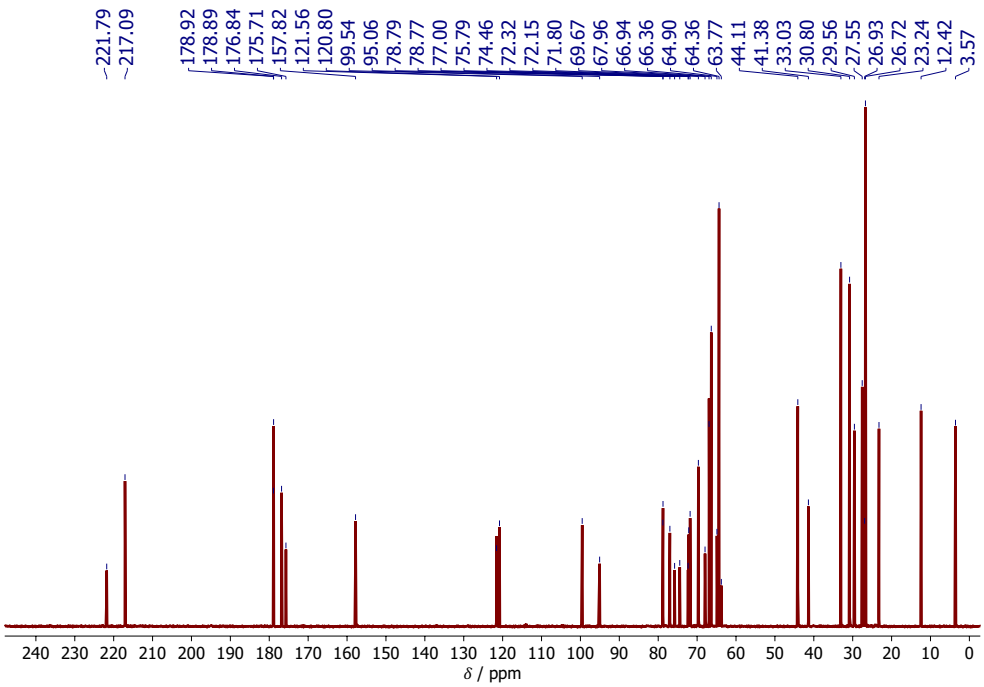

Figure S.3: ^{13}C PFG NMR spectrum of mixture III with gradient strength $G = 2{:}55$ G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

K-medians Algorithm and Silhouette Score

In the present work, we propose to use K-medians clustering, which is a variant of the K-means algorithm that is more robust towards outliers.[6,7] In the K-medians algorithm, the center of each cluster is calculated by the *median* of all data points associated with this cluster, and the following objective function is minimized for a specified number of clusters, i.e., pseudo-components, K:

$$J = \sum_{p=1}^{P} \sum_{k=1}^{K} r_{p;k} \left\| \boldsymbol{x}_p - \boldsymbol{c}_k \right\|_1 \tag{S.1}$$

where P is the total number of peaks in the ^{13}C NMR spectrum of the studied mixture. $\boldsymbol{x}_p$ contains the input data for peak p as described in the manuscript, and $\boldsymbol{c}_k$ represents the center of the kth cluster. $r_{p;k}$ is a binary indicator that captures to which cluster k peak p

is assigned: if peak p is assigned to cluster k, then $r_{p;k} = 1$, otherwise $r_{p;k} = 0$. $\|\boldsymbol{x}_p - \boldsymbol{c}_k\|_1$ denotes the L_1 distance, i.e., the sum of the absolute distances in the individual coordinates (also called 'manhatten' or 'cityblock' distance), between $\boldsymbol{x}_p$ and $\boldsymbol{c}_k$.

K-medians clustering was performed using the 'kmeans' function in MATLAB 2021b[8] and setting the distance metric to 'cityblock' to use the L_1 distance. The algorithm thereby uses a component-wise median to determine the cluster centers, i.e., the median is calculated independently in each dimension. Since the algorithm is a local optimization algorithm, 1000 replicates were used, and only the solution with the lowest J, cf. Eq.(S.1), was kept[6,9] for each specified number of clusters K.

Since the number of clusters K, i.e., the number of pseudo-components that are to be distinguished in the studied mixture, is a priori unknown in most cases in practice, it also needs to be set by the algorithm. For this purpose, we used the so-called silhouette score s, which is a common metric for automatically selecting the most suitable number of clusters for a given clustering problem.[10] The silhouette score s was thereby first calculated individually for each data point $\boldsymbol{x}_p$ as follows using the MATLAB function 'silhouette' with the L_1 distance metric:

$$s(\boldsymbol{x}_p) = \frac{b(\boldsymbol{x}_p) - a(\boldsymbol{x}_p)}{\max\{b(\boldsymbol{x}_p); a(\boldsymbol{x}_p)\}} \tag{S.2}$$

where $a(\boldsymbol{x}_p)$ is the average distance of $\boldsymbol{x}_p$ from all other data points *in the same cluster* (to which $\boldsymbol{x}_p$ is assigned), and $b(\boldsymbol{x}_p)$ is the smallest average distance of $\boldsymbol{x}_p$ to all points in a different cluster; again, the L_1 distance was thereby used as distance metric. The definition of $a(\boldsymbol{x}_p)$ and $b(\boldsymbol{x}_p)$ was slightly adapted for the special case of a cluster that contains only a single data point as discussed and explained in the following section.

The silhouette score can, by definition, have values between -1 and 1, where -1 indicates that the data point $\boldsymbol{x}_p$ is 'totally dissimilar' to the other points in the same cluster, whereas a silhouette score of 1 indicates that the data point fits perfectly into the assigned cluster. By averaging the obtained silhouette scores of all data points associated with a cluster (via the arithmetic mean), a mean silhouette score for each cluster was obtained. Subsequently,

the mean silhouette scores of the clusters were again averaged (via the arithmetic mean) to obtain an overall silhouette score $\bar{s}(K)$, which only depends on the assumed total number of clusters K, i.e., the number of pseudo-components considered here. This two-step averaging process was chosen to ensure that clusters with different numbers of assigned data points are weighted equally for the calculation of the final silhouette score $\bar{s}(K)$.

For selecting the appropriate number of clusters, K-medians clustering was performed with values of K ranging from 2 to P, i.e., up to the total number of peaks in the ^{13}C NMR spectrum of the mixture, and in each case, the overall silhouette score $\bar{s}(K)$ was calculated; then, the number of clusters K with the highest $\bar{s}(K)$ was adopted.

Calculation of Silhouette Coefficients for Single Data Points

For calculating the individual silhouette scores s for each data point, the MATLAB function 'silhouette' was used, which we, however, had to adapt as described in the following. The reason for this is that in the special case of a cluster that contains *only a single data point*, which is denoted as $\boldsymbol{x}_p^*$ in the following, the silhouette score $s(\boldsymbol{x}_p^*)$ is not well defined since there are no distances $a(\boldsymbol{x}_p^*)$ within the cluster that could be calculated here. While this case might not be relevant in many other situations, in particular, if the number of data points N greatly exceeds the expected number of clusters K ($N >> K$), it needs to be considered for the application considered here: there are, in fact, components that show only a single peak in an NMR spectrum, e.g., 1,4-dioxane or benzene in proton-decoupled ^{13}C NMR spectroscopy, to name only two of many examples.

The default setting in MATLAB for the calculation of the silhouette score $s(\boldsymbol{x}_p^*)$, in this case, is to set $s(\boldsymbol{x}_p^*) = 1$, i.e., to assume a perfect assignment. This, in turn, leads to a model that favors solutions with an unreasonably high number of clusters (in our case: pseudo-components). To circumvent this issue, we used the experimental uncertainty $e_{p;95\%}$ of $\boldsymbol{x}_p^*$ for $a(\boldsymbol{x}_p^*)$, i.e., the intra-cluster distance, if the respective cluster contains $\boldsymbol{x}_p^*$ as the only data point. We furthermore defined $b(\boldsymbol{x}_p^*)$ as the minimal L_1 distance to any other data point in

this case. The intuition behind this is as follows: a data point with a small error, i.e., small $a(x_p)$, but with a large distance to all other data points, i.e., large $b(x_p)$, is likely to represent a (pseudo-)component that shows only a single peak in the NMR spectrum; hence, defining a cluster consisting of the respective data point only should, in this case, result in a high silhouette score $s(x_p)$. On the other hand, a data point with a rather large error bar, i.e., large $a(x_p)$, that is close to any other data point, i.e., low $b(x_p)$, is not so likely to represent a separate cluster, which should, in this case, result in a small or even negative silhouette score $s(x_p)$.

In Figure S.4 we demonstrate that the default behavior of the MATLAB function 'silhouette' would lead to the largest overall silhouette score $\bar{s}(K)$ if the assumed number of clusters K matches the total number of peaks P in the ^{13}C NMR spectrum. Hence, the clustering algorithm would always define the maximum possible number of pseudo-components, where all pseudo-components consist of only a single structural group and would show only a single peak in the ^{13}C NMR spectrum; such a result is, however, highly unrealistic. Figure S.4 demonstrates this using mixture I from the manuscript as an example, where the overall silhouette score $\bar{s}(K)$ continuously increases with increasing K.

Figure S.4: Overall silhouette scores $s(K)$ for the clustering of peaks in the ^{13}C NMR spectrum of mixture I with the K-medians algorithm for different numbers of clusters K as calculated with the default MATLAB setting.

Prediction of Molar Masses and Normalized Diffusion Coefficients

There are different methods for the prediction of molar masses from self-diffusion coefficients in the literature; Ref.[11] gives a good overview. We, therefore, briefly recapitulate only those concepts that are relevant for the development of our method in the following.

Good predictions can be obtained by internal calibration methods, where multiple reference components are added to the sample that contains the unknown component. Oftentimes a power-law is then fitted to the reference components in the sample, which is subsequently used for the prediction of the molar mass of the unknown component.[11,12] Of course, this requires that the reference components, among other things, are ideally inert and sufficiently soluble in the studied solvent,[12] and it requires the addition of reference components to the mixture of interest. Therefore, in Ref.[13] an external calibration method for the prediction of molar masses was developed, which requires only one known component in the mixture.

The authors thereby introduced the concept of 'normalized di usion coe cients':

$$\log(D_{x;norm}) = \log(D_{ref;x}) \quad \log(D_{ref}) + \log(D_x) \tag{S.3}$$

where $\log(D_{x;norm})$ is the normalized self-di usion coe cient of the unknown component (labeled 'x' here), D_{ref} and D_x are the measured self-di usion coe cients of the reference and unknown component in the sample, respectively, and $D_{ref,x}$ is the known value of the self-di usion coe cient of the reference component that was determined by measuring only the reference component in the same solvent. We note that $D_{ref,x}$ only has to be determined once for each reference component in a speci c solvent and then can be used for the determination of molar masses of unknown components.

Solvent-speci c power-laws (for di erent shapes of unknown components) are then tted to the normalized di usion coe cients of a large number of components. In consequence, Eq.(S.3) can be seen as a method to link the measured self-di usion coe cient of an unknown component (in the actual sample) to a hypothetical sample to which the power-law was tted, which then enables a good prediction of molar masses without requiring several reference components.

In the following, we show that the concept of normalized di usion coe cients is similar to what we use in Eq. (4) in the manuscript, where we assume that the ratio of the self-di usion coe cients of an unknown component to that of a reference component in a mixture is the same as their ratio at in nite dilution in the solvent. We also show in the following that both approaches are directly linked to the concept of relative di usion coe cients (cf. Eq. (3) in the manuscript).

Starting with Eq.(S.3), rearranging and applying logarithmic rules yields:

$$\log\left(\frac{D_x}{D_{ref}}\right) = \log\left(\frac{D_{x,norm}}{D_{ref,x}}\right) \tag{S.4}$$

Taking the exponent of Eq.(S.4) results in:

$$\frac{D_x}{D_{ref}} = \frac{D_{x,norm}}{D_{ref,\,x}} \tag{S.5}$$

In the following, we assume that the reference state is at in nite dilution and switch indices to our notation (x ! Ø):

$$\frac{D_Ø}{D_{ref}} = \frac{D^1_{Ø;norm}}{D^1_{ref,\,x}} \tag{S.6}$$

The resulting Eq.(S.6) is equivalent to the concept of relative di usion (Eq.(3) in the manuscript) at two di erent concentrations, or to Eq.(4) in the manuscript. In contrast to Ref.,[13] we did not t a solvent-speci c power-law; instead, we directly applied the SEGWE[14,15] model, which was developed for describing di usion coe cients at in nite dilution. Furthermore, the SEGWE model has been demonstrated to perform reasonably well using just one universal t parameter for di erent solvents.[14–16]

Concentration Dependence of Relative Di usion Coe cients

To verify the validity of Eq. (4) from the manuscript, i.e., that the ratio of the di usion coe cients of two components (a known reference component and a pseudo-component Ø here) is approximately constant for di erent compositions, two aqueous systems were studied here as examples. Table S.2 gives an overview of these systems and speci es the composition of two mixtures that were prepared for each system. In system A, 2-propanol was chosen as reference component, for which a value for the di usion coe cient at in nite dilution in water at 298.15 K of $D^1_{ref} = 0{:}99 \cdot 10^{-9}\,m^2s^{-1}$ was taken from Ref.[17] (as in the manuscript). In system B, acetone was chosen as reference component, for which $D^1_{ref} = 1{:}3 \cdot 10^{-9}\,m^2s^{-1}$ in water at 298.15 K was taken from Ref.[18] Figure S.5 shows the ratio $\frac{D_Ø}{D_{ref}}$ for the two systems measured by PFG NMR (cf. Section "PFG NMR Spectroscopy"). For all components, the

arithmetic mean of the self-diffusion coefficients of the respective peaks of the components were taken.

Table S.2: Overview of the studied aqueous mixtures for verifying Eq. (4) in the manuscript. All mixtures additionally contain the solvent water.

System	Component i	x_i / mol mol^{-1}
A	2-propanol	0.050
	malic acid	0.011
	2-propanol	0.010
	malic acid	0.050
B	acetone	0.010
	acetic acid	0.090
	acetone	0.090
	acetic acid	0.011

The ratio of the self-diffusion coefficients of the reference component and the pseudo-component stays nearly constant, irrespective of the different concentrations of the components in the studied mixtures.

Figure S.5: Measured dependence of the ratio $\frac{D_{\square}}{D_{ref}}$ on the mole fraction of the reference component, cf. Table S.2 and Eq. (4) in the manuscript.

Identification and Quantification of Structural Groups

Most of the considered structural groups, cf. Table 1 in the manuscript, contain only one carbon nucleus that shows a peak in the respective region of the ^{13}C NMR spectrum, which we denote by $z_g = 1$ for group g. There are two exceptions: first, the alkenyl groups ('CH=CH / C=C'), which contain two carbon nuclei that usually show peaks in the same region of the NMR spectrum, i.e., $z_g = 2$; and second, the (alkyl + ketone) groups ('CH3CO / CH2CO'), which also contain two carbon nuclei, but for which one can expect one peak in the region 0-60 ppm in the ^{13}C NMR spectrum (of the 'CH_3 / CH_2' part) and another peak in the region > 180 ppm (of the 'CO' part), and, hence, $z_g = 1$ for each of the two regions. As a consequence, the concentration of 'CH3/CH2' groups was calculated from the peak area in the assigned regions that exceeds the peak area in the region 160-180 ppm for each pseudo-component. Also note that if a 'CH3' group is detected in a pseudo-component, 'CH3CO' is chosen, otherwise 'CH2CO'.

Determination of Water-free Composition of Pseudo-components

From the clustering of structural groups to pseudo-components and the peak areas A_p, the ratio of structural groups in each pseudo-component, i.e., a group mole fraction $x_{g;k}$, can be calculated for every pseudo-component. In turn, together with the molar mass M_k of each pseudo-component k, as predicted by the SEGWE model based on the PFG NMR experiments, this enables the determination of the total number of groups n_k in each pseudo-component:

$$n_k = \frac{M_k}{\sum_{g=1}^{G} x_{g;k} M_g} \tag{S.7}$$

where M_g is the molar mass of group g, cf. Table S.3.

Table S.3: Molar mass M_g of all considered groups in this work, cf. Table 1 in the manuscript.

Group	M_g / g mol^{-1}
CH3	15.04
CH2	14.03
CH	13.02
C	12.01
OH	17.01
CH=CH [a]	26.04
C=C [a]	24.02
COOH	45.02
CHO	29.02
CH3CO [a]/CH2CO [a]	43.05/42.04

[a] To obtain the correct number of NMR-active nuclei z_k in the pseudo-component $z_g = 2$ has to be used for these groups since they contain two carbon atoms.

From this, the absolute number of each structural group g in pseudo-component k can be calculated:

$$n_{g;k} = x_{g;k} \, n_k \tag{S.8}$$

From this, in turn, the absolute number of NMR-active nuclei (here ^{13}C) z_k in each pseudo-

component can be calculated together with z_g, which is the number of NMR-active nuclei in structural group g:

$$z_k = \sum_{g=1}^{G} \nu_{g;k} z_g \tag{S.9}$$

The mole fraction x_k of each pseudo-component k in the water-free solution (which shows no signal in ^{13}C NMR), can then be determined using the quantitative results from the ^{13}C NMR spectrum:

$$x_k = \frac{\frac{\sum_{g=1}^{G} A_{g;k}}{z_k}}{\sum_{k=1}^{K} \frac{\sum_{g=1}^{G} A_{g;k}}{z_k}} \tag{S.10}$$

, where $A_{g;k}$ is the total area of all peaks associated to group g in pseudo-component k. Note that, with Eq. (S.10) also the mole fraction of the known reference component in the water-free solution is obtained, whereby $z_k = z_{ref}$ is also known.

Structural Group Composition

Composition of True Components

Table S.4 shows the composition of all components studied in this work regarding the groups of original UNIFAC. [19,20] Note that the UNIFAC nomenclature uses 'THF,'[19] as an abbreviation for cyclic ether groups. Since we found this misleading, we use 'cy-CH2O' instead.

Table S.4: Components considered in this work and their composition regarding groups from the UNIFAC table.[19,20] The numbers in parentheses are the identifiers for the sub-groups and the corresponding main-groups.

Component	UNIFAC groups
acetone	1 x 'CH3' (1,1) 1 x 'CH3CO' (18,9)
acetic acid	1 x 'CH3' (1,1) 1 x 'COOH' (42,20)
acetonitrile	1 x 'CH3CN' (40,19)
ascorbic acid	1 x 'CH2' (2,1) 2 x 'CH' (3,1) 4 x 'OH' (14,5) 1 x 'C=C' (70,2) 1 x 'COO' (77,41)
1,4-butanediol	4 x 'CH2' (2,1) 2 x 'OH' (14,5)
citric acid	2 x 'CH2' (2,1) 1 x 'C' (4,1) 1 x 'OH' (14,5) 3 x 'COOH' (42,20)
cyclohexanone	4 x 'CH2' (2,1) 1 x 'CH2CO' (19,9)
1,4-dioxane	2 x 'CH2' (2,1) 2 x 'cy-CH2O' (27,13)
glucose	1 x 'CH2' (2,1) 4 x 'CH' (3,1) 5 x 'OH' (14,5) 1 x 'CHO' (26,13)
malic acid	1 x 'CH2' (2,1) 1 x 'CH' (3,1) 1 x 'OH' (14,5) 2 x 'COOH' (42,20)
1-propanol	1 x 'CH3' (1,1) 2 x 'CH2' (2,1) 1 x 'OH' (14,5)
2-propanol	2 x 'CH3' (1,1) 1 x 'CH' (3,1) 1 x 'OH' (14,5)
water	1 x 'H2O' (16,7)
xylose	4 x 'CH' (3,1) 4 x 'OH' (14,5) 1 x 'cy-CH2O' (27,13)

Predicted Molar Masses and Composition of Pseudo-components

Tables S.5-S.7 show the predicted absolute numbers of the structural groups g in each pseudo-component k, denoted by $\nu_{g;k}$, in the three test mixtures, cf. Table 2 in the manuscript, as well as the predicted molar masses of the pseudo-components, M_k. Note that the stoichiometry and molar mass of the component that was considered as the known reference component here (Φ_1), which was needed for the determination of the stoichiometry of the other pseudo-components, is not included.

Table S.5: Absolute numbers $\nu_{g;k}$ of structural groups g according to UNIFAC[19,20] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of de ned pseudo-components for test mixture I, cf. Table 2 in the manuscript.

	Φ_2	Φ_3	Φ_4
M_k	48.98	145.21	72.11
$\nu_{CH3;k}$	0.782	-	1.147
$\nu_{CH2;k}$	-	6.464	-
$\nu_{CH;k}$	-	-	-
$\nu_{C;k}$	-	-	-
$\nu_{OH;k}$	-	3.206	-
$\nu_{CH=CH;k}$	-	-	-
$\nu_{C=C;k}$	-	-	-
$\nu_{COOH;k}$	-	-	1.219
$\nu_{CHO;k}$	-	-	-
$\nu_{CH3CO;k}$	0.865	-	-
$\nu_{CH2CO;k}$	-	-	-

Table S.6: Absolute numbers $\nu_{g;k}$ of structural groups g according to UNIFAC[19,20] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of de ned pseudo-components for test mixture II, cf. Table 2 in the manuscript.

	Θ_2	Θ_3
M_k	94.47	212.08
$\nu_{CH3;k}$	-	-
$\nu_{CH2;k}$	3.879	1.898
$\nu_{CH;k}$	-	1.680
$\nu_{C;k}$	-	0.759
$\nu_{OH;k}$	-	2.846
$\nu_{CH=CH;k}$	-	0.203
$\nu_{C=C;k}$	-	-
$\nu_{COOH;k}$	-	2.238
$\nu_{CHO;k}$	-	-
$\nu_{CH3CO;k}$	-	-
$\nu_{CH2CO;k}$	0.953	-

Table S.7: Absolute numbers $\nu_{g;k}$ of structural groups g according to UNIFAC[19,20] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of de ned pseudo-components for test mixture III, cf. Table 2 in the manuscript.

	Θ_2	Θ_3	Θ_4	Θ_5	Θ_6	Θ_7	Θ_8
M_k	51.71	71.06	85.26	115.96	148.90	248.36	313.85
$\nu_{CH3;k}$	0.853	1.111	2.111	-	-	-	-
$\nu_{CH2;k}$	-	-	1.409	4.688	6.625	1.812	1.617
$\nu_{CH;k}$	-	-	0.712	-	-	3.681	3.135
$\nu_{C;k}$	-	-	-	-	-	-	-
$\nu_{OH;k}$	-	-	1.440	-	3.291	4.577	4.752
$\nu_{CH=CH;k}$	-	-	-	-	-	0.435	-
$\nu_{C=C;k}$	-	-	-	-	-	-	0.858
$\nu_{COOH;k}$	-	1.208	-	-	-	1.907	3.308
$\nu_{CHO;k}$	-	-	-	-	-	-	-
$\nu_{CH3CO;k}$	0.903	-	-	-	-	-	-
$\nu_{CH2CO;k}$	-	-	-	1.194	-	-	-

Discussion of Uncertainties

The result of the proposed method, namely, the predicted composition of a poorly specified mixture with regard to pseudo-components, can be influenced by different sources of errors or uncertainties. These sources are:

(a) Incorrect identification of structural groups

While the correct identification of the structural groups in a poorly specified mixture is of course the basis for a meaningful definition of pseudo-components, the influence of errors here can, in many cases, be expected to have only a minor influence on the application of the results in combination with group-contribution methods. This is due to the fact that the identification here is physics-based, namely, based on information on the chemical shift of peaks in the NMR spectra and on the substitution degree of carbon nuclei. This procedure results in incorrectly predicted structural groups usually being identified as very similar structural groups, with only a small influence on the modeling results.

(b) Experimental error of the quantitative NMR analysis

The experimental error of the quantitative NMR analysis was well below 5% in most cases here, cf. Section 3 Sample Preparation and NMR Spectroscopy, and, thus, of only minor influence on the results of the present work.

(c) Experimental error of the PFG NMR experiments

The experimental error of the PFG NMR experiments, resulting in uncertainties in the measured diffusion coefficients, was also very small, namely, in average in the order of 2%, cf. Figures 2, 4, and 6 in the manuscript.

(d) Errors introduced by the SEGWE model

Errors introduced by the SEGWE model have a direct influence on the molar masses predicted from the measured diffusion coefficients. In the original paper,[15] the authors

reported a root-mean-square deviation in the order of 15% for predicted diffusion coefficients. The rather large expected errors comply with the observations in Figures 8 – 10 of the manuscript. We can therefore consider the errors introduced by the SEGWE model as the main source of error for the results of the present work.

(e) Experimental error of the diffusion coefficient of the defined reference component
For the application of the proposed method, also the diffusion coefficient of a known reference component at infinite dilution in the solvent of the poorly specified mixture is required. In the present work, we have adopted the respective experimental values from the literature. Of course, also these values come with an uncertainty, which can introduce an additional error of the proposed method's results.

Additional Results

NMR Fingerprinting

Figure S.6 shows the results of the NMR fingerprinting in the form of group mole fractions x_g. In mixture I (Figure S.6 (a)), the group mole fractions are predicted very accurately. Small deviations can be attributed to experimental uncertainties of the NMR analysis. Also in mixture II (Figure S.6 (b)), the agreement is good in most cases. Small deviations can be found due to the misinterpretation of 'OH' and 'CH2' as 'cy-CH2O' groups. Furthermore the 'CHO' group is missed by our method. In mixture III (Figure S.6 (c)) the 'CH3CN' group (=acetonitrile) is missed and falsely predicted as 'C=C' and 'CH3' groups. Furthermore, a small amount of the ester group ('COO') is missed leading to an overprediction of the 'COOH' group. 'CH2CO' and 'CH3CO' groups can furthermore not be differentiated here, since no distinction between different pseudo-components is made here.

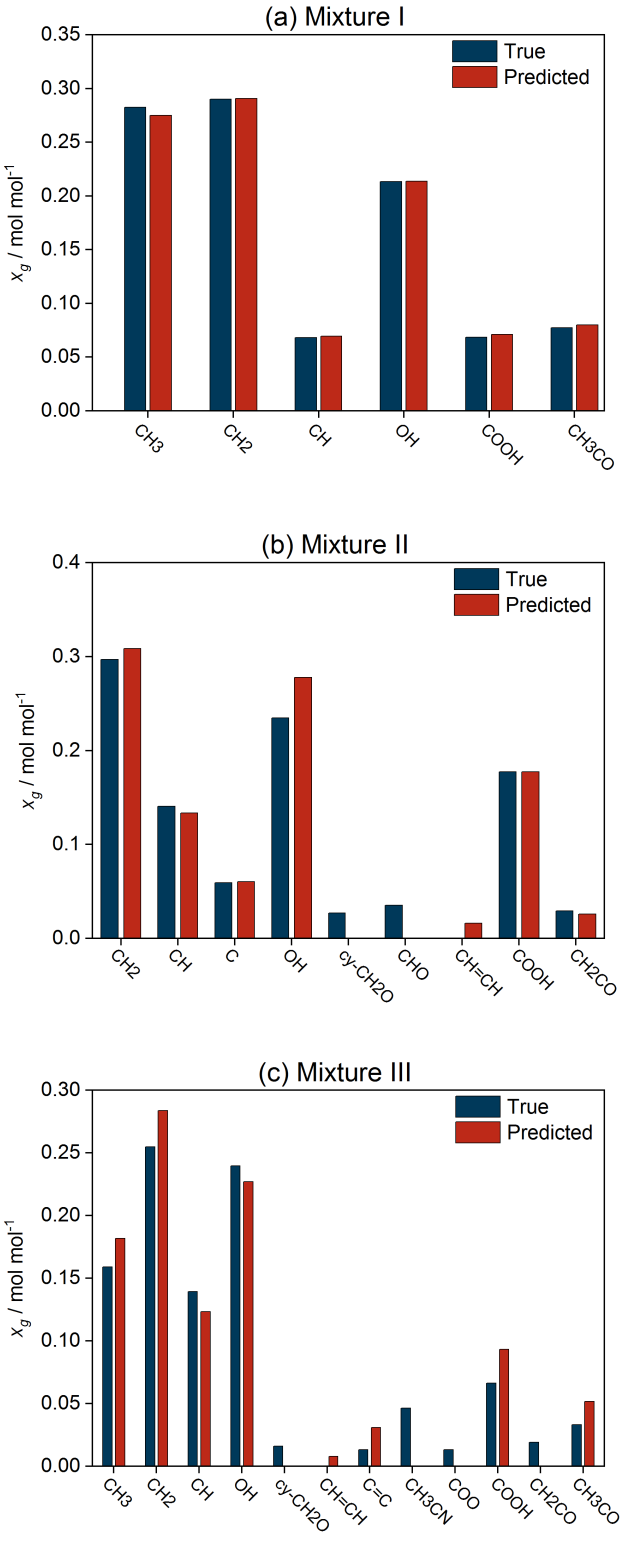

Figure S.6: Prediction of structural groups in test mixtures, cf. Table 2 in the manuscript.

Clustering with Prior Information

Figure S.7 shows the results of the clustering with the K-medians algorithm for mixture II from the manuscript, but here by fixing $K = 4$, which is the true number of components (except water and neglecting the anomers of glucose) in the mixture. Hence, in this case, a sort of prior information (on the number of components in the mixture) was used instead of automatically choosing K based on the overall silhouette score. The results show that, in this case, the clustering algorithm correctly assigns all peaks (structural) groups to the different pseudo-components.

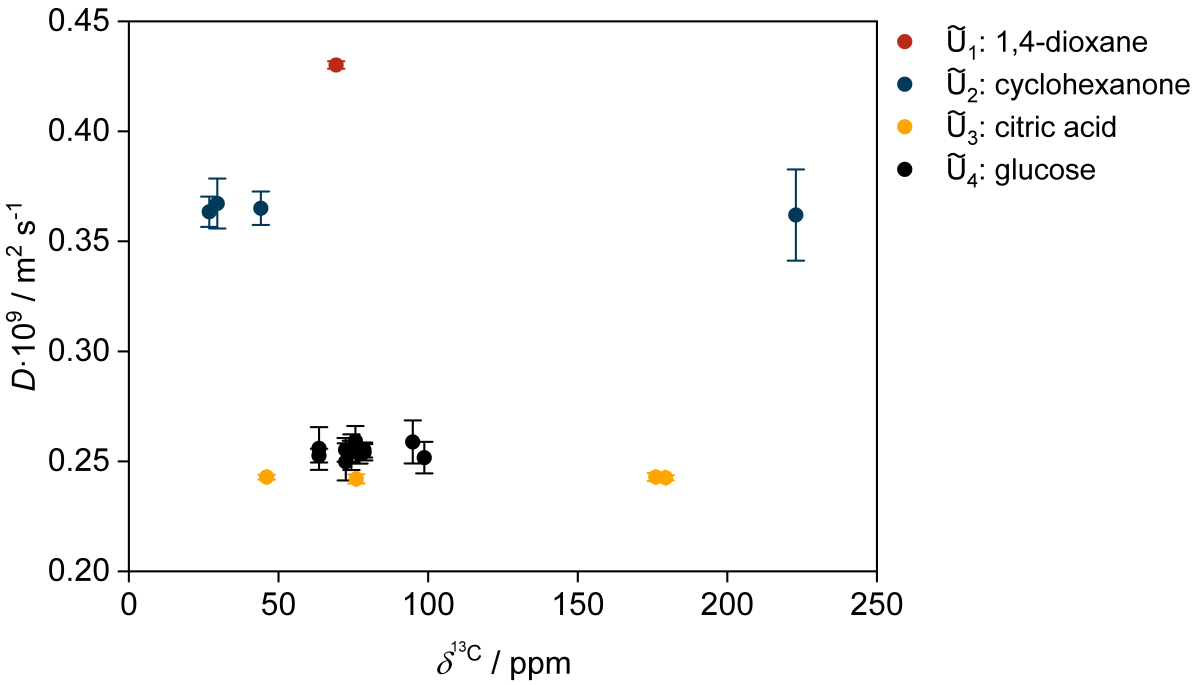

Figure S.7: DOSY map of mixture II with the result of the clustering of peaks (structural groups) by the K-medians algorithm and setting the number of clusters to $K = 4$. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate the 95 % confidence intervals based on a t-distribution.

In Figure S.8 results of the clustering with the K-medians algorithm for mixture III from the manuscript are shown but here by fixing $K = 10$, which is the true number of components (except water and neglecting the anomers of xylose, cf. manuscript) in the mixture. By using this prior knowledge, the clustering algorithm correctly assigns all peaks (structural) groups

to the different pseudo-components.

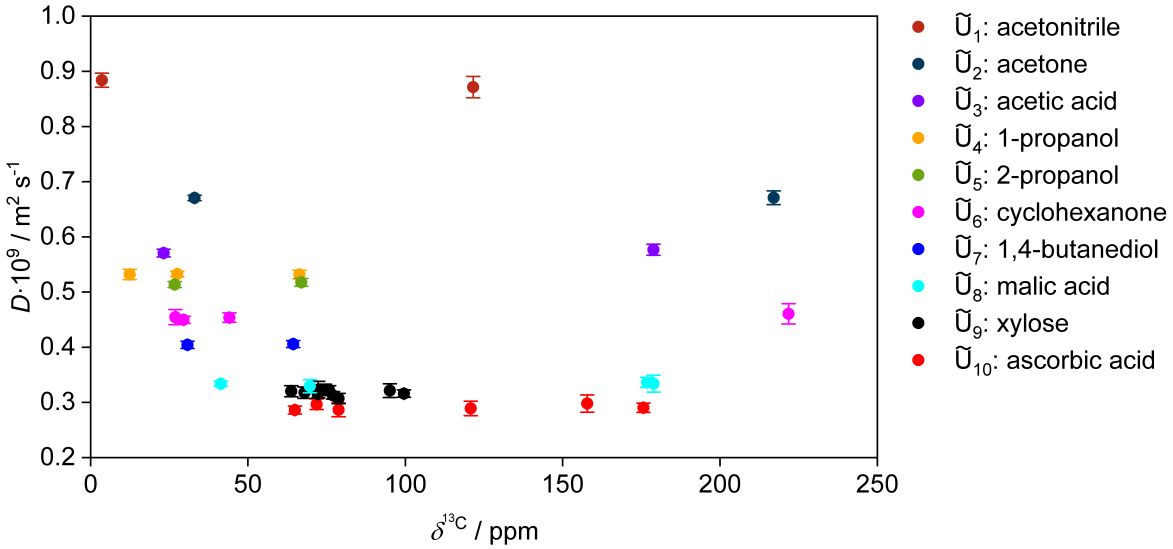

Figure S.8: DOSY map of mixture III with the result of the clustering of peaks (structural groups) by the K-medians algorithm and setting the number of clusters to $K = 10$. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate the 95 % confidence intervals based on a t-distribution.

Influence of Reference Component on Predicted Molar Masses

In Figure S.9, the prediction of the molar masses of the pseudo-components defined by the K-medians algorithm in mixture III, cf. Figure 6 in the manuscript, with the SEGWE model, is shown. In contrast to Figure 10 in the manuscript, xylose (instead of acetonitrile) was chosen as reference component. A value for the diffusion coefficient of xylose at infinite dilution in water at 298.15 K of $D_{\mathrm{ref}}^{\infty} = 7{:}495 \cdot 10^{-10}\mathrm{m}^2\mathrm{s}^{-1}$ was adopted from Ref.[21] and used in Eq. (4) from the manuscript.

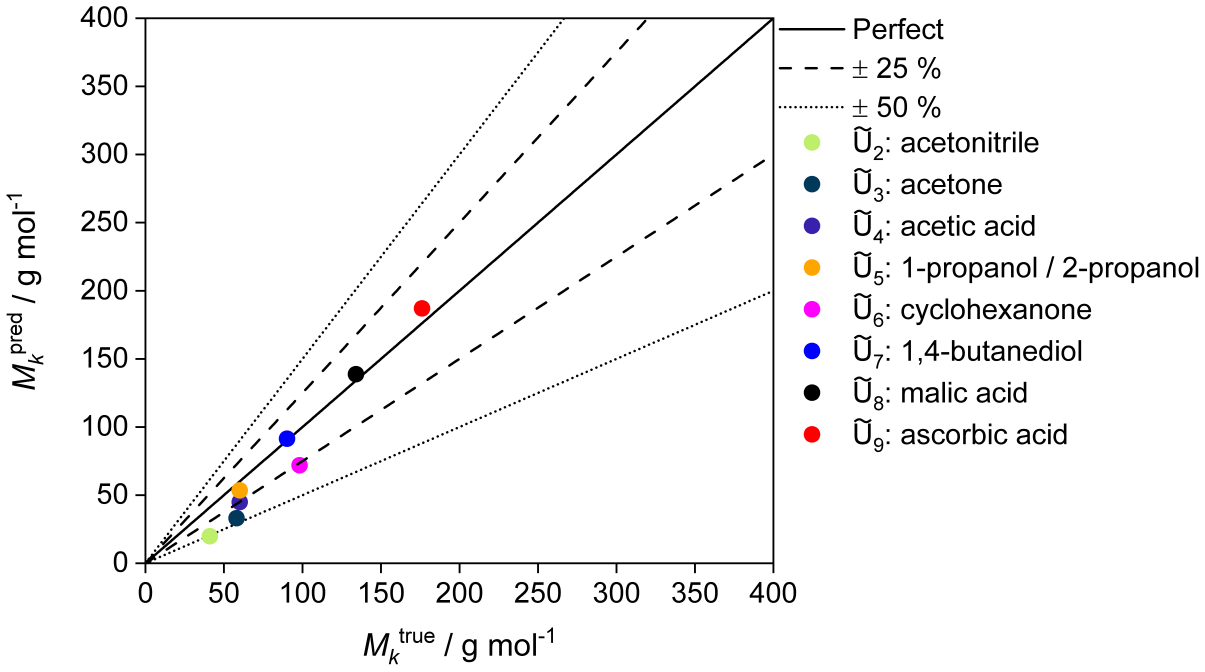

Figure S.9: Prediction of molar masses of pseudo-components in mixture III by considering xylose as reference component.

If we compare the results shown here to the results in Figure 10 in the manuscript, we observe an improved prediction of the molar masses of 1,4-butanediol, malic acid, and ascorbic acid in Figure S.9. However, the prediction of the molar masses of the rather small and less polar components, like acetone and acetonitrile, is slightly worse compared to Figure 10 in the manuscript. We assign these findings to the fact that we assume a constant ratio $\frac{D_{\tilde{U}}}{D_{\text{ref}}}$ for the extrapolation from finite concentrations to infinite dilution for all pseudo-components $\tilde{U}$. The results indicate that a (slightly) different ratio for the different components could improve the results. We can furthermore speculate that chemically similar species, e.g., highly polar components, like 1,4-butanediol, xylose, malic acid, and ascorbic acid, can be treated well using the same ratio, but that this does not hold for less similar components like acetonitrile. Hence, in principle, it might be possible to exploit such knowledge to refine the ratio for the different pseudo-components (based on the group-specific composition that is automatically obtained with our method) in future work.

References

(1) Claridge, T. D., Ed. *High-Resolution NMR Techniques in Organic Chemistry*, 3rd ed.; Elsevier, 2016.

(2) 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.

(3) Bellaire, D.; Großmann, O.; Münnemann, K.; Hasse, H. 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* **2022**, *166*, 106691.

(4) Yemloul, M.; Castola, V.; Leclerc, S.; Canet, D. Self-diffusion Coefficients Obtained from Proton-decoupled Carbon-13 spectra for Analyzing a Mixture of Terpenes. *Magnetic Resonance in Chemistry* **2009**, *47*, 635–640.

(5) Stejskal, E. O.; Tanner, J. E. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. *The Journal of Chemical Physics* **1965**, *42*, 288–292.

(6) Whelan, C.; Harrell, G.; Wang, J. Understanding the K-Medians Problem. Proceedings of the International Conference on Scientific Computing. 2015.

(7) Bradley, P. S.; Mangasarian, O. L.; Street, W. N. Clustering via Concave Minimization. Proceedings of the 9th International Conference on Neural Information Processing Systems. Cambridge, MA, USA, 1996; p 368–374.

(8) MATLAB, *version 9.11.0 (R2021b)*; The MathWorks Inc.: Natick, Massachusetts, 2021.

(9) Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed.; Springer-Verlag: Berlin, Heidelberg, 2006.

(10) Rousseeuw, P. J. Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* **1987**, *20*, 53–65.

(11) Evans, R. The Interpretation of Small Molecule Diffusion Coefficients: Quantitative use of Diffusion-ordered NMR Spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2020**, *117*, 33–69.

(12) Li, D.; Keresztes, I.; Hopson, R.; Williard, P. G. Characterization of Reactive Intermediates by Multinuclear Diffusion-Ordered NMR Spectroscopy (DOSY). *Accounts of Chemical Research* **2009**, *42*, 270–280.

(13) Neufeld, R.; Stalke, D. Accurate Molecular Weight Determination of Small Molecules via DOSY-NMR by Using External Calibration Curves with Normalized Diffusion Coefficients. *Chemical Science* **2015**, *6*, 3354–3364.

(14) Evans, R.; Deng, Z.; Rogerson, A. K.; McLachlan, A. S.; Richards, J. J.; Nilsson, M.; Morris, G. A. Quantitative Interpretation of Diffusion-Ordered NMR Spectra: Can We Rationalize Small Molecule Diffusion Coefficients? *Angewandte Chemie International Edition* **2013**, *52*, 3199–3202.

(15) Evans, R.; Poggetto, G. D.; Nilsson, M.; Morris, G. A. Improving the Interpretation of Small Molecule Diffusion Coefficients. *Analytical Chemistry* **2018**, *90*, 3987–3994.

(16) Großmann, O.; Bellaire, D.; Hayer, N.; Jirasek, F.; Hasse, H. Database for Liquid Phase Diffusion Coefficients at Infinite Dilution at 298 K and Matrix Completion Methods for their Prediction. *Digital Discovery* **2022**, *1*, 886–897.

(17) Pratt, K. C.; Wakeham, W. A.; Ubbelohde, A. R. J. P. 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* **1975**, *342*, 401–419.

(18) Tyn, M. T.; Calus, W. F. Temperature and Concentration Dependence of Mutual Diffusion Coefficients of Some Binary Liquid Systems. *Journal of Chemical & Engineering Data* **1975**, *20*, 310–316.

(19) Online Group Assignment for UNIFAC. http://www.ddbst.com/unifacga.html (Last accessed: 23.03.2023).

(20) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures. *AIChE Journal* **1975**, *21*, 1086–1099.

(21) Ueadaira, H.; Uedaira, H. Diffusion Coefficients of Xylose and Maltose in Aqueous Solution. *Bulletin of the Chemical Society of Japan* **1969**, *42*, 2140–2142.