Thomas Specht
Thermodynamic Modeling of Poorly Specified Mixtures using NMR Fingerprinting and Machine Learning
Scientific Report Series Volume 42
2023

Scientific Report Series
Laboratory of Engineering Thermodynamics (LTD)
RPTU Kaiserslautern
P.O. Box 3049
67663 Kaiserslautern
Germany

ISSN 2195-7606
ISBN 978-3-944433-41-7

© LTD all rights reserved

Thermodynamic Modeling of Poorly Specified Mixtures using NMR Fingerprinting and Machine Learning

Vom Fachbereich Maschinenbau und Verfahrenstechnik
der Rheinland-Pfälzischen Technischen Universität
Kaiserslautern-Landau
zur Verleihung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)
genehmigte
Dissertation

von
M. Sc. Thomas Specht
aus Koblenz

Dekan:	Prof. Dr. rer. nat. Roland Ulber
Berichterstatter:	Prof. Dr.-Ing. Hans Hasse
	Jun. Prof. Dr.-Ing. Fabian Jirasek
	Prof. Dr. Werner R. Thiel

Tag der mündlichen Prüfung: 16.06.2023

D 386

Danksagung

Diese Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Thermodynamik (LTD) der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau (ehemals Technische Universität Kaiserslautern). Zum Erfolg der Arbeit haben viele Leute beitragen, sodass ich hoffe, niemanden vergessen zu haben.

Zunächst möchte ich mich bei Prof. Dr.-Ing. Hans Hasse für die hervorragende Betreuung bedanken. Durch seine offene Art und Weise sind viele wertvolle Diskussionen entstanden, ohne die diese Arbeit heute so nicht vorliegen würde. Ohne deinen Vertrauensvorschuss und die Freiheit während der Promotion hätte ich mich nicht so entfalten können.

Ebenso möchte ich JP. Dr.-Ing. Fabian Jirasek danken, der mich während der Promotion immer unterstützt hat und stets offen für neue Vorschläge war. Danke auch dafür, dass du dir bei jeglicher Art von Problemen Zeit genommen hast, diese zu diskutieren. Ich möchte sowohl JP. Dr.-Ing. Fabian Jirasek als auch Prof. Dr.-Ing. Hans Hasse weiterhin für die berufliche Förderung über die Promotion hinaus danken. Ich danke weiterhin Prof. Dr. Werner R. Thiel für die Begutachtung meiner Arbeit und Prof. Dr.-Ing. Erik von Harbou für die Übernahme des Vorsitzes der Prüfungskommission und auch dafür das er mich letztendlich an den LTD gebracht hat.

Mein Dank gilt dem gesamten LTD-Team. Ich möchte insbesondere Marlies Mangold, Jennifer Bergmann und Ilona Stein für die hervorragende Organisation am Lehrstuhl danken. Danke auch, dass ihr mich immer an mein Postfach erinnert habt :-). Daniel Fröscher möchte ich für die schnelle Behebung von jeglichen IT-Problemen danken und der großen Unterstützung in der digitalen Lehre. Ich möchte auch dem gesamten Laborteam danken, das mir immer mit Rat und Tat zur Seite stand.

Weiterhin möchte ich mich bei allen Studierenden bedanken, die eine Arbeit bei mir absolviert haben und damit wertvolle Beiträge zur Promotion geliefert haben. Mein besonderer Dank gilt auch allen hilfswissenschaftlichen Mitarbeitern, die mich während meiner Promotion unterstützt haben. Mein Dank gilt insbesondere Johannes Stüber, der während der gesamten Promotion angestellt war und mich das ein oder andere Mal mit seinen unkonventionellen Lösungsvorschlägen weitergebracht hat und ohne den diese Arbeit ebenfalls nicht möglich gewesen wäre.

Danke an alle Kollegen am LTD, die immer zu einer sehr schönen Arbeitsumgebung beigetragen haben und die super Atmosphäre am Lehrstuhl ausmachen. Besonderer Dank gilt meinem Bürokollegen Johnnie Phuong, der keiner (thermodynamischen) Diskussion aus dem Weg gegangen ist und mir immer mit Rat und Tat zur Seite stand.
Ich möchte meiner Familie und insbesondere meinen Eltern danken, die mich im Studium immer unterstützt und mir damit den Weg zur Promotion geebnet haben. Zuletzt möchte ich meiner Freundin Laura für den Rückhalt und Support, den du mir während der ganzen Promotion gegeben hast, danken. Danke, dass du mich immer von allen Zweifeln befreit hast, für die vielen schönen Momente in den letzten Jahren und dafür, dass du mir immer vor Augen führst, worauf es im Leben wirklich ankommt.

Irvine, Kalifornien, August 2023

Thomas Specht

Abstract

Poorly specified mixtures, i.e., mixtures of unknown or incompletely known composition, are common in many fields of process engineering. Dealing with such mixtures in process design is challenging as their properties cannot be described with classical thermodynamic models, which require a full specification. As a workaround, pseudo-components can be introduced, which are generally defined using ad-hoc assumptions. In the present thesis, a new framework is developed for the thermodynamic modeling of such mixtures using nuclear magnetic resonance (NMR) experiments in combination with machine-learning (ML) methods. In the framework, a characterization of a mixture in terms of structural groups (“NMR fingerprint”) is obtained by using the ML concept of support vector classification. Based on the group-specific fingerprint, quantum-chemical descriptors of the unknown part of the mixture as well as activity coefficients can already be predicted. Furthermore, a meaningful definition of pseudo-components is achieved by clustering the structural groups into pseudo-components with the K-medians algorithm based on their self-diffusion coefficients measured by pulsed-field gradient (PFG) NMR. It is demonstrated that the characterization of poorly specified mixtures in terms of pseudo-components can be combined with several thermodynamic group-contribution methods. The resulting thermodynamic models were applied to various poorly specified mixtures and used for solving two typical tasks from conceptual fluid separation process design: the solvent screening for liquid-liquid extraction processes and the simulation of open evaporation processes. The predictions with the methods developed here show very good agreement with the results obtained for the fully specified mixtures.

Kurzfassung

Schlecht spezifizierte Mischungen, d. h. Mischungen mit unbekannter oder unvollständig bekannter Zusammensetzung, treten in vielen Bereichen der Verfahrenstechnik auf. Solche Mischungen stellen insbesondere bei der Prozessauslegung eine Herausforderung dar, da ihre Eigenschaften nicht mit klassischen thermodynamischen Modellen beschrieben werden können, die eine vollständige Spezifikation der Zusammensetzung verlangen. Oftmals werden deshalb Pseudokomponenten definiert, wobei in der Regel allerdings pauschale Annahmen über die chemische Natur und Anzahl dieser getroffen werden. In der vorliegenden Dissertation wird ein Framework eingeführt, das die thermodynamische Modellierung schlecht spezifizierter Mischungen mit Hilfe von Kernspinresonanz (NMR)-Experimenten in Kombination mit Methoden des maschinellen Lernens (ML) ermöglicht. Die Charakterisierung einer Mischung erfolgt dabei auf Basis von Strukturgruppen ("NMR fingerprint") unter Verwendung des ML-Konzepts der Support-Vektor-Klassifikation. Basierend auf solch einer gruppenspezifischen Charakterisierung können bereits quantenchemische Deskriptoren des unbekannten Teils der Mischung sowie Aktivitätskoeffizienten abgeschätzt werden. Anschließend wird eine rationale Definition von Pseudokomponenten erzielt, indem die strukturellen Gruppen mit dem K-Medians-Algorithmus auf der Grundlage ihrer mittels Pulsed-Field-Gradient (PFG)-NMR gemessenen Selbstdiffusionskoeffizienten zu Pseudokomponenten geclustert werden. Es wurde gezeigt, dass die so erhaltene Charakterisierung von schlecht spezifizierten Mischungen mit verschiedenen thermodynamischen Gruppenbeitragsmethoden kombiniert werden kann. Die sich daraus ergebenden thermodynamischen Modelle wurden für die Lösung zweier typischer Aufgaben aus dem konzeptionellen Design von Trennprozessen verwendet: Für das Lösungsmittel-Screening für Flüssig-Flüssig-Extraktionsprozesse und für die Simulation von offenen Verdampfungsprozessen. Die Vorhersagen mit dem entwickelten Ansatz zeigen eine sehr gute Übereinstimmung mit den Ergebnissen für voll spezifizierte Mischungen.

Contents

1 Introduction 1

2 NMR Fingerprinting based on ^{1}H and ^{13}C NMR 5

- 2.1 Introduction 5
- 2.2 Data and Methods 8
  - 2.2.1 NMR Spectra of Pure Components 8
  - 2.2.2 NMR Spectra of Mixtures 11
  - 2.2.3 Definition of Input and Output Data 11
  - 2.2.4 Support Vector Classification 12
  - 2.2.5 Multi-label Classification 14
    - 2.2.5.1 Classification Scores 14
    - 2.2.5.2 Training, Hyperparameter Optimization, Evaluation with Pure Component Data 15
    - 2.2.5.3 Application to Mixture Data 16
- 2.3 Results 17
  - 2.3.1 Structural Group Analysis Based on Pure Component Spectra . . 17
    - 2.3.1.1 Group-identification Method 17
    - 2.3.1.2 Group-assignment Method 20
  - 2.3.2 Structural Group Analysis in Mixture Spectra 23
    - 2.3.2.1 Group-identification Method 24
    - 2.3.2.2 Group-assignment Method 25
- 2.4 Conclusions 30

3 NMR Fingerprinting based on ^{1}H, ^{13}C, and ^{13}C DEPT NMR 33

- 3.1 Introduction 33
- 3.2 Overview of the Workflow 34
- 3.3 Data and Methods 37
  - 3.3.1 Generation of Input and Output Data for Training 37
  - 3.3.2 Collection of Pure-component NMR Data 38
  - 3.3.3 Training of Support Vector Classification 41
  - 3.3.4 Experimental Methods 42

3.4 Results and Discussion . . . 43
- 3.4.1 Prediction of Structural Groups from Pure-component Spectra . . 43
- 3.4.2 Prediction of Structural Groups from Mixture Spectra . . . 44
  - 3.4.2.1 Results for Aqueous Mixtures . . . 45
  - 3.4.2.2 Results for Organic Mixtures . . . 47

3.5 Conclusions . . . 51

4 Definition of Pseudo-components 53

4.1 Introduction . . . 53

4.2 Overview of the Method . . . 54
- 4.2.1 NMR Experiments . . . 58
- 4.2.2 Identification and Quantification of Structural Groups . . . 58
- 4.2.3 Clustering of Structural Groups to Pseudo-components . . . 60
  - 4.2.3.1 Determination of Self-diffusion Coefficients of Structural Groups . . . 60
  - 4.2.3.2 Clustering Algorithm . . . 61
- 4.2.4 Prediction of Molar Masses . . . 62

4.3 Overview of Applications . . . 63

4.4 Results and Discussion . . . 64
- 4.4.1 Prediction of Structural Groups and Clustering to Pseudo-Components . . . 64
  - 4.4.1.1 Mixture I . . . 65
  - 4.4.1.2 Mixture II . . . 66
  - 4.4.1.3 Mixture III . . . 68
- 4.4.2 Prediction of Molar Masses of Pseudo-components . . . 70

4.5 Conclusions . . . 73

5 Estimation of σ-Profiles and Activity Coefficients 75

5.1 Introduction . . . 75

5.2 Materials and Methods . . . 77
- 5.2.1 Materials . . . 77
- 5.2.2 Methods . . . 78
  - 5.2.2.1 Sample Preparation and NMR Spectroscopy . . . 78
  - 5.2.2.2 Assignment and Quantification of Chemical Groups . . . 78
  - 5.2.2.3 Calculation of Activity Coefficients . . . 80

5.3 Results and Discussion . . . 80

5.4 Conclusions . . . 87

6 Conceptual Fluid Separation Process Design 89

6.1 Introduction . . . 89

6.2 Methods . . . 90
  6.2.1 Overview of the Workflow . . . 90
  6.2.2 NMR Fingerprinting and Pseudo-component Method . . . 92
  6.2.3 Simulation of Liquid-liquid Extraction Processes . . . 93
  6.2.4 Prediction of Residue Curves . . . 93
6.3 Overview of Studied Mixtures . . . 94
6.4 Results and Discussion . . . 96
  6.4.1 Solvent Screening for Liquid-liquid Extraction Processes . . . 96
  6.4.2 Simulation of Open Evaporation Processes . . . 100
6.5 Conclusions . . . 104

7 Conclusions 107

Literature 109

Appendix 125

A Supporting Information for Chapter 2 125

A.1 Experimental Methods . . . 125
  A.1.1 Preparation of Samples and Acquisition of NMR Spectra . . . 126
  A.1.2 Quantitative Evaluation and Comparison to Ground Truth . . . 126
A.2 NMR Spectra of Mixtures . . . 127
A.3 Influence of NMR Spectra Binning on Validation Score . . . 131
A.4 Handling of Special Cases . . . 131
  A.4.1 Classification Scores . . . 131
A.5 Step-by-step Example for the Group-assignment Method . . . 132
A.6 Linearly and Non-linearly Separable Data Sets . . . 137
A.7 Performance of the Group-identification Method for Different Solvents . . 137
A.8 Additional Results of the Group-assignment Method . . . 138
A.9 Information on Studied Pure Components . . . 144
  A.9.1 Priority Rules for Structural Group Division . . . 144
  A.9.2 Division of All Considered Pure Components . . . 144

B Supporting Information for Chapter 3 207

B.1 Data . . . 207
  B.1.1 Processing of Pure-component Spectra . . . 207
    B.1.1.1 NMRShiftDB . . . 207
    B.1.1.2 BMRB . . . 207
  B.1.2 Further Processing Steps . . . 208
B.2 Data Overview by Consideration of Substitution Degrees of Carbon Atom 209

B.3 Modeling and Computational Details 210
  B.3.1 Application of Method to Mixture Data 211
B.4 Experimental Methods 212
  B.4.1 Preparation of Samples and Acquisition of NMR Spectra 213
  B.4.2 Determination of Substitution Degree 213
  B.4.3 Determination of Labile Protons 214
  B.4.4 Quantitative Evaluation and Comparison to Ground Truth 216
B.5 NMR Spectra of Mixtures 218
B.6 Additional Results for Variant of Method without Information about Labile Protons 225
  B.6.1 Prediction of Structural Groups from Pure-component Spectra . . 225
  B.6.2 Prediction of Structural Groups from Mixture Spectra 226

C Supporting Information for Chapter 4 231
C.1 Experimental Methods 231
  C.1.1 Chemicals 231
  C.1.2 NMR Analysis 232
    C.1.2.1 Sample Preparation and NMR Spectroscopy 232
    C.1.2.2 PFG NMR Spectroscopy 232
C.2 Distinction between Substitution Degrees with DEPT NMR 233
C.3 PFG NMR Spectra and Assignment of Peaks 234
C.4 K-medians Algorithm and Silhouette Score 236
  C.4.1 Calculation of Silhouette Coefficients for Single Data Points . . . 237
C.5 Prediction of Molar Masses and Normalized Diffusion Coefficients 238
C.6 Concentration Dependence of Relative Diffusion Coefficients 240
C.7 Identification and Quantification of Structural Groups 242
C.8 Determination of Water-free Composition of Pseudo-components 242
C.9 Structural Group Composition 244
  C.9.1 Composition of True Components 244
  C.9.2 Predicted Molar Masses and Composition of Pseudo-components 246
  C.9.3 Discussion of Uncertainties 249
C.10 Additional Results 250
  C.10.1 NMR Fingerprinting 250
  C.10.2 Clustering with Prior Information 252
  C.10.3 Influence of Reference Component on Predicted Molar Masses . . 253

D Supporting Information for Chapter 5 255
D.1 Information on Water Mass Fraction 255

D.2 Group Assignment 260
D.3 Influence of the Molar Mass $M_{\tilde{\mathrm{U}}}$ of the Pseudo-Component Ũ 262
D.4 Stoichiometry of the Pseudo-Component Ũ and Estimated Composition . 263
D.4.1 Unknown Water Mass Fraction 263
D.4.2 Known Water Mass Fraction 270
D.5 Averaging of σ-profiles 277
D.6 Additional σ-profiles 278
D.7 Computational Details 282
D.8 NMR Spectra of the Systems 283
D.9 Step-by-step Example for NEAT Application 287
D.10 Temperature Dependency of Activity Coefficients 289

E Supporting Information for Chapter 6 291
E.1 Calculation of Liquid-liquid Equilibria 291
E.2 Prediction of Residue Curves 292
E.3 Prediction of Feed Composition 293
E.4 Predicted Composition of Pseudo-components 295
E.5 Results for Pure-component Vapor Pressures 302
E.6 Additional Results for the Prediction of Residue Curves 306

List of Symbols

Latin symbols

A	cavity surface area
A_g	total area associated to structural group g
b	offset in SVC
C	hyperparameter in SVC
c_1, c_2	probe-specific fitting parameters in modified Stejskal-Tanner equation
d	decision function of SVC
$\mathcal{D}$	data set
D_p	self-diffusion coefficient of peak p
$e_{p,95\%}$	experimental uncertainty of D_p specified by 95% confidence interval based on a t-distribution
$F_{1,g}$	F_1 score for structural group g / weighted F_1 score for structural group g
F_1^{macro}	arithmetic mean of F_1 scores
g	structural group
G	total number of structural groups
G	gradient strength
G^{E}	Gibbs excess energy
$I_{0,p}$	measured height of peak p in absence of diffusion
I_p	measured height of peak p
K	total number of pseudo-components
k_{B}	Boltzmann constant
M_i	molar mass of component i
M	number of mixtures
n	mole number
N	number of data points
N	number of components
N_{A}	Avogadro constant
N_g	number of pure components in which structural group g is present
N_g^s	number of pure components in which structural group g induces a peak in section s

p	pressure
P	total number of peaks in ^{13}C NMR spectrum
p_i^{S}	vapor pressure of pure component i
P_g	precision of structural group g
p_i	partial pressure of component i
R_g	recall of structural group g
$\overline{s}(K)$	overall silhouette score for K pseudo-components
$S^{^{13}\mathrm{C}}$	total number of sections in ^{13}C NMR spectrum
$S^{^{1}\mathrm{H}}$	total number of sections in ^{1}H NMR spectrum
T	temperature
T_1	spin–lattice relaxation time
U	unknown component
$\tilde{\mathrm{U}}$	pseudo-component
V	cavity volume
W	water
$\boldsymbol{w}$	normal weight vector in SVC
$\boldsymbol{x}$	input vector for SVC
$\boldsymbol{\mathcal{X}}^{\mathrm{mix}}$	set of vectors for clustering peaks in a mixture
$x_{\mathrm{U}}^{\mathrm{total}}$	sum of mole fraction of all unknown components
$x_{g,k}$	mole fraction of structural group g in pseudo-component k
x_g	group mole fraction
x_i	mole fraction of component i
x_k^*	water-free mole fraction of pseudo-component k
$\boldsymbol{x}_p$	input vector for peak p for clustering
$\boldsymbol{y}$	output vector for SVC
$\boldsymbol{Y}$	output matrix for SVC
Y_{T}	yield of target component
y_i	gas-phase mole fraction of component i
y_i	binary output SVC
z_g	number of NMR-active nuclei in structural group g in the same chemical shift region

Subscripts and Superscripts

0	initial value
$'$	raffinate phase
$''$	extract phase
∞	at infinite dilution
A	group-assignment method

ald	aldehyde
ar	aromatic
g	structural group
hb	hydrogen bonding
i	component
I	group-identification method
k	pseudo-component
ket	ketone
L	liquid
mix	mixture
nhb	non-hydrogen bonding
p	peak
pred	predicted
pure	pure component
ref	reference component
rel	relative
s	section
S	solvent
T	target component
T	transposed
U	unknown component
$\tilde{\mathrm{U}}$	pseudo-component
W	water
x	number of bonded protons

Greek symbols

α	ratio of molar masses of solvent and solute in the SEGWE model
α_{T}	separation factor of target component
β	evaporation ratio
Δ	diffusion time
δ	duration of gradient pulse
δ	chemical shift
η	dynamic viscosity
γ	gyromagnetic ratio
γ	activity coefficient
γ	hyperparameter in SVC
ν	stochiometric coefficient
$p(\sigma)$	normalized σ-profile

ϕ	transformation function
ρ_{eff}	lumped parameter in SEGWE model
σ	screening charge density
τ	correction constant in modified Stejskal-Tanner equation due to the usage of bipolar gradients
ξ	slack variable in SVC

Abbreviations

BMRB	Biological Magnetic Resonance Data Bank
COSMO-RS	conductor-like screening model for real solvents
CV	cross-validation
DDB	Dortmund Data Bank
DEPT	distortionless enhancement by polarization transfer
DFT	density-functional theory
DOSY	diffusion-ordered spectroscopy
FN	false negative
FP	false positive
GC-COSMO-RS (OL)	group-contribution version of COSMO-RS
HMBC	heteronuclear multiple bond correlation
HOSE	hierarchical organization of spherical environments
HSQC	heteronuclear single quantum coherence
IR	infrared
ML	machine learning
MS	mass spectrometry
NEAT	NMR spectroscopy for the estimation of activity coefficients of target components in poorly specified mixtures
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NRTL	non-random two-liquid
P, S, T, Q	primary, secondary, tertiary, quaternary
PENDANT	polarization enhancement nurtured during attached nucleus testing
PFG	pulsed-field gradient
SDBS	Spectral Database for Organic Compounds
SEGWE model	Stokes-Einstein Gierer-Wirtz estimation model
SLE	solid-liquid equilibrium
SMARTS	SMILES arbitrary target specification
SMILES	simplified molecular-input line-entry system

SVC	support vector classification
T	target component
TMSP-d4	sodium 3-(trimethylsilyl)tetradeuteriopropionate
TP	true positive
U	unknown component
UNIFAC	universal quasichemical functional group activity coefficients
UNIFAC (DO)	modified UNIFAC (Dortmund)
UNIQUAC	universal quasichemical
$\tilde{\mathrm{U}}$	pseudo-component
W	water

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. The present thesis deals with this situation and thereby focuses on liquid mixtures of which the composition is not (completely) known, which are called "poorly specified mixtures" here, and proposes a new, rational framework for their characterization and thermodynamic modeling without relying on ad-hoc assumptions.

A poor mixture specification can have different origins and manifestations. In the ideal picture, 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 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 addressed in this thesis.

Well-known examples of such mixtures are petroleum oil fractions [1–3], polymerization products [4, 5], fermentation broths [6], or mixtures arising during waste-water treatment [7]. 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 [8–13]. 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" [14, 15]. 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.

Nuclear magnetic resonance (NMR) spectroscopy is a powerful tool for the structural elucidation and quantification of components in liquid mixtures [16–21]. To facilitate the identification of unknown *pure* components from NMR spectra, several tools are available and have partly been implemented in commercial software, such as MestReNova (Mestrelab Research), ACD/Labs (Advanced Chemistry Development Inc.), and CMC-se (Bruker) [22]. These tools usually rely on 1D and 2D NMR spectra of the sample and, in all cases, require the knowledge of the empirical formula (the ratio of the composing elements) of the component to be identified, which has to be determined in additional analysis, such as mass spectrometry (MS) [22].

Compared to the case of pure components, the comprehensive elucidation of unknown components in *mixtures* from NMR spectra is a much more difficult and time-consuming task [23]. Its solution generally requires chemical intuition and good knowledge of NMR spectra and there are many examples in the literature that describe how this task was solved for complex mixtures [16–21]. One approach that is applied in the field of metabolomics is dereplication, which can be used as a tool to provide the user with a set of most likely present components by comparing the sample spectra against a local database of pure-component spectra as demonstrated in Ref. [24]. Hence, only components that are part of the predefined database can be identified by dereplication.

The methods introduced in this thesis follow a different approach yielding a *structural group-specific* characterization of an unknown sample and thereby circumventing the inherent problems of a *component-specific* elucidation. The group-specific characterization of a mixture based on NMR experiments is called “NMR fingerprint” here. The information on the group speciation of a mixture can be retrieved from NMR spectra more readily than information on the component speciation. There are actually many relevant tasks for which a fingerprinting of unknown mixtures regarding the structural groups they contain is sufficient. For instance, group-specific characterizations are used as a first step to structure elucidation or for the prediction of reactivities [25].

Furthermore, based on an NMR fingerprint, pseudo-components in a poorly specified mixture can be defined in a rational way, namely, by the clustering of the structural groups based on self-diffusion coefficients measured by pulsed-field gradient (PFG) NMR. Since the self-diffusion coefficients also encode information about the molar mass of the pseudo-components, also the size of the pseudo-components can thereby be estimated.

The obtained characterization of poorly specified mixtures in terms of pseudo-components is all that is needed to calculate different thermodynamic properties by group-contribution methods [26–30], most notably activity coefficients [31, 32], which describe the non-ideality in the liquid phase and are, therefore, the basis for modeling phase equilibria and simulating thermal separation processes.

The present thesis is organized as follows:

In Chapters 2 and 3, NMR fingerprinting methods based on different basic NMR experiments using support vector classification are presented, whereby the method developed in Chapter 2 relies only on ^{1}H and ^{13}C NMR spectra. The method in Chapter 3 incorporates additional information from ^{13}C distortionless enhancement by polarization transfer (DEPT) NMR experiments. In Chapter 4, a methodology for the rational definition of pseudo-components based on an NMR fingerprint and a clustering approach is introduced. In Chapter 5, the determination of quantum-chemical descriptors and activity coefficients in poorly specified mixtures is demonstrated using a simple pseudo-component approach. In Chapter 6, a rigorous application of the NMR fingerprinting approach and the pseudo-component method for thermodynamic modeling and simulation of thermal separation processes involving poorly specified mixtures is presented.

2 NMR Fingerprinting based on ^{1}H and ^{13}C NMR

2.1 Introduction

With NMR spectroscopy, different NMR-active nuclei can be studied, the most common ones are ^{1}H and ^{13}C. In principle, each chemically non-equivalent proton or carbon in a sample leads to an individual peak in the ^{1}H or ^{13}C NMR spectrum, respectively. However, overlapping peaks are common, particularly in ^{1}H NMR spectroscopy. The position of each peak in a spectrum, also called the chemical shift, depends on the chemical environment of the corresponding nucleus, i.e., on the neighboring atoms or, in other words, the structural group to which the nucleus belongs. Structural groups can be defined in different ways. For several common groups, characteristic chemical shift ranges are reported in form of chemical shift tables for ^{1}H and ^{13}C NMR [25, 33]. These tables can be used for identifying the structural groups in unknown samples from NMR spectra. However, in many cases, no unambiguous decision can be made, since the characteristic chemical shift ranges of most structural groups overlap with those of other groups. Hence, for reliably identifying structural groups, it is usually necessary to study NMR spectra and chemical shift tables of different nuclei, and to use chemical intuition. Furthermore, the reported chemical shift tables are biased by the personal experience of the respective authors or are even not consistent, see, e.g., Refs. [25, 33]. Meanwhile, also computational approaches, such as methods based on density-functional theory (DFT) [34] and the hierarchical organization of spherical environments (HOSE) [35] method are available that can be used to predict the NMR spectra of a component from its chemical structure. However, their application for the reverse problem, i.e., the analysis of structural groups (or components) from NMR spectra, is not straightforward.

In this chapter, an automated method is introduced for identifying structural groups in unknown samples from NMR spectra. It is based on machine learning (ML), a rapidly developing branch of science that aims to extract information from data to perform specific tasks without requiring explicit user-defined rules (even though user-defined rules can in principle be included in ML methods). ML has already been used in combination with NMR spectroscopy many times, e.g., for the purpose of pattern recognition [36, 37], metabolite fingerprinting [38], or in medical diagnostics [39]. In addition, Ref. [40] gives an overview of ML methods for structure verification and discrimination with NMR spectroscopy, for which an a-priori estimation of the molecular structure of the component to be identified is required. Furthermore, Ref. [41] applied support vector regression for determining the content of self-defined component classes ("saturates", "aromatics" , and "polars") in crude oil based on ^{13}C NMR spectra after training their method to mixture data. This approach was extended by Ref. [42] using data from infrared (IR), ^{1}H, and ^{13}C NMR spectroscopy. Ref. [43] used ^{13}C NMR spectra for the identification of natural product classes in unknown samples using ML.

A few automated methods for the identification of structural groups in pure components have been described in the literature before: Refs. [44, 45] used discriminant functions to identify structural groups in unknown pure components with ^{13}C NMR spectroscopy alone. This chapter takes not only a different mathematical approach for the discrimination, but uses also ^{1}H NMR spectra, besides ^{13}C NMR spectra. Furthermore, it is also applied to mixtures. In a recent work [46], different analytical techniques, namely MS and IR spectroscopy, were used for the identification of structural groups from over 7000 spectra of pure components and applied to binary mixtures using a deep-learning approach.

Most importantly, all approaches reported in the literature to date are restricted to the *identification* of structural groups in unknown samples and, hence, yield only *qualitative* information on the composition of the sample. Such information can be of great practical interest, e.g., for process and reaction control or as the first step for structure elucidation. However, for many purposes, *quantitative* information is required, e.g., for modeling reaction and phase equilibria during process design and optimization. In NMR spectroscopy, quantitative information is obtained by integration of the peaks in the NMR spectra. Therefore, an approach for quantifying the structural groups in unknown samples has to identify the groups *and assign* them to peaks in the NMR spectra to enable a targeted integration. To the best of the authors' knowledge, no generic approach that automates the quantitative structural group characterization of structural groups in unknown samples based on NMR spectroscopy has been described previously in the literature.

Two different methods to elucidate structural groups in unknown samples are described in this chapter. Both methods are based on information from a ^{1}H and a ^{13}C NMR spectrum of the sample of interest. The first method yields a qualitative characterization of an unknown sample: it was developed and trained to *identify* structural groups and is therefore called *group-identification method* in the following. In the present version of this method, 13 structural groups are considered. The second method was developed and trained to identify the same structural groups in a sample *and to assign* them to peaks in the ^{13}C NMR spectrum of the sample. This method is therefore called *group-assignment method* in the following and constitutes the basis for a quantitative characterization of an unknown sample as described above. Of course, also in the group-assignment method, the task of group-identification is implicitly solved. There are, however, good reasons why this is not done exactly in the same way in both methods. Therefore, the group-identification method is presented separately here. Details on the differences are described below.

Both methods are based on the ML technique of support vector classification (SVC) [47]. A shallow ML approach is used here since the respective methods often work better on rather small data sets, as they are common in natural science and engineering and also present here, and are usually easier to interpret than deep-learning approaches [48]. SVC methods are binary classifiers that separate data points into classes (here: structural groups) by considering an input (here: ^{1}H and ^{13}C NMR spectrum of the sample). In the most simple – but unfortunately rare – situation of a linearly separable data set and only two distinguishable classes, a hard-margin SVC method maximizes the margin, i.e., the distance between the “nearest” representatives of the two classes [47]. However, SVC methods are flexible approaches that can be extended to also handle non-linearly separable data and to the distinction of more than two classes, which is a prerequisite for their application for solving the problems that are addressed here. Both methods introduced here were trained to NMR spectra of almost 1000 pure components from the Spectral Database for Organic Compounds (SDBS) [49]. As new routes are explored in the present chapter, the complexity of the task was limited: only organic compounds that contain the elements C, H, and O are considered. This class is still very large and was deemed suitable and interesting enough for the present tests. Furthermore, the focus here is on molecules of moderate size, not on macromolecules. After the training, both methods were tested by predicting the structural groups in pure components from NMR spectra from the SDBS database that were excluded from the training set. Furthermore, NMR spectra of three ternary mixtures were recorded in the present chapter and used for testing whether the methods work also for mixtures. The mixtures that were studied in the present chapter were selected to cover most of the considered structural groups, while containing components that were not in the (training) data set. No other information than a single ^{1}H and a single ^{13}C NMR spectrum of each sample was thereby used.

2.2 Data and Methods

2.2.1 NMR Spectra of Pure Components

For training and evaluation of the proposed methods, ^{1}H and ^{13}C NMR spectra of 985 pure components were taken from the Spectral Database for Organic Compounds (SDBS) [49]. Mixtures of enantiomers were considered as pure components since they cannot be distinguished in NMR spectroscopy. With few exceptions that are explained in Appendix A, all components that consist only of carbon (C), hydrogen (H), and oxygen (O), have a molar mass of up to 160 g mol^{-1}, contain not more than eight carbon atoms per molecule, and could be divided into the structural groups distinguished in this chapter, cf. Table 1. A list of all considered components is given in Table A.5 in Appendix A.

There are many different possibilities to divide components into structural groups. The choice, which groups to consider and how to exactly define them, depends on the nature of the problem that is considered; e.g., one may consider CH_3, CH_2, and CH as individual groups, or lump them together. In this chapter, 13 common structural groups are considered, which are listed in Table 1, covering a wide range of chemical diversity. The definition of structural groups is thereby inspired by the “main-groups” of a very popular group-contribution method for predicting thermodynamic properties, modified UNIFAC (Dortmund) [50]. All structural groups considered here comprise a single carbon nucleus, i.e., they yield a single peak in the proton-decoupled ^{13}C NMR spectrum. The respective group split for all components considered in the present chapter is given in Table A.5 in Appendix A. In most cases, the group split is unambiguous; in a few cases, additional rules are required, which are described in detail in Appendix A.

Table 1: Structural groups distinguished in the present chapter. Each group contains exactly one carbon.

Group label	Description
CH_3	Methyl group
CH_x	Linear alkyl group, $x \in \{0,1,2\}$
$cyCH_x$	Cyclic alkyl group, $x \in \{0,1,2\}$
CH_xOH	Aliphatic alcohol group, $x \in \{0,1,2,3\}$
CH_xO	Ether group, $x \in \{0,1,2,3\}$
$CH_x{=}$	Aliphatic double bond carbon, $x \in \{0,1,2\}$
$CH_x^{ar}{=}$	Aromatic carbon, $x \in \{0,1\}$
$RO{-}CH_x^{ar}{=}$	Aromatic carbon with oxygen substituent, $x \in \{0,1\}$
COOR	Carbonyl group in an ester/lactone/anhydride
$ROOCH_x$	Alkyl group attached to an ester/lactone oxygen, $x \in \{0,1,2,3\}$
COOH	Carboxyl group
CO^{ald}	Carbonyl group in an aldehyde
CO^{ket}	Carbonyl group in a ketone

As described above, each structural group has a characteristic range in which its peaks appear in NMR spectra [25, 33]. In Figure 1, the ranges in the ^{13}C NMR spectrum are graphically represented for the 13 structural groups and the 985 pure components considered in this chapter. The color code in Figure 1 represents the number of pure components in the data set that contain the respective structural group g that shows a peak in a specific section s of the ^{13}C NMR spectrum. That number is labeled here as N_g^s.

The data set is very imbalanced: some structural groups, such as CH_3 and CH_x, are part of basically all considered components, whereas other structural groups, such as COOH and CO^{ket}, are part of much less components. Moreover, the ranges of chemical shift, in which the peaks of a specific structural group appear, overlap substantially: in all regions of the ^{13}C NMR spectrum, peaks of at least two different structural groups are observed. Hence, the definition of fixed ranges in the ^{13}C NMR spectrum and simple assignment of all peaks in a specific range to a specific structural group does not lead to a satisfactory characterization of an unknown sample.

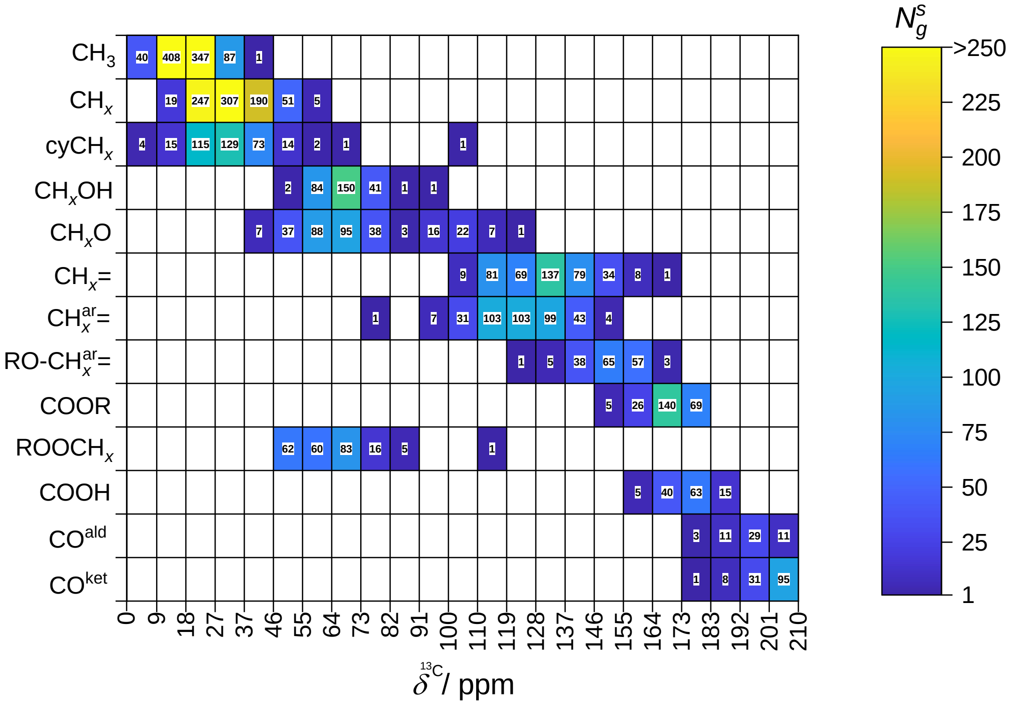


Figure 1: Positions of the peaks of the considered 985 pure components in the ^{13}C NMR spectrum. The color code and the numbers inside the cells represent N_g^s, which is the number of components that contain a specific structural group g (row) that induces a peak in a specific section s of the spectrum (column). White cells refer to $N_g^s = 0$.

In this chapter, the ^{1}H NMR and ^{13}C NMR spectra of all components were divided into discrete sections, characterized by a grid in the chemical shift, that was equally spaced both for ^{1}H NMR and ^{13}C NMR. The number of the sections is labeled here with $S^{^1\mathrm{H}}$ and $S^{^{13}\mathrm{C}}$, respectively. $S^{^1\mathrm{H}}$ and $S^{^{13}\mathrm{C}}$ are hyperparameters of the methods introduced here and were determined as described in more detail below. For too large numbers of $S^{^1\mathrm{H}}$ or $S^{^{13}\mathrm{C}}$, the discretization is too fine and there will be not enough data in each bin to enable reasonable learning. On the other hand, for too small numbers of $S^{^1\mathrm{H}}$ or $S^{^{13}\mathrm{C}}$, the discretization is too broad and the input data contain basically no information. For ^{1}H NMR spectra, the chemical shift range that was considered was 0-10 ppm; for ^{13}C NMR spectra, that range was 0-210 ppm. Consequently, each peak in an NMR spectrum of the considered components was assigned to a specific section according to the respective chemical shift as reported in the SDBS database [49], cf. Section 2.2.3. If numerical values for the peak positions were given in the SDBS database, they were adopted. In the few cases in which only a range of chemical shifts for a peak was reported in the SDBS database, the mean of the upper and lower limit was used as position of the peak. Peaks outside the defined ranges (< 0 ppm or > 10 ppm for ^{1}H NMR spectra and < 0 ppm or > 210 ppm for ^{13}C NMR spectra) were assigned to the respective nearest sections.

2.2.2 NMR Spectra of Mixtures

Besides pure component spectra, also ^{1}H and ^{13}C NMR spectra of three ternary mixtures were measured in this chapter and used for testing the presented methods regarding qualitative and quantitative structural group analysis. These mixture spectra were not used for training the methods. An overview of the chemicals as well as detailed descriptions of the experimental procedure, the data preprocessing, and the qualitative and quantitative evaluation of the NMR spectra are given in Appendix A.

2.2.3 Definition of Input and Output Data

Supervised machine-learning approaches, such as SVC, aim at learning how to map inputs $\boldsymbol{x}$ to outputs $\boldsymbol{y}$ by training to a labeled set of input-output pairs [47]. Two types of outputs were used in the present chapter. They are labeled here with $\boldsymbol{y}$ for the group-identification method and with $\boldsymbol{Y}$ for the group-assignment method. Hence, two data sets $\mathcal{D}_{\mathrm{I}}^{\mathrm{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ and $\mathcal{D}_{\mathrm{A}}^{\mathrm{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^N$ were used to train and evaluate the group-identification method and the group-assignment method, respectively. Both data sets consist of $N = 985$ input-output pairs and were derived from the considered pure component NMR spectra as described in the following.

The inputs $\boldsymbol{x}$ are the same in both data sets; their definition follows directly from the previously described binning of the ^{1}H and ^{13}C NMR spectra described above, cf. Section 2.2.1. Each ^{1}H and ^{13}C NMR spectrum was thereby translated into a bit vector of length $S^{^1\mathrm{H}}$ or $S^{^{13}\mathrm{C}}$, respectively, as follows: starting from 0 ppm in the spectrum and the first entry of the bit vector, the entry of the bit vector was set to 1 if at least one peak was observed in the respective section of the spectrum. This conversely means that each zero-value entry represents a section in the spectrum in which no peak is present. The order in which the information from the respective sections is translated into a bit vector has no influence on the solution of the optimization problem; however, the order chosen here is obviously intuitive. For each component i, the bit vector representing its ^{1}H NMR spectrum was appended to the bit vector representing its ^{13}C NMR spectrum. Hence, for all 985 considered pure components i, a bit vector $\boldsymbol{x}_i$ of length $S^{^{13}\mathrm{C}} + S^{^1\mathrm{H}}$ was obtained.

The output for the group-identification method, which aims at indicating whether specific structural groups are present in the sample or not, consists of one bit vector $\boldsymbol{y}_i$ of length $G = 13$ (since 13 structural groups are distinguished here, cf. Table 1) for each pure component i. $\boldsymbol{y}_i$ contains the information which structural groups are present in component i according to the group-division scheme used in this chapter, cf. Table A.5 in Appendix A. Following the order of the structural groups in Table 1, the entries of

$\boldsymbol{y}_i$ were set to 1 if the respective structural group is part of component i, and set to 0 otherwise. The order of the structural groups, as the one of the sections in the NMR spectra, can be chosen arbitrarily without influencing the solution of the optimization problem.

The output for the group-assignment method, which aims at identifying and assigning structural groups to peaks in the ^{13}C NMR spectrum, consists of a bit matrix $\boldsymbol{Y_i}$ of dimension $S^{^{13}\mathrm{C}} \times G$ for each component i. $\boldsymbol{Y}_i(s,g) = 1$ indicates that in section s of the ^{13}C NMR spectrum of component i at least one peak induced by the structural group g is observed. Conversely, $\boldsymbol{Y}_i(s,g) = 0$ indicates that no peak caused by structural group g is observed in section s of the ^{13}C NMR spectrum of component i.

The data sets $\mathcal{D}_{\mathrm{I}}^{\mathrm{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ and $\mathcal{D}_{\mathrm{A}}^{\mathrm{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^{N}$ were divided in three subsets in this chapter: a training set to train the methods, a validation set to optimize the hyperparameters of the methods, and a test set to evaluate their predictive performances. Details on this procedure are given in the following sections.

Besides the two data sets from pure component NMR spectra, two additional data sets $\mathcal{D}_{\mathrm{I}}^{\mathrm{mix}} = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1}^{M}$ and $\mathcal{D}_{\mathrm{A}}^{\mathrm{mix}} = \{(\boldsymbol{x}_j, \boldsymbol{Y}_j)\}_{j=1}^{M}$ consisting of $M = 3$ input-output pairs derived from mixture NMR spectra were used for testing the group-identification method and the group-assignment method, respectively, but not for training or validation of the methods. Hence, $\mathcal{D}_{\mathrm{I}}^{\mathrm{mix}} = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=1}^{M}$ and $\mathcal{D}_{\mathrm{A}}^{\mathrm{mix}} = \{(\boldsymbol{x}_j, \boldsymbol{Y}_j)\}_{j=1}^{M}$ were considered as test sets only. The definition of the input and output vectors/matrices for the studied mixtures was analogous to the procedure for the pure components, except that the positions of the peaks were extracted from the experimentally recorded NMR spectra. A detailed description of the procedure is given in Appendix A.

2.2.4 Support Vector Classification

SVC methods are by default binary margin classifiers that find a decision hyperplane that separates training data points i into two distinguishable classes $y_i \in \{-1, 1\}$ by considering their input $\boldsymbol{x}_i$. The output y_i of an SVC method is constituted by the class of i.

Real-world data are often not completely separable with a linear classifier. In this case, soft-margin SVC can be applied that allows outliers, i.e., data points in the training set that are incorrectly labeled by the algorithm. These outliers are penalized. To find the decision hyperplane in linear soft-margin SVC, the following constrained convex optimization problem has to be solved [47]:

$$
\begin{aligned}
&\min_{\boldsymbol{w},b,\boldsymbol{\xi}} && \frac{1}{2}\cdot\|\boldsymbol{w}\|^2 + C\cdot\sum_{i=1}^{N}\xi_i && && \text{(1a)}\\
&\text{subject to} && y_i(\boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i \geq 0 && i = 1,...,N && \text{(1b)}\\
& && \xi_i \geq 0 && i = 1,...,N && \text{(1c)}
\end{aligned}
$$

where the decision hyperplane is defined by the normal weight vector $\boldsymbol{w}$ and the offset b. Minimizing $\frac{1}{2}\|\boldsymbol{w}\|^2$ in the objective function (Eq. (1a)) represents maximizing the margin between the classes as in the hard-margin case, while the term $C\cdot\sum_{i=1}^{N}\xi_i$ penalizes the outliers. ξ_i is called slack variable and represents the severity of each violation by an outlier. C is a hyperparameter that weights the two contributions to the objective function. Small values of C lead to a classifier that only slightly penalizes outliers in the training data set, whereas large values of C yield a classifier that is very restrictive to incorrectly classified training data points.

After training the algorithm, i.e., solving the optimization problem in Eq. (1) for the training set, the decision function $d(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\cdot\boldsymbol{x}_i + b$, is used for the classification of unseen data points i: if $d(\boldsymbol{x}_i) > 0$, i is assigned to the positive class ($y_i = 1$), if $d(\boldsymbol{x}_i) < 0$, i is assigned to the negative class ($y_i = -1$) [47].

The concept of SVC can also be applied to non-linear classification problems, i.e., if the data are intrinsically not separable by a linear classifier, by using the so-called kernel trick [47]. Examples for linearly separable and non-linearly separable data sets are shown in Figure A.6 in Appendix A. The input data are thereby transferred into a feature space via a transformation function $\phi(\boldsymbol{x})$ [47]. Instead of an explicit transformation of the input data into the feature space, a kernel function $K(\boldsymbol{x}_i,\boldsymbol{x}_j)$ [47] is applied on pairs of input vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j) \tag{2}$$

A kernel basically measures the pairwise similarity between two input vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the feature space (after the transformation). In this chapter, the so-called radial basis function (RBF) kernel was used, which calculates the dot product from Eq. (2) as:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\gamma\cdot\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \tag{3}$$

where γ is a hyperparameter. If the Kernel trick is applied, the decision function for data point i is given by: $d(\boldsymbol{x}_i) = \boldsymbol{w}^{\mathrm{T}}\cdot\phi(\boldsymbol{x}_i) + b$.

2.2.5 Multi-label Classification

The above-described soft-margin SVC method is only applicable for binary classifications. In this chapter, 13 different classes (structural groups) are to be distinguished. Furthermore, in general, multiple classes (different structural groups) can be assigned to each data point (pure component or mixture). Therefore, the so-called one-vs-rest strategy [51], in which multiple binary SVC methods (units) are jointly considered, was applied here.

For the group-identification method, a separate (binary) SVC unit for each structural group (13 in total) was trained to indicate whether the respective structural group is present in the considered component or mixture, or not.

For the group-assignment method, the one-vs-rest strategy was applied separately *for each section* of the ^{13}C NMR spectrum. Hence, for each section, a set of (binary) SVC units was trained to indicate the presence or absence of each structural group. The data set for each set of SVC units (for a specific section of the ^{13}C NMR spectrum) thereby only contains the input-output pairs obtained from those pure component NMR spectra that exhibit a peak in the respective section. Hence, in principle a total of $S^{^{13}\mathrm{C}} \times G$ binary SVC could be trained. However, only for those structural groups that can in principle, based on the training set, show peaks in the respective section of the ^{13}C NMR spectrum, an SVC unit was trained. I.e., the number of SVC units that were trained for each section varies. As an example, for the section “>201 ppm” in Figure 1, only two SVC units were trained, one for identifying CO$^{\mathrm{ald}}$ groups and one for identifying CO$^{\mathrm{ket}}$ groups, since for all other structural groups, no positive examples for this section were available in the training data set.

2.2.5.1 Classification Scores

Classification scores are used to evaluate a method’s capability to correctly assign classes to data points. In this chapter, the F_1 score was used to evaluate the performance of the group-identification method for predicting the correct structural groups g (classes) based on the NMR spectra of pure components (input). F_1 scores of 1 correspond to perfect predictions. For the group-assignment method, the evaluation was done for each section s of the ^{13}C NMR spectrum. The F_1 score is defined as the harmonic mean of precision P_g and recall R_g and was calculated for each structural group g:

$$F_{1,g} = 2 \cdot \frac{P_g \cdot R_g}{P_g + R_g} \tag{4}$$

The precision P_g for group g is defined as the ratio of the number of data points that are correctly labeled with group g (True Positive, TP_g) and the total number of data points that are labeled with g, i.e., the sum of TP_g and the number of data points that are incorrectly labeled with group g (False Positive, FP_g):

$$P_g = \frac{TP_g}{TP_g + FP_g} \tag{5}$$

The recall R_g for group g, on the other hand, is defined as the ratio of TP_g and the sum of TP_g and the number of data points that are incorrectly not labeled with group g (False Negative, FN_g):

$$R_g = \frac{TP_g}{TP_g + FN_g} \tag{6}$$

The F_1 scores for all groups g are summarized in the F_1^{macro} score, which is defined as the mean of the individual $F_{1,g}$ scores [52]. Only the $F_{1,g}$ scores for which meaningful values could be obtained, i.e., the scores for those groups g for which a classifier was trained, were thereby considered.

2.2.5.2 Training, Hyperparameter Optimization, Evaluation with Pure Component Data

As described above, the data sets $\mathcal{D}_{\text{I}}^{\text{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ and $\mathcal{D}_{\text{A}}^{\text{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^{N}$ derived from the pure component NMR spectra were divided in three subsets in this chapter: a training set, a validation set, and a test set. The training set was used for fitting the parameters of the SVC method. The validation set was used for hyperparameter optimization to prevent overfitting on the training set. The test set was used for testing the predictive performance of the trained classifiers on unseen data points.

Group-Identification Method

For the group-identification method, nested cross-validation [53–55] with an outer loop with ten folds and an inner loop with five folds was applied. Hence, in the outer loop, 10% of the data points in $\mathcal{D}_{\text{I}}^{\text{pure}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ were defined as test data and not used for training or hyperparameter optimization. The remaining 90% of the data, in turn, were split in 80% training and 20% validation data.

In the inner loop, the hyperparameters $S^{^1\text{H}}$ and $S^{^{13}\text{C}}$ were specified and the remaining hyperparameters C and γ were determined to maximize the average F_1^{macro} score on the validation data (averaged over the five folds in the inner loop) after training the method to the training data. After setting C and γ to the optimized values, the group-identification method was then trained to all data in the inner loop and applied to classify the test data, that had been withheld in the outer loop. This procedure was repeated ten times such that each data point served once as test data point, i.e., a predicted classification for each data point was obtained, which could then be compared to the true class label.

The whole procedure of nested cross-validation for the group-identification method was repeated for different combinations of $S^{^1\mathrm{H}}$ and $S^{^{13}\mathrm{C}}$, i.e., for different discretizations of the NMR spectra. The scanned grids for all hyperparameters are summarized in Table 2.

Table 2: Ranges in which the hyperparameters were optimized for the group-identification method. The grids for $S^{^1\mathrm{H}}$ and $S^{^{13}\mathrm{C}}$ were evenly distributed in a linear scale; the grids for C and γ were evenly distributed in a logarithmic scale. For C and γ, the procedure for the group-assignment method was the same, while $S^{^1\mathrm{H}}$ and $S^{^{13}\mathrm{C}}$ were not optimized but set to 14 and 23, respectively.

Hyperparameter	Type	Grid	No. of points
$S^{^1\mathrm{H}}$	Integer	10 – 20	11
$S^{^{13}\mathrm{C}}$	Integer	15 – 25	11
C	Float	10^{-3} – 10^{2}	30
γ	Float	10^{-5} – 10^{0}	30

Group-Assignment Method

For the group-assignment method in combination with the pure component data $\mathcal{D}_\mathrm{A}^\mathrm{pure} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^N$, the applied procedure of nested cross-validation was analog, except that $S^{^1\mathrm{H}} = 14$ and $S^{^{13}\mathrm{C}} = 23$ were used, as this combination worked well for the group-identification method, cf. Section 2.3. Furthermore, only five folds were used in both the inner and outer loop to split the data into training, validation, and test sets.

Computational Details

Both methods were implemented in Python 3.7 using scikit-learn 0.22 [55] on a Linux system (Linux Ubuntu, Intel (R) Xeon (R) 2.4 GHz). The implementation is based on the library libsvm [56]. Because the studied data set is very imbalanced, cf. Figure 1, the option of weighted SVC in scikit-learn [55, 56] was used to optimize the hyperparameter C. The data splits were performed randomly, but using the learning library scikit-multilearn [57] and the multi-label data stratification technique based on Ref. [58] to reduce imbalances between classes in the different folds, where applicable.

2.2.5.3 Application to Mixture Data

Both methods were also tested for the elucidation of structural groups in unknown mixtures. Therefore, the group-identification method and the group-assignment method were trained to the complete pure component data sets $\mathcal{D}_\mathrm{I}^\mathrm{pure} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$ or $\mathcal{D}_\mathrm{A}^\mathrm{pure} = \{(\boldsymbol{x}_i, \boldsymbol{Y}_i)\}_{i=1}^N$, respectively, using a five-fold cross-validation to optimize the hyperparameters C and γ on a logarithmic grid (150 x 150) with the same boundaries as stated in Table 2 while setting $S^{^1\mathrm{H}} = 14$ and $S^{^{13}\mathrm{C}} = 23$. Furthermore, in this case, an additional post-processing step was included to assure the physical consistency of the predictions. For the group-identification method, this post-processing step basically prevents that the method predicts more structural groups in a sample than different peaks are present in its ^{13}C NMR spectrum (it is supposed that each different structural group results in a distinct peak, which is a reasonable assumption in ^{13}C NMR spectroscopy) and that the method predicts at least one structural group if at least one peak is present in the ^{13}C NMR spectrum. If required, the predictions were corrected in the post-processing step: if too many structural groups were predicted, the structural groups with the smallest positive decision function value were rejected as positive classes until physical consistency was achieved; if no structural group was predicted although a peak in the ^{13}C NMR spectrum was observed, the group with the highest decision function value was accepted as positive class. For the group-assignment method, these rules were applied for each section of the ^{13}C NMR spectrum. In Appendix A, a step-by-step example for the application of the group-assignment method to mixture spectra is presented, which includes the description of the application of the post-processing rules. This is the most sophisticated of the studied scenarios; a transfer to the application of the group-identification method to mixture spectra, or of both methods to pure component spectra is straightforward.

2.3 Results

2.3.1 Structural Group Analysis Based on Pure Component Spectra

2.3.1.1 Group-identification Method

The highest average F_1^{macro} *validation* score (averaged over the ten folds in the outer loop) of the group-identification method if applied to the considered pure component NMR spectra was obtained by dividing the ^{1}H NMR spectra and ^{13}C NMR spectra into 14 and 23 sections, respectively, i.e., for $S^{^1\text{H}} = 14$ and $S^{^{13}\text{C}} = 23$, as shown in Figure A.4 in Appendix A. However, Figure A.4 also demonstrates that the group-identification method is very robust regarding the binning of the NMR spectra within the studied grids: the results of the F_1^{macro} validation scores vary only between 0.86 and 0.90; the lowest scores were found for the smallest studied values for $S^{^{13}\text{C}}$ ($S^{^{13}\text{C}} = 15$).

In the following, the predictive performance of the group-identification method for pure component NMR spectra is evaluated using $S^{^1\text{H}} = 14$ and $S^{^{13}\text{C}} = 23$ only and study the F_1 *test* scores for the considered structural groups, which indicate how reliably a specific group g is identified by the method. Figure 2 shows the F_1 test scores for all considered structural groups g as a function of the number of pure components N_g in the data set, in which the respective structural group g is present.

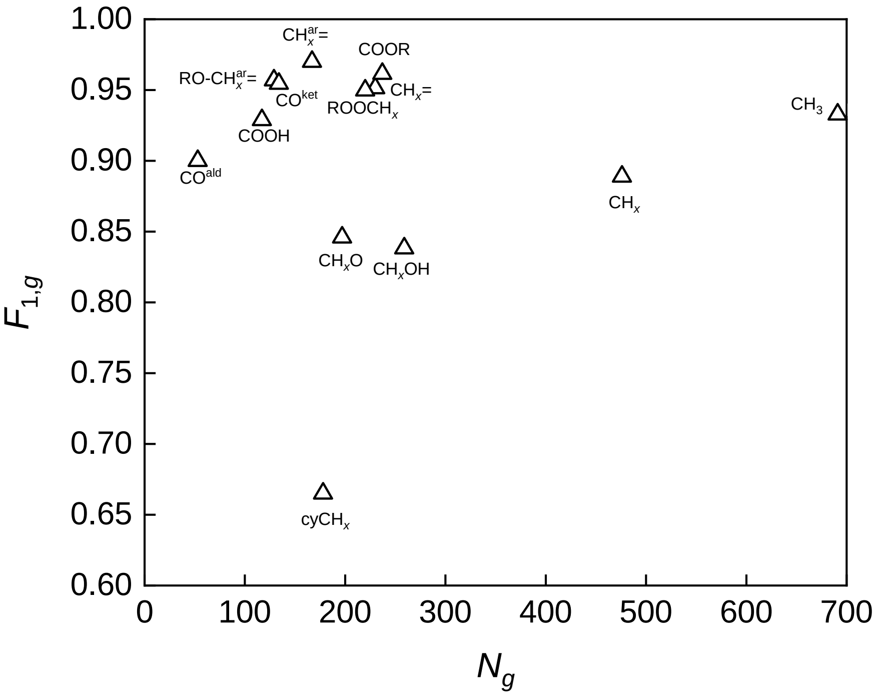


Figure 2: F_1 test scores of the group-identification method for the considered structural groups g and the NMR spectra of the considered pure components. N_g denotes the number of pure components in the data set in which the respective structural group g is present.

Obviously, the F_1 score for a specific structural group g does not depend on the number of pure components N_g in the data set in which this group is present. For instance, the CH_3 group is the most frequent group in the data set and present in nearly 700 of the 985 considered components, but does not exhibit a significantly greater F_1 score than most other structural groups; the same holds for the second most frequent group in the data set CH_x. This is considered as a strength of the method, proving that it is not overfitted or biased towards frequent structural groups, but also shows good results for less frequent groups.

Overall, the group-identification method shows a good performance in identifying structural groups from pure component NMR spectra with F_1 scores greater than 0.8 for all groups except for $cyCH_x$. The different performances for different structural groups can mostly be attributed to the presence or absence of characteristic peaks for the respective groups in the NMR spectra. For instance, the $cyCH_x$ group shows peaks over a wide range of chemical shifts in the ^{13}C NMR spectrum, overlapping with ranges of mainly five other structural groups (CH_3, CH_x, CH_xOH, CH_xO, and $ROOCH_x$), cf. Figure 1.

Also, the protons of the cyCH$_x$ group can yield peaks at a great variety of chemical shifts in the ^{1}H NMR spectrum ranging from very low chemical shifts (<0.4 ppm) for three-membered rings (e.g., cyclopropane) [33] to higher chemical shifts that overlap with those of the protons in, e.g., CH$_x$ and CH$_3$ groups. Hence, the identification of cyCH$_x$ groups based only on a ^{13}C and a ^{1}H NMR spectrum is extremely difficult, in some cases basically impossible. Taking into account these difficulties, the obtained scores of the group-identification method for the cyCH$_x$ group are remarkable. The performance can be further improved by including information from additional NMR experiments, such as recording distortionless enhancement by polarization transfer (DEPT) NMR spectra, as demonstrated in Chapter 3. DEPT spectra allow the distinction of carbon nuclei based on their multiplicity, which yields distinguishable peaks for, e.g., CH$_3$ and cyCH$_x$ groups. Another option to react to this finding would be to refrain from trying to identify the cyCH$_x$ group from the ^{13}C and ^{1}H spectra, and to reduce the number of groups, e.g., by merging the CH$_x$ and the cyCH$_x$ group.

For other structural groups that also show substantial overlapping of their peaks' characteristic ranges with those of other groups in the ^{13}C NMR spectrum, such as CH$_x$= and CH$_x^{\mathrm{ar}}$= groups, cf. Figure 1, excellent scores were obtained. Both groups show characteristic peaks in the ^{1}H NMR spectrum, which presumably helps the method to identify them. Another example is the successful distinction of COOH and COOR groups, which is difficult based on ^{13}C NMR alone, as the respective characteristic ranges strongly overlap, but which is especially important regarding the prediction of reactivities and fluid properties in general [59]. The high scores for these groups can also be attributed to the consideration of ^{1}H NMR spectra in addition to ^{13}C NMR here, since the related peaks in ^{1}H NMR spectroscopy are often distinct [25, 33]. These findings underpin the importance of combining information from different NMR experiments as realized in this chapter and in contrast to previous studies. It is noted that most of the NMR spectra considered here were recorded with the solvent CDCl$_3$; only the spectra of 92 components were recorded in other solvents. As the chemical shift of a structural group is sensitive to the used solvent, one could assume issues with the reliability of the group-identification method for samples recorded in other solvents than CDCl$_3$. Therefore, in Figure A.7 in Appendix A, the influence of the solvent on the performance of the group-identification method is studied. The results indicate that the method is quite robust regarding different solvents. However, in future work, additional information on the solvent can be incorporated as input data in the approach, e.g., by considering an additional bit vector that encodes the solvent.

2.3.1.2 Group-assignment Method

Figure 3 shows the performance of the group-assignment method when applied to the considered pure component spectra. A separate F_1 test score for each combination of structural group g and section s in the ^{13}C NMR spectrum is considered, termed $F_{1,g}^{s}$, and thereby evaluated how reliably the method identifies a structural group *and* assigns it to the correct peak (section) in the ^{13}C NMR spectrum. Only structural group/section combinations with at least three positive examples in the data set were thereby taken into account, cf. shaded areas in Figure 3, since the results for combinations with less positive examples are guaranteed to be inconclusive due to the procedure of nested cross-validation where the data set was divided in three subsets. The higher the number of positive examples per combination, the higher the probability to obtain meaningful test scores. To obtain interpretable scores, at least one positive example in each subset is required. However, it is not guaranteed that with three positive examples per structural group/^{13}C NMR section combination, the respective test scores in Figure 3 are always reasonable, since a homogeneous data splitting was not enforced, i.e., all three positive examples for a specific combination could be part of the same set, even if the multilabel data stratification technique that was used tries to avoid this [58]. For more details, cf. Section 2.2.5.2.

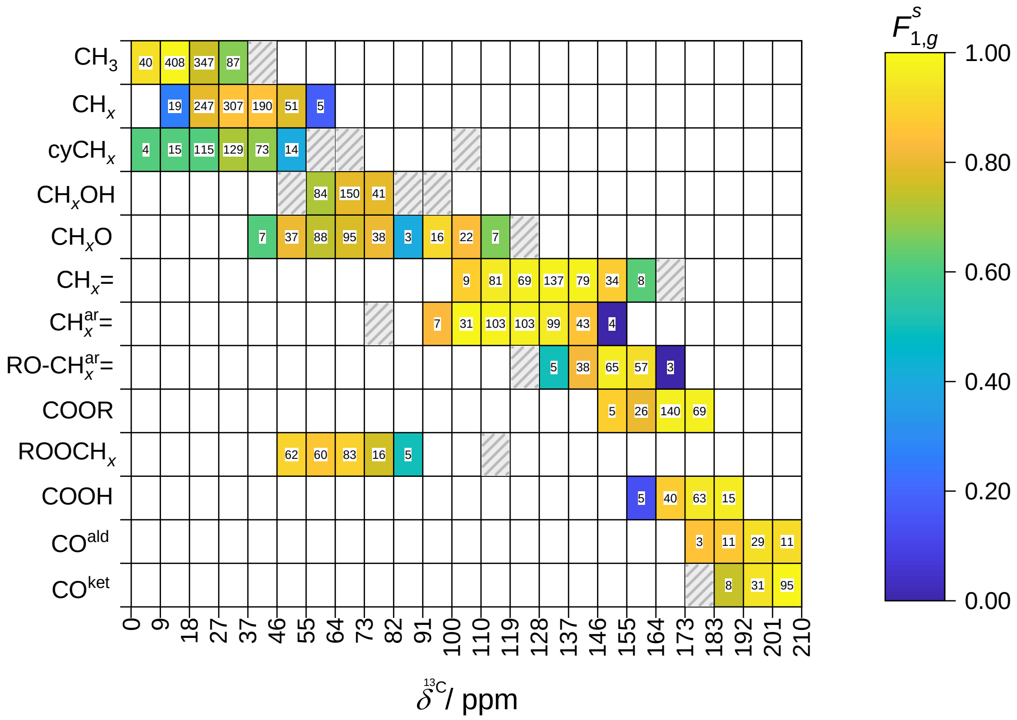


Figure 3: F_1 test scores (indicated by the color code) of the group-assignment method for the considered structural groups g and the sections s of the ^{13}C NMR spectra of the considered pure components. The numbers inside the cells indicate the number of components N_g^s in the data set that contain the respective structural group g (row) inducing a peak in the respective section s of the ^{13}C NMR spectrum (column). Shaded areas indicate structural group/section combinations with less than three positive examples in the data set, which were not taken into account.

The color code in Figure 3 represents the F_1 scores whereas the numbers inside the cells denote the number of positive examples for each structural group/section combination in the data set, i.e., the number of components N_g^s in the data set that contain the respective structural group g (row), which in turn induces a peak in the respective section s of the ^{13}C NMR spectrum (column). For most structural group/section combinations, high F_1 scores are observed, i.e., the method successfully identifies and assigns structural groups to the correct sections in the ^{13}C NMR spectrum. Even structural groups whose peaks in the ^{13}C NMR spectrum strongly overlap, such as CH$_x$= and CH$_x^{ar}$= groups, can be distinguished with high success rate. Low F_1 scores are observed for structural group/section combinations with very few positive examples in the data set. This observation is consistent with the expectations, as it is difficult for a purely data-driven approach, as applied here, to learn to properly classify from rare examples. In Figures A.8-A.10 in Appendix A, the performance of the group-assignment method for the application to NMR spectra of larger, more complex molecules (up to 306 g mol^{-1}) is shown. All these components, and the respective spectra, were not part of the training data set;

please note that the methods were trained to data for pure components up to 160 g mol^{-1} only, cf. above. It is emphasized that the applicability of the methods introduced in the present chapter is not restricted to small molecules as a result of the consideration of structural groups instead of components; the molar mass of a component should not have a significant impact on the performance of the methods.

The group-assignment method can of course also be used for simply identifying structural groups in an unknown sample (without assigning the groups to peaks in the ^{13}C NMR spectrum). Figure 4 shows the performance of the group-assignment method applied to the pure component spectra for this purpose by considering the F_1 test scores for the sole identification of the groups g irrespective of the section in which the respective peaks are observed.

In analogy to Figure 2, these scores are termed $F_{1,g}$ in Figure 4. $F_{1,g}$ for a specific group g was calculated by summing all $F_{1,g}^s$ test scores of group g and weighing with the number of positive examples of the respective group g in section s (N_g^s), cf. Figure 3:

$$F_{1,g} = \frac{\sum\limits_{s=1}^{S^{13}\mathrm{C}} F_{1,g}^s \cdot N_g^s}{\sum\limits_{s=1}^{S^{13}\mathrm{C}} N_g^s} \tag{7}$$

Only the $F_{1,g}^s$ test scores for those group g / section s combinations for which at least three positive examples were available in the data set were thereby considered to obtain interpretable results, as discussed above.

For most structural groups, the group-assignment method shows slightly worse performance than the group-identification method, cf. Figure 2. This is hardly surprising, since the “building blocks” of the group-assignment method (a separate classifier for each combination of structural group and section in the ^{13}C NMR spectrum) were trained on much smaller data sets than the “building blocks” of the group-identification method (a separate classifier for each structural group), as explained in Section 2.2.5. Hence, if only the identification of structural groups in an unknown sample is required, the group-identification method should be employed. In future work, both methods could be combined by using the output of the group-identification method, i.e., the obtained information on the presence of structural groups in an unknown sample, as additional input for the group-assignment method to further improve its performance.

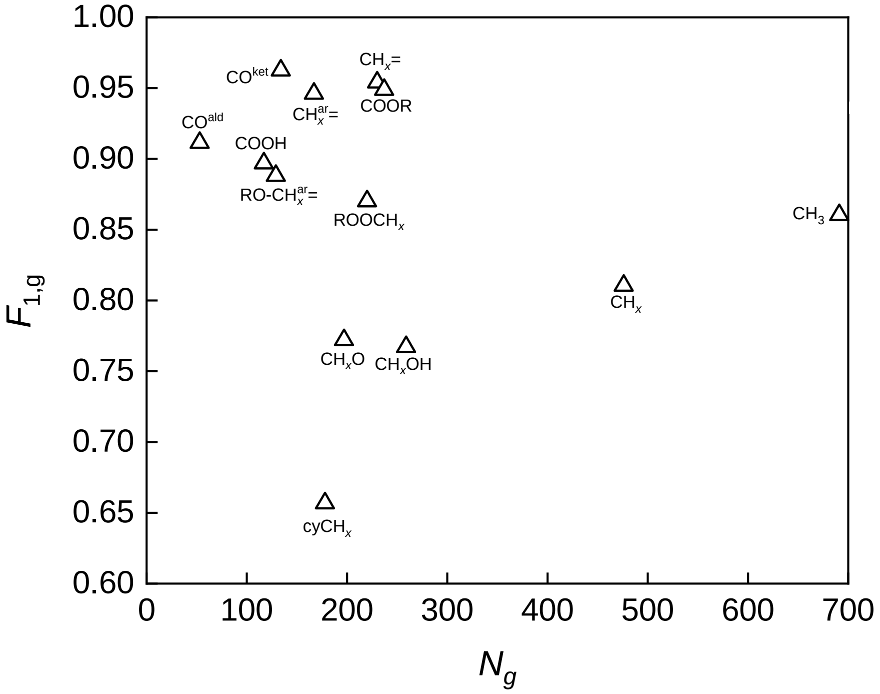


Figure 4: F_1 test scores of the group-assignment method for the sole identification of the considered structural groups g and the NMR spectra of the considered pure components. For each group g, the respective score was calculated by summing all $F_{1,g}^s$ test scores for group g (and all relevant sections s) and weighing with the number of positive examples for the respective structural group / section combination, cf. Figure 3.

2.3.2 Structural Group Analysis in Mixture Spectra

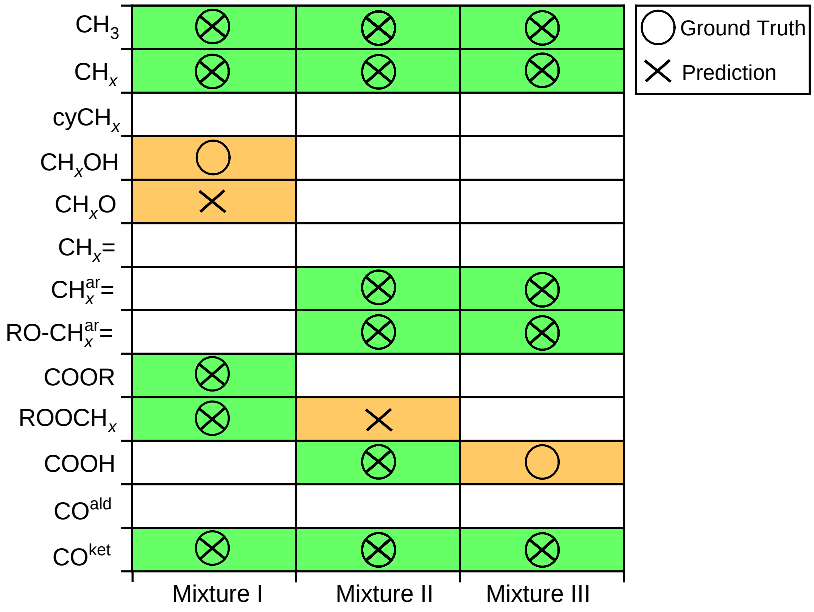
In the following, the evaluation for the application of both proposed methods for the elucidation of structural groups based on NMR spectra of mixtures is shown. Table 3 summarizes information on the mixtures studied here. It is emphasized again that both methods were trained to pure component data only, cf. Section 2.2.5.3. Mixture I includes only components that are also part of the pure component data set. Mixture II contains two such components (acetone and 4-hydroxybenzoic acid) and one component that was not part of the pure component data set (ibuprofen). Mixture III contains two components that were not part of the pure component data set (ibuprofen and tert-butylhydroquinone).

Table 3: Overview of the mixtures studied in this chapter. In the last column, all structural groups in the respective mixture according to the group division scheme used here are given (the order of the groups follows the order in Table 1 and does not correspond to the individual components).

Mixture	Components i	x_i / mol mol^{-1}	Structural groups
I	1-Propanol	0.29	CH_3, CH_x, CH_xOH,
	Acetone	0.51	COOR, $ROOCH_x$,
	Ethyl acetate	0.20	CO^{ket}
II	Ibuprofen	0.03	CH_3, CH_x, $CH_x^{ar}{=}$,
	Acetone	0.93	$RO{-}CH_x^{ar}{=}$, COOH,
	4-Hydroxybenzoic acid	0.04	CO^{ket}
III	Ibuprofen	0.03	CH_3, CH_x, $CH_x^{ar}{=}$,
	Acetone	0.93	$RO{-}CH_x^{ar}{=}$, COOH,
	tert-Butylhydroquinone	0.04	CO^{ket}

2.3.2.1 Group-identification Method

Figure 5 shows the results for the application of the group-identification method to the ^{1}H and ^{13}C NMR spectra of the studied mixtures: the structural groups that are predicted by the method for each mixture are compared to the groups that are in fact present in the mixtures. Overall, excellent agreement of the predictions with the ground truth is observed: most groups are correctly predicted by the group-identification method. In mixture I, the only error is that the method predicts an ether group (CH_xO) instead of an alcohol group (CH_xOH). This can presumably be explained by the massive shifting of the peak of the hydroxyl proton of the CH_xOH group in the ^{1}H NMR spectrum, cf. Figure A.1 in Appendix A. Also for the other mixtures, very good predictions are obtained with the group-identification method; in mixture II, it only falsely predicts the presence of a $ROOCH_x$ group, whereas it misses a COOH group in mixture III.



	Mixture I	Mixture II	Mixture III
CH_3	⊗	⊗	⊗
CH_x	⊗	⊗	⊗
$cyCH_x$			
CH_xOH	○		
CH_xO	×		
$CH_x=$			
$CH_x^{ar}=$		⊗	⊗
$RO\text{-}CH_x^{ar}=$		⊗	⊗
COOR	⊗		
$ROOCH_x$	⊗	×	
COOH		⊗	○
CO^{ald}			
CO^{ket}	⊗	⊗	⊗

Figure 5: Results of the application of the group-identification method to the NMR spectra of mixtures I - III, cf. Table 3. Green areas indicate correct predictions, orange areas indicate mistakes.

The results demonstrate the broad applicability of the proposed group-identification method. After training the method to pure component data only, it gives excellent predictions also for mixtures. The results also show that it is not required that the mixture components are (as pure components) part of the training set of the method.

2.3.2.2 Group-assignment Method

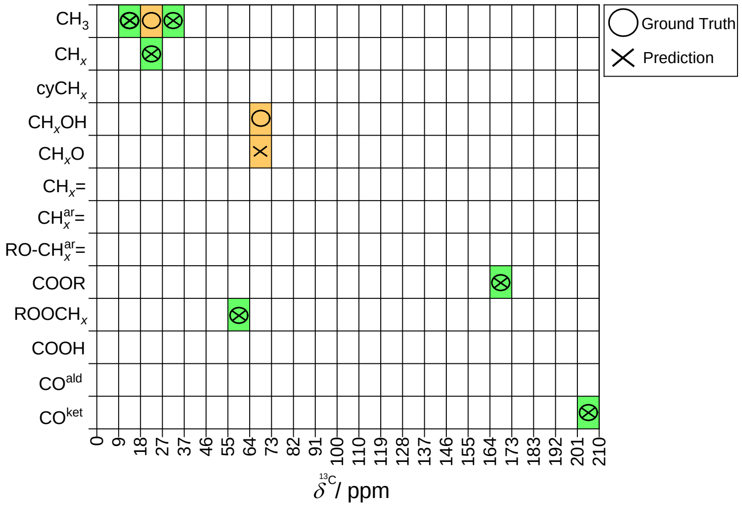
In Figure 6, the results for the application of the group-assignment method to identify the structural groups in the mixtures I - III based on a ^{1}H and a ^{13}C NMR spectrum of each mixture and, additionally, to assign the identified structural groups to sections of the ^{13}C NMR spectrum are shown.

In Figure 6 (a), the results for mixture I are shown. Almost all structural groups in this mixture are correctly identified by the method and assigned to the correct sections in the ^{13}C NMR spectrum. Just like the group-identification method, the group-assignment method thereby confuses the peaks of an alcohol group (CH_xOH) with those of an ether group (CH_xO). Furthermore, one CH_3 group is overseen by the group-assignment method: all peaks in the respective section of the ^{13}C NMR spectrum are assigned to CH_x groups.

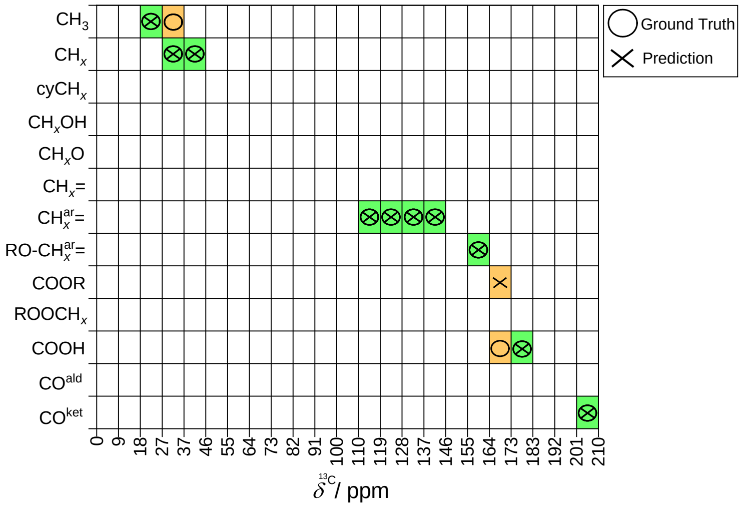
In Figure 6 (b), the respective results for mixture II are shown. The method confuses a carboxyl group (COOH) with an ester group (COOR), presumably due to the peak of the proton of the CH_x group at 3.71 ppm in the ^{1}H NMR spectrum, which falls in the characteristic region for the proton(s) of the ROOCH$_x$ group, cf. Figure A.2 in Appendix A. The identification of carboxyl groups in mixtures is particularly challenging, since the peaks of carboxyl protons can significantly shift depending on the composition. The situation could be improved by integrating information on the uncertainty of the positions of the peaks in further work. Except for a CH$_3$ group that is missed by the method, all remaining groups are correctly identified and assigned to sections in the ^{13}C NMR spectrum.

Figure 6 (c) shows the results for mixture III. Although two components of this mixture (ibuprofen and tert-butylhydroquinone) were not considered in the training step, i.e., their pure component spectra were not part of the training data, the overall performance of the group-assignment method for analyzing the structural groups is very good. As in mixture I and II, the method oversees a CH$_3$ group and also has some difficulties to correctly identify and assign the different aromatic carbons in the mixture, cf. regions ranging from 137-155 ppm in the ^{13}C NMR spectrum. This could be caused by the upfield shifting of the aromatic protons of tert-butylhydroquinone close to the characteristic regions of olefinic protons, cf. Figure A.3 in Appendix A. In Figure A.11 in Appendix A, additional results of the application of the group-assignment method to further mixtures, including aqueous mixtures, are shown.

(a) Mixture I



(b) Mixture II



(c) Mixture III

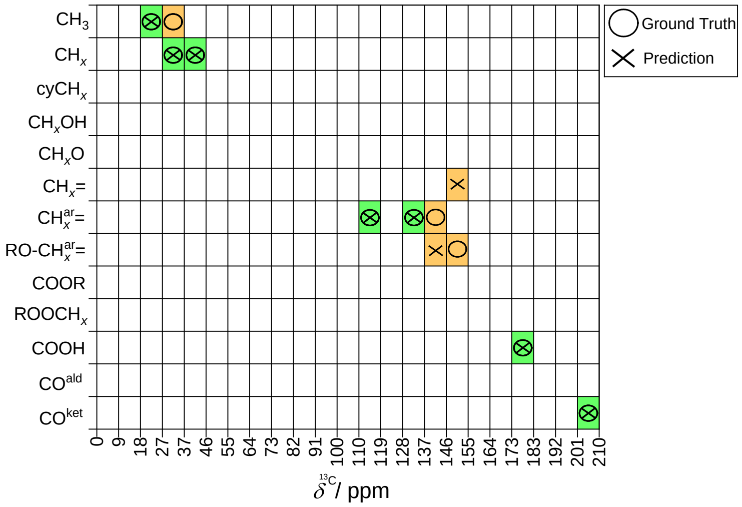
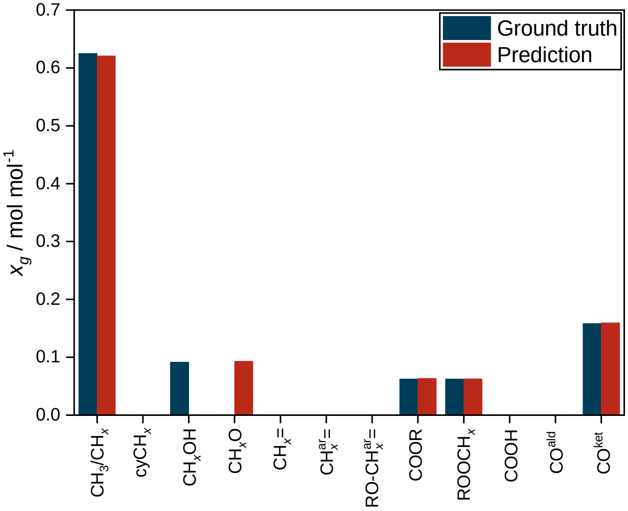


Figure 6: Prediction of structural groups in mixtures I - III (cf. Table 3) and assignment to sections in the ^{13}C NMR spectrum with the group-assignment method. Green areas indicate correct predictions, orange areas indicate mistakes.

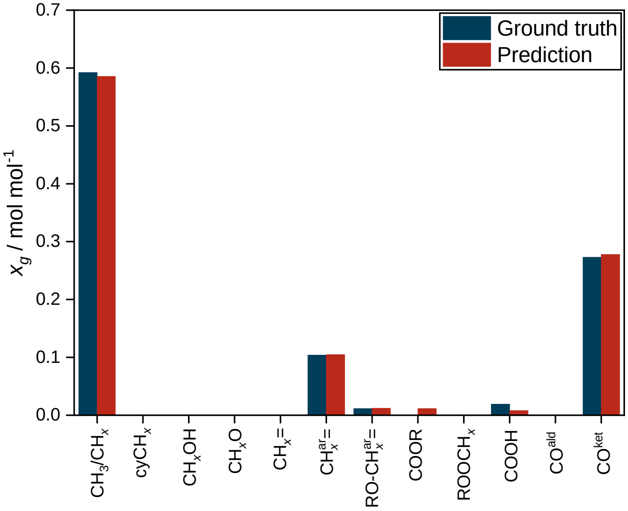
Overall, excellent predictions with the group-assignment method if applied to NMR spectra of mixtures are obtained. These predictions can not only serve as qualitative analysis of unknown mixtures, i.e., to answer the question which structural groups are present in a mixture (which can be even better handled by the presented group-identification method), but also constitute the basis for a quantitative elucidation of unknown mixtures, i.e., to answer the question which structural groups are present in a mixture *in which concentration*. Based on the results of the group-assignment method, quantitative predictions can be obtained by section-wise integration of the ^{13}C NMR spectrum of the studied mixture. Figure 7 shows such quantitative predictions in the form of group mole fractions for mixtures I - III obtained with the group-assignment method and compares them to the true group composition of the mixtures, which is available from sample preparation, cf. Section A.1.2 in Appendix A for details.

In Figure 7(a), the results for mixture I are shown. For most structural groups, excellent predictions for the group mole fractions are achieved. The group-assignment method only confuses the CH_xOH group with the CH_xO group, cf. Figure 6(a), which is also reflected in the quantitative results here. No predictions for the individual mole fractions of the CH_3 group and the CH_x group are reported here, as the current version of the group-assignment method is not able to differentiate between peaks within the same section of the NMR spectrum. Therefore, only the sum of both group mole fractions (CH_3/CH_x) is represented. A comparison between the prediction for this summed group mole fraction and the ground truth shows high accuracy.

(a) Mixture I



(b) Mixture II



(c) Mixture III

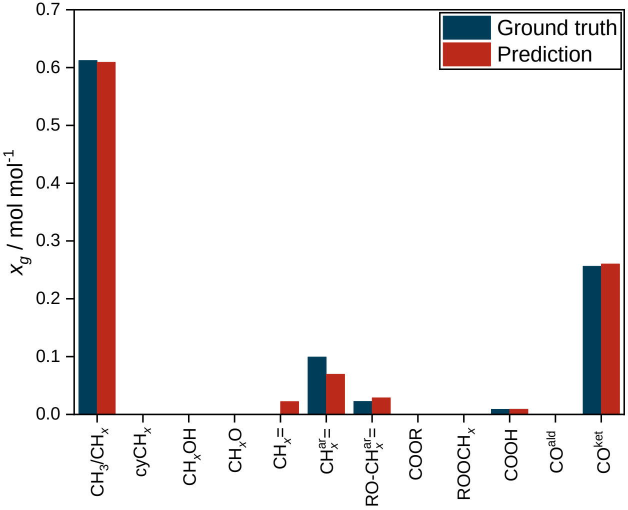


Figure 7: Prediction of mole fractions of structural groups in mixtures I - III (cf. Table 3) obtained from the results of the group-assignment method and a section-wise integration of the ^{13}C NMR spectrum, cf. Section A.1.2 in Appendix A.

The knowledge of the sum of the mole fractions of CH_3 and CH_x groups is sufficient for many purposes, as one can argue that these two groups are very similar not only pertaining to their peaks in ^{1}H and ^{13}C NMR spectroscopy, but also regarding their influence on the properties of components (or mixtures) they are part of. It is likely that for estimating the properties of components (or mixtures) with thermodynamic group-contribution methods, the individual mole fractions of CH_3 and CH_x groups are of minor importance whereas a good prediction of the sum of the two group mole fractions is crucial. As an example, consider modified UNIFAC (Dortmund) [50], which is one of the most successful thermodynamic group-contribution methods. In modified UNIFAC (Dortmund) CH_3 and CH_x belong to the same "main-group", which means that the same interaction parameters are used for both groups. Figure 7(b) shows the quantitative predictions for mixture II. Excellent agreement between the predictions and the ground truth was observed for all structural groups, only a small number of COOR groups is mistakenly predicted. In Figure 7(c), the results for mixture III are shown, which are similar to those for mixture II. The mole fractions of most groups are predicted very well, except for a small amount of CH_x= groups that is mistakenly indicated by the group-assignment method. Overall, excellent agreement between prediction and ground truth was found.

The obtained quantitative group-specific characterizations can, e.g., directly be used for the prediction of fluid properties of unknown mixtures with thermodynamic group-contribution methods, as recently demonstrated in Refs. [59–63] and Chapters 5 and 6. It is emphasized that the introduced methods in this chapter do not aim at outperforming human spectroscopists. In contrast, the methods are supposed to support chemists or engineers in the evaluation of NMR spectra; especially if a great many of spectra has to be evaluated such an automated approach is a huge asset.

2.4 Conclusions

In this chapter, two methods have been introduced to elucidate the structural groups in unknown pure components and mixtures based on one ^{1}H NMR spectrum and one ^{13}C NMR spectrum of the considered sample and using the concept of support vector classification. The first method, called group-identification method, yields qualitative information on the presence or absence of 13 different structural groups in an unknown sample. The second method, called group-assignment method, paves the way for a quantitative analysis of unknown samples by identifying the same 13 different structural groups in the sample and additionally assigning the identified groups to peaks in the ^{13}C NMR spectrum of the sample; information on the concentration of the structural groups can then be obtained in a simple manner by integration of the peaks in the ^{13}C NMR spectrum, as demonstrated in the present chapter. Both methods were trained to ^{1}H and ^{13}C NMR spectra of nearly 1000 pure components for which the required information, i.e., the constituent structural groups, were available. The predictive performance of both methods for pure components was evaluated using nested cross-validation and high classification scores for unseen data points, which were withheld during training and validation, were obtained. Additionally, the applicability of both methods to mixtures was demonstrated using three ternary mixtures as examples that were experimentally studied in this chapter. Although both methods were trained to pure component data only, high success rates were found for mixtures as well. The proposed methods pave the way for an automated reliable qualitative and quantitative characterization of unknown components or mixtures, which occur frequently in practice. The present chapter demonstrates that based on conventional 1D NMR spectra, which can easily be obtained in practice, reasonable predictions for the structural groups in unknown samples can be achieved. The obtained group-specific characterizations can, e.g., be used for the prediction of fluid properties of unknown mixtures with thermodynamic group-contribution methods, which is especially simple with the approach here as the definition of structural groups here is closely related to that of modified UNIFAC (Dortmund), indicating a great influence of the proposed methods on process design and optimization in chemical engineering. Chapter 3 demonstrates that the developed methodology can easily be extended to incorporate additional information, e.g., from ^{13}C DEPT NMR spectra.

3 NMR Fingerprinting based on ^{1}H, ^{13}C, and ^{13}C DEPT NMR

3.1 Introduction

Chapter 2 has built the foundations for automatically identifying structural groups from NMR spectra using the machine-learning concept of support vector classification. In this chapter, the NMR fingerprinting concept is substantially extended to also include information from ^{13}C DEPT NMR spectra in addition to ^{1}H and ^{13}C NMR spectra of the sample. The ^{13}C DEPT NMR spectra thereby provide direct information on the substitution degree of the carbon atoms and are used here to differentiate, e.g., between 'CH_3' and 'CH_2' groups in the sample.

Furthermore, in this chapter, SMARTS are incorporated in the NMR fingerprinting framework. SMARTS [64] is an acronym for SMILES [65] arbitrary target specification strings, which are based on the simplified molecular-input line-entry system (SMILES), which, in turn, is a system to represent components by simple text strings. SMARTS are used here for a rigorous definition of the distinguished structural groups in the NMR fingerprinting framework and enable a fully automated training workflow. SMARTS also provide great flexibility in defining the groups so that the approach can be straightforwardly tailored to a specific application.

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

The new method was trained on spectra of 2839 pure components from two data banks, namely, the NMRShiftDB [66] and the Biological Magnetic Resonance Data Bank (BMRB) [67], and was tested using rigorous nested cross-validation (CV). Furthermore, a data augmentation technique was developed to substantially increase the training data set and address the fact that no comprehensive data bank for NMR spectra of mixtures is available. Finally, several test mixtures were analyzed with an 80 MHz benchtop NMR spectrometer, and the recorded spectra were used as input for testing the new approach in a practical setting.

The developed method here is made available on an interactive website (`https://nmr-fingerprinting.de`), which enables testing and applying the method presented here through a graphical user interface (GUI) without the need for any program installation. The user supplies the spectral information and gets the corresponding NMR fingerprint.

3.2 Overview of the Workflow

Figure 8 visualizes the workflow of the method developed in this chapter, which can be used via the website (`https://nmr-fingerprinting.de`). The method's goal is identifying structural groups in an unknown sample and assigning them to peaks in the ^{13}C NMR spectrum of the sample. In the present version, the NMR fingerprinting method can differentiate thirteen structural groups, which are summarized in Table 4. The groups are the same as in Ref. [68], cf. also Chapter 2, but SMARTS strings for each group are also provided here.

For using the method, a ^{1}H NMR spectrum, a ^{13}C NMR spectrum, and ^{13}C DEPT 90/135 NMR spectra of the studied sample are required. Specifically, the chemical shift of all peaks in the ^{1}H NMR (except the peaks of labile protons) and in the ^{13}C NMR spectrum as well as the substitution degree of each carbon atom, which can be automatically determined considering the signs of the peaks in the DEPT 90/135 NMR spectra [69], are needed for defining the input of the method.

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

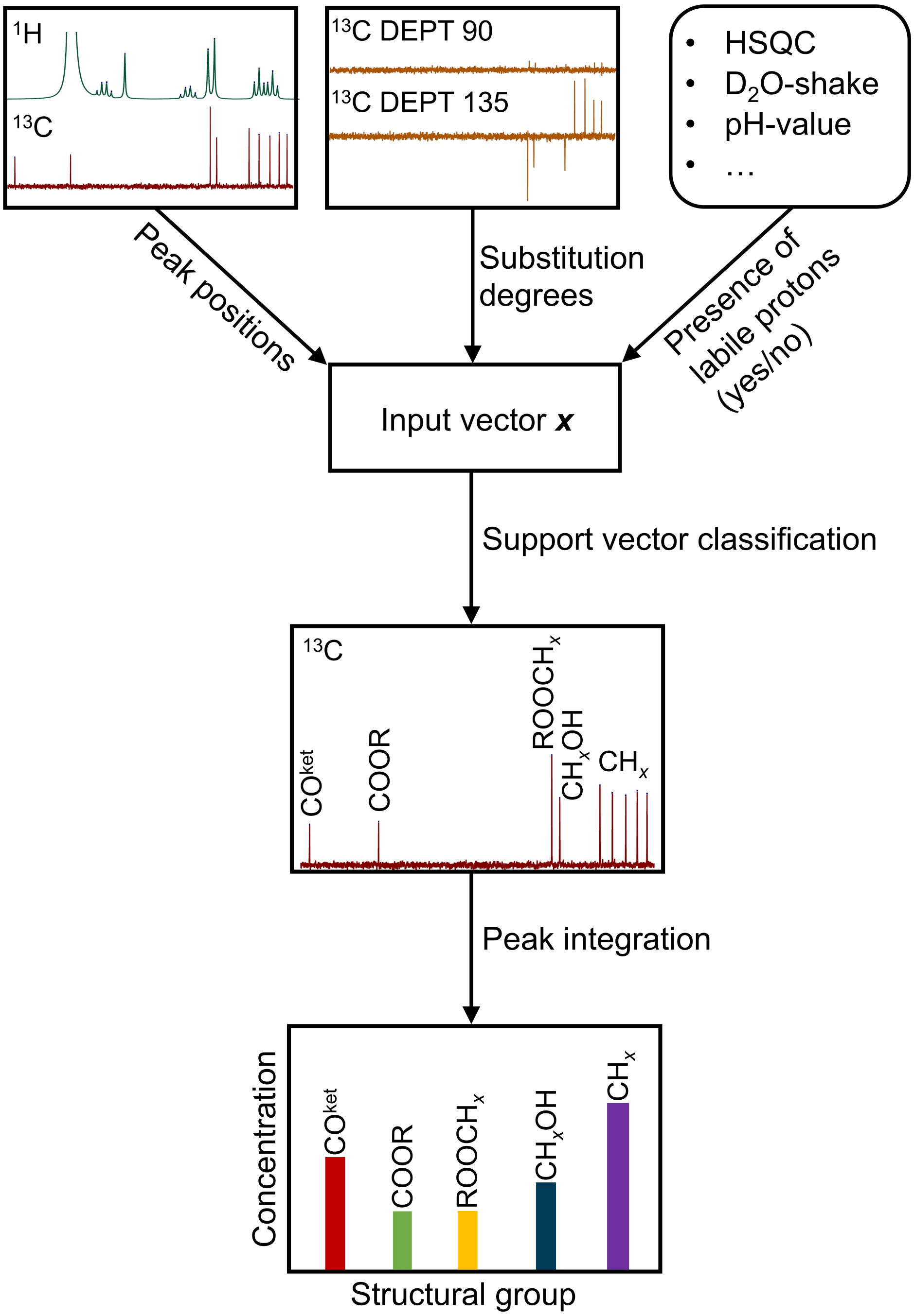


Figure 8: Workflow of the developed NMR fingerprinting method using an SVC for predicting structural groups based on spectral information from ^{1}H, ^{13}C, and ^{13}C DEPT NMR spectra as well as using binary information about the presence or absence of labile protons in the sample. In Appendix B, a variant of the method that does not require information about labile protons is presented.

Table 4: Structural groups distinguished in the present chapter and the respective SMARTS strings for their representation. Each group contains exactly one carbon atom and x is the number of protons directly bonded to the carbon atom.

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

The spectral information, as well as the information on the presence or absence of labile protons, are used for defining the input vector $\boldsymbol{x}_i$ of a sample i for the SVC, whereby the NMR spectra are in general equidistantly binned and the peaks are assigned to the respective sections in the spectra, cf. Section 3.3.1 for details. The information about the presence or absence of labile protons in the sample can come from different sources. In many cases, one will know if, e.g., carboxylic acids or alcohols are present in the sample. However, even if this information is not available a priori, there are multiple ways to its determination. For instance, labile protons can often be recognized from the ^{1}H NMR spectrum since they usually show characteristic, very broad peaks. Alternatively, HSQC NMR spectra can be used, where peaks of labile protons show no correlation with any carbon nucleus in the sample. Another established method is the so-called D_2O-shake, i.e., the addition of a small amount of deuterated water to the sample, which will cause the peaks stemming from labile protons to vanish or at least significantly change their position and/or amplitude if compared to a ^{1}H NMR spectrum before adding D_2O. Also, simple pH measurements can detect labile protons, specifically the ones from carboxylic acids.

The SVC method was trained on labeled data for 2839 pure components to predict the structural groups summarized in Table 4, cf. Section 3.3.3 and Appendix B for details. Finally, by integrating the peaks in the ^{13}C NMR spectrum, the concentrations of all identified structural groups can be directly obtained.

3.3 Data and Methods

3.3.1 Generation of Input and Output Data for Training

For training the method, only experimental data for pure components were used and retrieved from two NMR data banks, cf. Section 3.3.2. Furthermore, synthetic mixture data were generated to augment the training set by combining the processed experimental pure-component data in the nested CV, cf. below for details. In general, for training the method and evaluating it on the pure-component data, all peaks of labile protons were removed from the ^{1}H NMR spectra.

For generating the input data for the training, the substitution degree of each carbon atom in each pure component, i.e., primary (P), secondary (S), tertiary (T), or quaternary (Q), was determined automatically based on the structure of the component using RDKit [70], cf. Appendix B for details. Furthermore, the ^{13}C and ^{1}H NMR spectra of all pure components from the data set were divided into equidistant discrete sections. For the ^{13}C NMR spectra, the chemical shift range from 0-210 ppm was considered and divided into $S^{^{13}\mathrm{C}} = 21$ sections of 10 ppm width. For the ^{1}H NMR spectra, the chemical shift range from 0-10 ppm was considered and divided into $S^{^{1}\mathrm{H}} = 20$ sections of 0.5 ppm width.

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

The general goal of the developed method is to identify and assign the structural groups g from Table 4 to sections s in the ^{13}C NMR spectrum of a component i. Hence, as the output of the method, the matrix $\boldsymbol{Y}_i$ was defined for each component i from the data set, which is a bit matrix of dimension $S^{^{13}\mathrm{C}}$ x G, where $\boldsymbol{Y}_i(s, g) = 1$ indicates the presence of at least one peak induced by structural group g in section s and G is the total number of distinguished groups ($G = 13$, cf. Table 4). The output matrix for each component was generated automatically using the respective SMARTS strings, cf. Table 4, and the RDKit package [70], cf. Appendix B for details.

3.3.2 Collection of Pure-component NMR Data

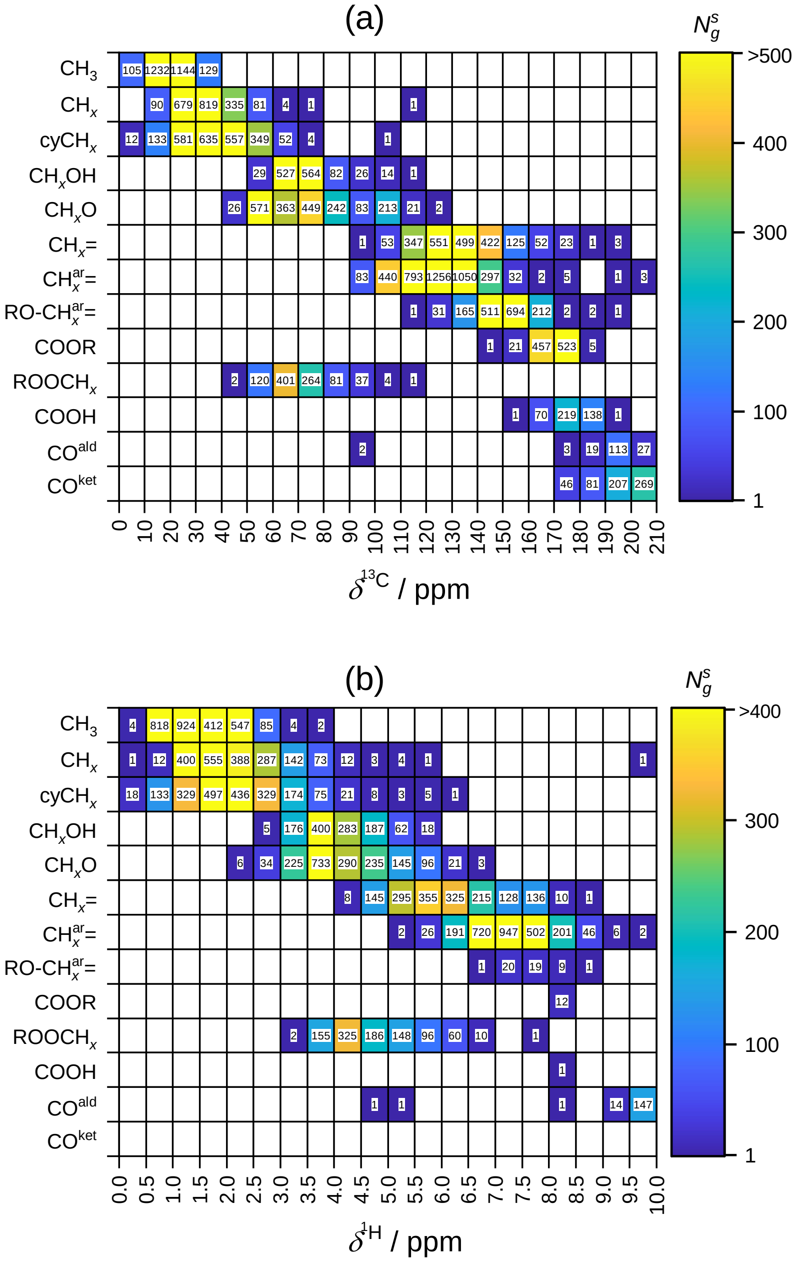
Raw NMR spectra of 2839 pure components were adopted from the BMRB [67] data bank and the NMRShiftDB [66] data bank and used for training and evaluation of the developed method, whereby spectra from the NMRShiftDB were preferred if data for a given component were available in both data banks. However, not all components and respective spectra reported in these NMR data banks were used. Therefore, besides removing erroneous spectra (cf. Appendix B for details), the following criteria had to be met:

1. Both a ^{13}C *and* a ^{1}H spectrum of the component are available.
2. The component is composed only of carbon (C), hydrogen (H), and/or oxygen (O).
3. The component can unambiguously be segmented into structural groups from the list in Table 4.

The first restriction could be relaxed in future work by the augmentation of incomplete data sets by predicted NMR spectra, e.g., using hierarchical organization of spherical environments (HOSE) methods [35], density-functional theory (DFT) calculations [34], or ML approaches [71]. Also, considering additional elements will be engaging in future work but will require additional analytical data, e.g., from other NMR experiments or other analytical techniques. The third restriction directly depends on the availability of NMR spectra of components containing the group of interest to ensure meaningful training of the method.

For all ^{13}C NMR spectra and ^{1}H NMR spectra from the data set, the chemical shifts of all peaks were extracted with only one exception: the peaks in the ^{1}H NMR spectra originating from protons directly bonded to oxygen. The reason for this is as follows: such protons are often labile due to exchange with other protons in the sample and, depending on the conditions during acquisition (like temperature and composition of the sample), their position in the NMR spectrum can vary strongly [69], which makes their position less informative for the proposed method. More details on the processing of the data are given in Appendix B.

Figure 9 (a) gives an overview of the data set containing the input and output data of the considered 2839 pure components, that is called $\mathcal{D}^{\text{pure}}$ in the following. Figure 9 (a) thereby indicates the frequency of the 13 distinguished structural groups in the considered components and also in which sections of the ^{13}C NMR spectrum the respective peaks appear. The assignment of groups to sections in the spectra is not unique: peaks of at least two different structural groups are found in each section. Furthermore, there is a substantial imbalance regarding the frequency of structural groups in the data set and the frequency of peaks in the different sections of the NMR spectrum. Figure B.1 in Appendix B shows that by taking into account the information on the different substitution degrees, multiple assignments of groups to a single section are substantially reduced.



(a)

	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180	180–190	190–200	200–210
$\mathrm{CH_3}$	105	1232	1144	129																	
CH_x		90	679	819	335	81	4	1				1									
cyCH_x	12	133	581	635	557	349	52	4			1										
$\mathrm{CH}_x\mathrm{OH}$						29	527	564	82	26	14	1									
$\mathrm{CH}_x\mathrm{O}$					26	571	363	449	242	83	213	21	2								
CH_x=										1	53	347	551	499	422	125	52	23	1	3	
$\mathrm{CH}_x^{\mathrm{ar}}$=										83	440	793	1256	1050	297	32	2	5		1	3
RO-$\mathrm{CH}_x^{\mathrm{ar}}$=												1	31	165	511	694	212	2	2	1	
COOR															1	21	457	523	5		
ROOCH_x					2	120	401	264	81	37	4	1									
COOH																1	70	219	138	1	
$\mathrm{CO^{ald}}$										2								3	19	113	27
$\mathrm{CO^{ket}}$																		46	81	207	269

(b)

	0.0–0.5	0.5–1.0	1.0–1.5	1.5–2.0	2.0–2.5	2.5–3.0	3.0–3.5	3.5–4.0	4.0–4.5	4.5–5.0	5.0–5.5	5.5–6.0	6.0–6.5	6.5–7.0	7.0–7.5	7.5–8.0	8.0–8.5	8.5–9.0	9.0–9.5	9.5–10.0
$\mathrm{CH_3}$	4	818	924	412	547	85	4	2												
CH_x	1	12	400	555	388	287	142	73	12	3	4	1								1
cyCH_x	18	133	329	497	436	329	174	75	21	8	3	5	1							
$\mathrm{CH}_x\mathrm{OH}$						5	176	400	283	187	62	18								
$\mathrm{CH}_x\mathrm{O}$					6	34	225	733	290	235	145	96	21	3						
CH_x=									8	145	295	355	325	215	128	136	10	1		
$\mathrm{CH}_x^{\mathrm{ar}}$=											2	26	191	720	947	502	201	46	6	2
RO-$\mathrm{CH}_x^{\mathrm{ar}}$=														1	20	19	9	1		
COOR																	12			
ROOCH_x							2	155	325	186	148	96	60	10		1				
COOH																	1			
$\mathrm{CO^{ald}}$										1	1						1		14	147
$\mathrm{CO^{ket}}$																				

Figure 9: Positions of the peaks of the 2839 pure components from the data set (a) in the ^{13}C NMR spectrum and (b) in the ^{1}H NMR spectrum. The color code and the numbers inside the cells denote N_g^s, which is the number of components in the data set that contain the structural group g (row) that induces a peak in the section s of the spectrum (column). White cells refer to $N_g^s = 0$.

Figure 9 (b) shows the respective information on the data set for the ^{1}H NMR spectrum. Similar to the ^{13}C NMR spectrum, the assignment of structural groups to the sections is not unique, so at least two different structural groups are found in each section. Furthermore, although there is a clear tendency for some groups to show peaks most frequently in some areas of the spectrum, all groups cover several chemical shift sections.

3.3.3 Training of Support Vector Classification

The core of the developed method is an SVC with a radial basis function (RBF) kernel, cf. Eq. (3), implemented in scikit-learn 1.2.0 [55], which was trained on the pure-component data set $\mathcal{D}^{\text{pure}}$.

During the development of the SVC method, $\mathcal{D}^{\text{pure}}$ was repeatedly divided into three subsets by a double-loop approach in the frame of a nested CV strategy [54]: a training set for fitting the model parameters, a validation set for optimizing the hyperparameters, and a test set for evaluating the predictive performance of the method. Details on how the data splits were performed are given in Appendix B. Nested CV was applied for each section of the ^{13}C NMR spectrum separately, whereby the data set for each section contained only those data points that induce a peak in the respective section of the ^{13}C NMR spectrum.

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

Optimizing the hyperparameters in the inner loop was carried out using a Bayesian optimization algorithm [72] to reduce computation time. Furthermore, only the scores for structural group/section combinations with at least ten positive examples in the data set were considered for the optimization in the inner loop, which was done to reduce the influence of atypical data points (outliers) on the developed method, cf. Appendix B for details.

Since an SVC is a priori only applicable to distinguish between two classes, i.e., an SVC is a priori a binary classifier, but multiple classes need to be assigned to each data point here, the so-called one-vs-rest strategy [51] was employed in this chapter.

For this purpose, multiple binary SVCs (called “units” in the following) were trained, one for predicting the presence or absence of each of the considered structural groups. The raw output of each binary SVC is its so-called decision function value, where the sign of the value indicates if a structural group is identified by the algorithm (positive value) or not (negative value), and the absolute value is proportional to the confidence of the method in the prediction [47]. Therefore, the decision which structural group g was assigned to a specific section s of the ^{13}C NMR spectrum was made by considering the values of the decision function d_g^s of all binary SVC units, whereby all groups with $d_g^s > 0$ were identified, cf. Appendix B for details. For applying the final SVC method to NMR spectra of mixtures, this procedure was slightly adapted as described in detail in Appendix B.

The predictive performance of the SVC was evaluated here using the so-called F_1-score $F_{1,g}$ for each structural group g, cf. Eq. (4).

Additionally, a “final” method was trained following the same procedure, but a CV with a split of the data set into 90% training and 10% validation data in each run, cf. Appendix B for details. The final method was not used to calculate the reported F_1-scores on the pure-component data but only applied to the experimentally studied mixtures in this chapter.

3.3.4 Experimental Methods

^{1}H NMR, ^{13}C NMR, and ^{13}C DEPT NMR spectra with pulse angles of 90 and 135 degrees were recorded of four test mixtures on an 80 MHz (proton frequency) benchtop NMR spectrometer (Spinsolve 80 Carbon Ultra) from Magritek. The experimental time for recording the ^{1}H NMR spectra of one sample was about 0.25 h. The experimental time for recording the ^{13}C NMR and ^{13}C DEPT NMR spectra for one sample was about 13 h. Note that the focus of the present chapter was not on time efficiency of the measurements; furthermore, obtaining a sufficient signal-to-noise ratio depends strongly on the concentration of the samples. Quantitative information was obtained by integrating the respective peaks in the quantitative ^{13}C NMR spectrum, if applicable. In addition, the substitution degree of each carbon atom was determined from the ^{13}C DEPT NMR spectra of the respective mixture. Details are given in Appendix B.

3.4 Results and Discussion

3.4.1 Prediction of Structural Groups from Pure-component Spectra

Figure 10 shows the results for the F_1 scores of the method for identifying structural groups and assigning the respective peaks to the different sections in the ^{13}C NMR spectrum of the pure-component data set of this chapter. White cells indicate group/section combinations with zero positive examples in the data set, and shaded cells indicate group/section combinations with one to nine positive examples in the data set. However, the sections with so little data were not considered in the evaluation for the reasons explained in the previous section.

Overall, high F_1 scores (> 0.8) were obtained for all group/section combinations with generally higher F_1 scores for group/section combinations with larger numbers of positive examples in the data set. In the region 0-40 ppm, excellent results were obtained: the $\mathrm{CH_3}$, $\mathrm{CH_x}$, and cyCH_x groups can be distinguished reliably. Excellent accuracy was also obtained in the region 190-210 ppm, where, e.g., the $\mathrm{CO^{ald}}$ group can be differentiated from the $\mathrm{CO^{ket}}$ and COOH groups, cf. Figure 9. Some other groups, such as $\mathrm{CH}_x\mathrm{OH}$, $\mathrm{CH}_x\mathrm{O}$, and ROOCH_x, are harder to distinguish but are still reasonably predicted. In Figure B.11 in Appendix B, additional results for a variant of the method that does not use information about the presence or absence of labile protons are shown, which are slightly worse, particularly for the COOH and $\mathrm{CH}_x\mathrm{OH}$ groups.

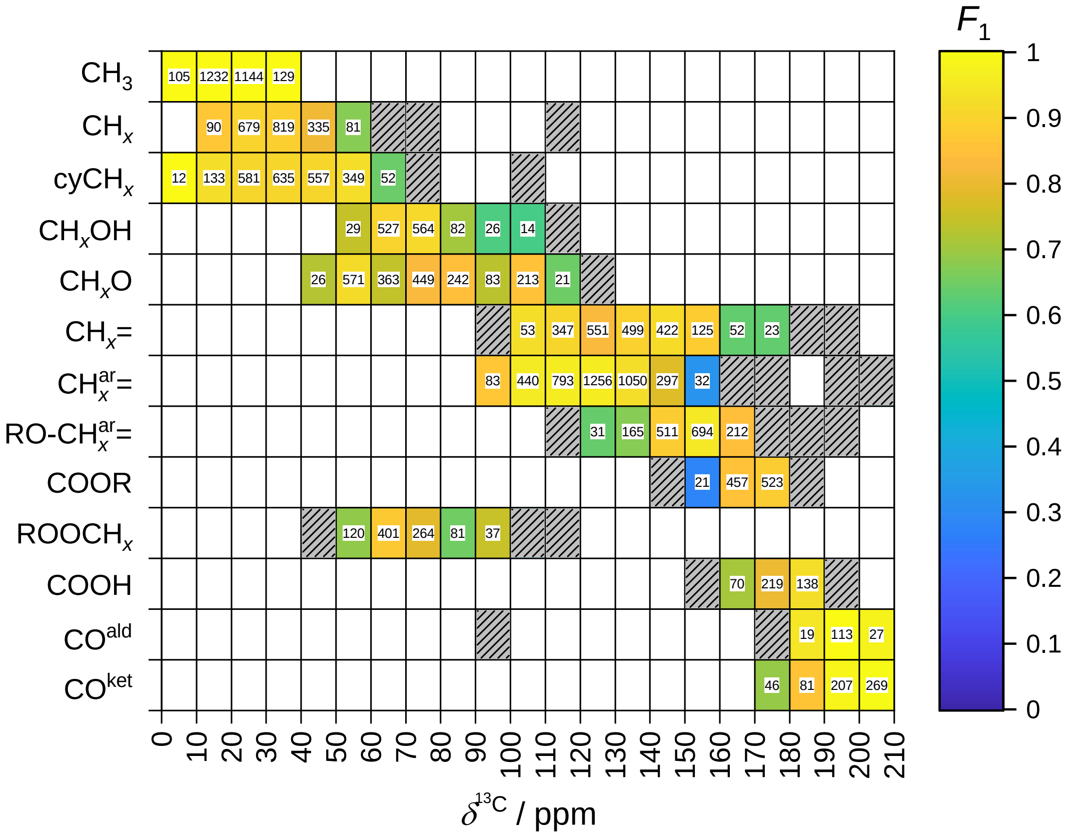


Figure 10: F_1 test scores (indicated by the color code) of the method for structural groups g and sections s of the ^{13}C NMR spectra of the pure components in the data set. The numbers inside the cells indicate the number of components N_g^s in the data set that contain the respective structural group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

3.4.2 Prediction of Structural Groups from Mixture Spectra

In the following, the results of applying the NMR fingerprinting method to four test mixtures are shown. Two aqueous and two organic mixtures were chosen for this purpose; the components used for preparing these mixtures were selected randomly but in a way that most of the structural groups (except CH_xO groups) covered by the developed method, cf. Table 4, were represented in at least one of the mixtures. Table 5 summarizes information on the mixtures studied as examples here.

Table 5: Overview of the test mixtures studied in this chapter.

Mixture	Components i	x_i / mol mol^{-1}
I	2-butanone	0.0136
	ethyl acetate	0.0145
	water	0.9719
II	cyclohexanone	0.0193
	malic acid	0.0198
	1-propanol	0.0197
	water	0.9412
III	1-octanol	0.9023
	tert-butylhydroquinone	0.0977
IV	acetone	0.1587
	butanal	0.1313
	oleic acid	0.7100

3.4.2.1 Results for Aqueous Mixtures

Since the signal-to-noise ratio of the aqueous mixtures was relatively low due to the high dilution, cf. Table 5, and the low magnetic field strength of the benchtop NMR spectrometer, a signal enhancement strategy, namely based on the nuclear overhauser effect (NOE), was used for obtaining the shown results; in consequence, no quantitative results were obtained here.

Figure 11 shows the results for applying the method to mixture I. All structural groups in the mixture were correctly identified and assigned to the respective peaks in the ^{13}C NMR spectrum. In Figure B.12 in Appendix B, results for the variant of the method without using prior information about the presence or absence of labile protons are presented; the same holds for the other studied mixtures discussed in the following.

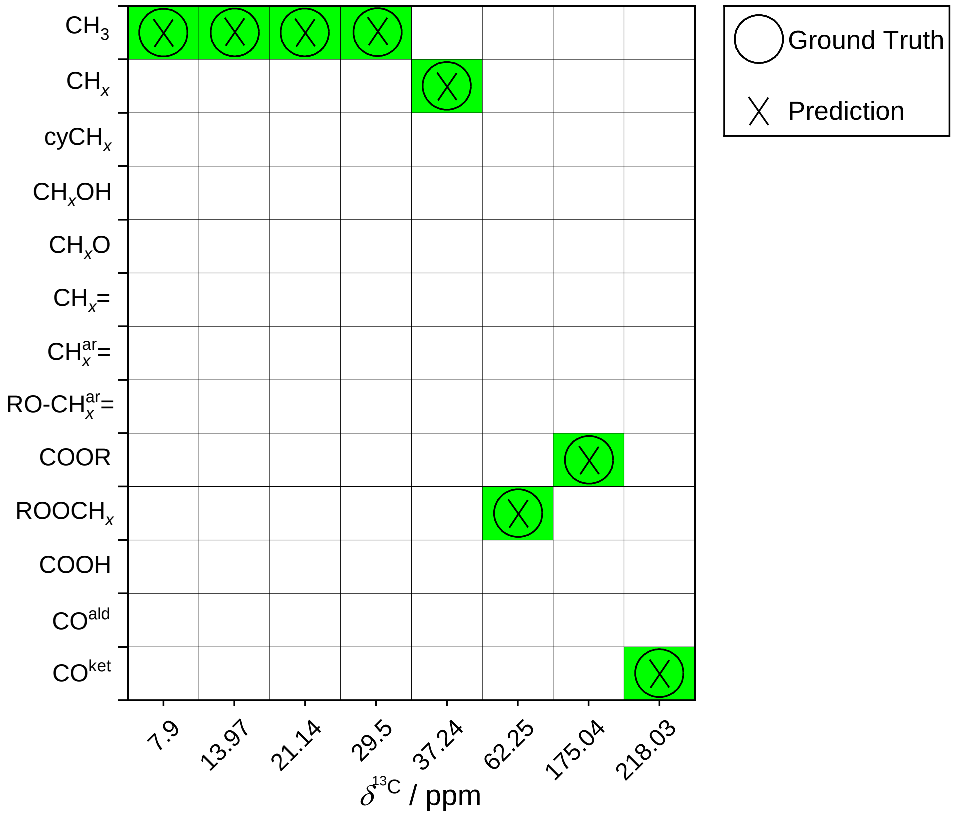


Figure 11: Results of the application of NMR fingerprinting to mixture I (cf. Table 5) for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

Figure 12 shows the results of applying the method to mixture II. In this case, the cyCH$_x$ groups were misinterpreted as CH$_x$ groups, which can be considered a minor error in many applications. All other groups are identified correctly.

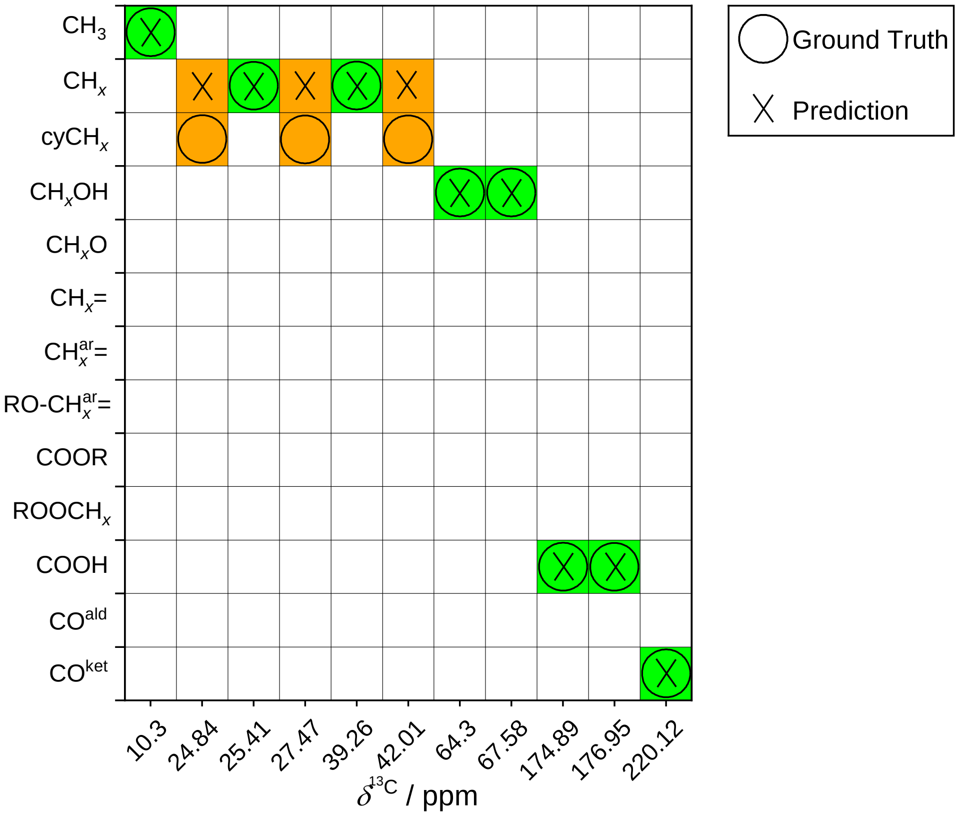


Figure 12: Results of the application of NMR fingerprinting to mixture II (cf. Table 5) for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions and orange color indicates mistakes. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

3.4.2.2 Results for Organic Mixtures

In Figure 13 (a), the results for identifying structural groups in mixture III are shown, which was accomplished correctly for all groups. Quantitative results, namely, the concentration of all identified structural groups in the form of group mole fractions x_g, are shown in Figure 13 (b). The differences between the predicted group mole fractions and the ground truth can mainly be attributed to the experimental error of the NMR spectra indicated by the signal-to-noise ratio.

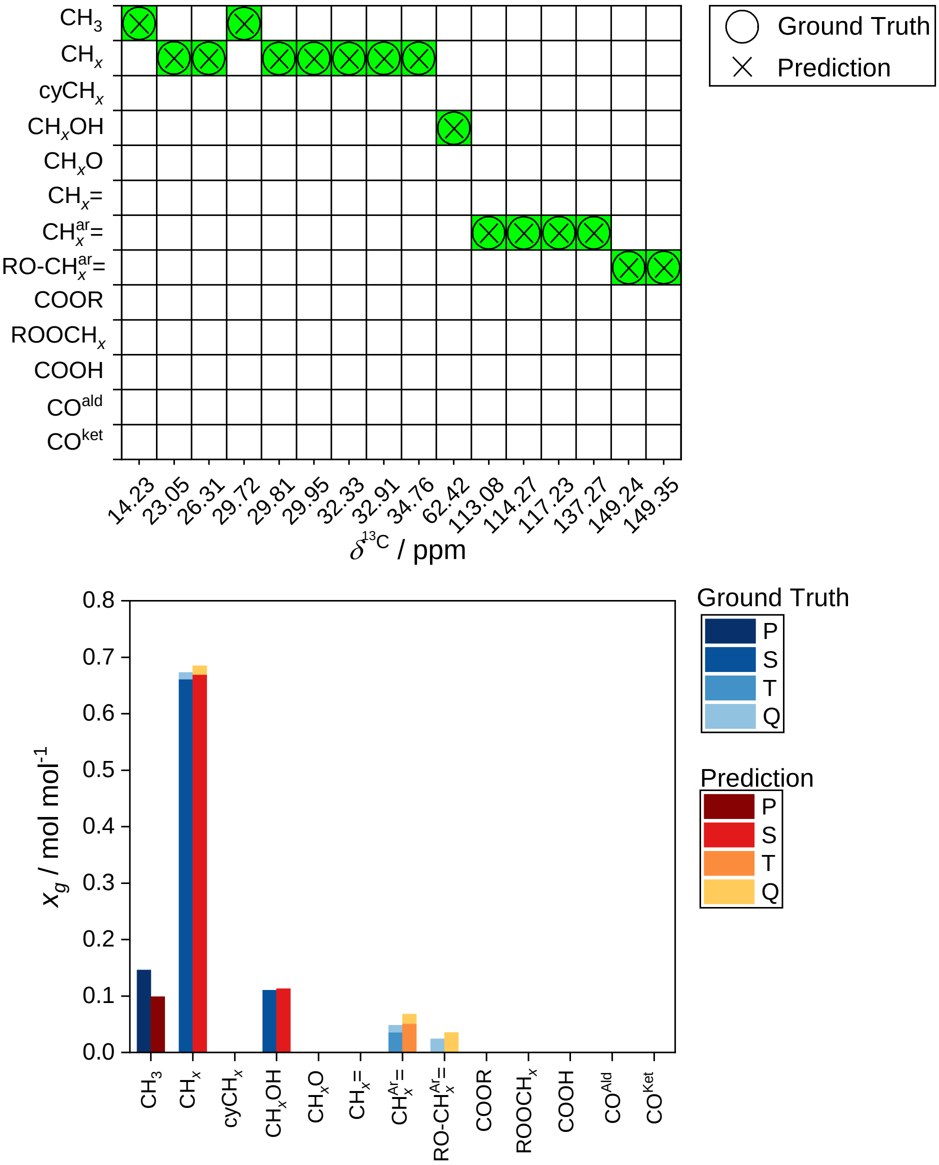


Figure 13: Results of the application of NMR fingerprinting to mixture III (cf. Table 5). (a): prediction of structural groups and assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated. (b): prediction of mole fractions of structural groups by integration of the peaks in the ^{13}C NMR spectrum; the results are shown individually for the different substitution degrees of the carbon atoms (P, S, T, Q).

Figure 14 (a) shows the respective results for the identification of structural groups in mixture IV. Again, all structural groups were predicted correctly. Also, the agreement between the predicted mole fractions of the structural groups and the ground truth is excellent, as shown in Figure 14 (b).

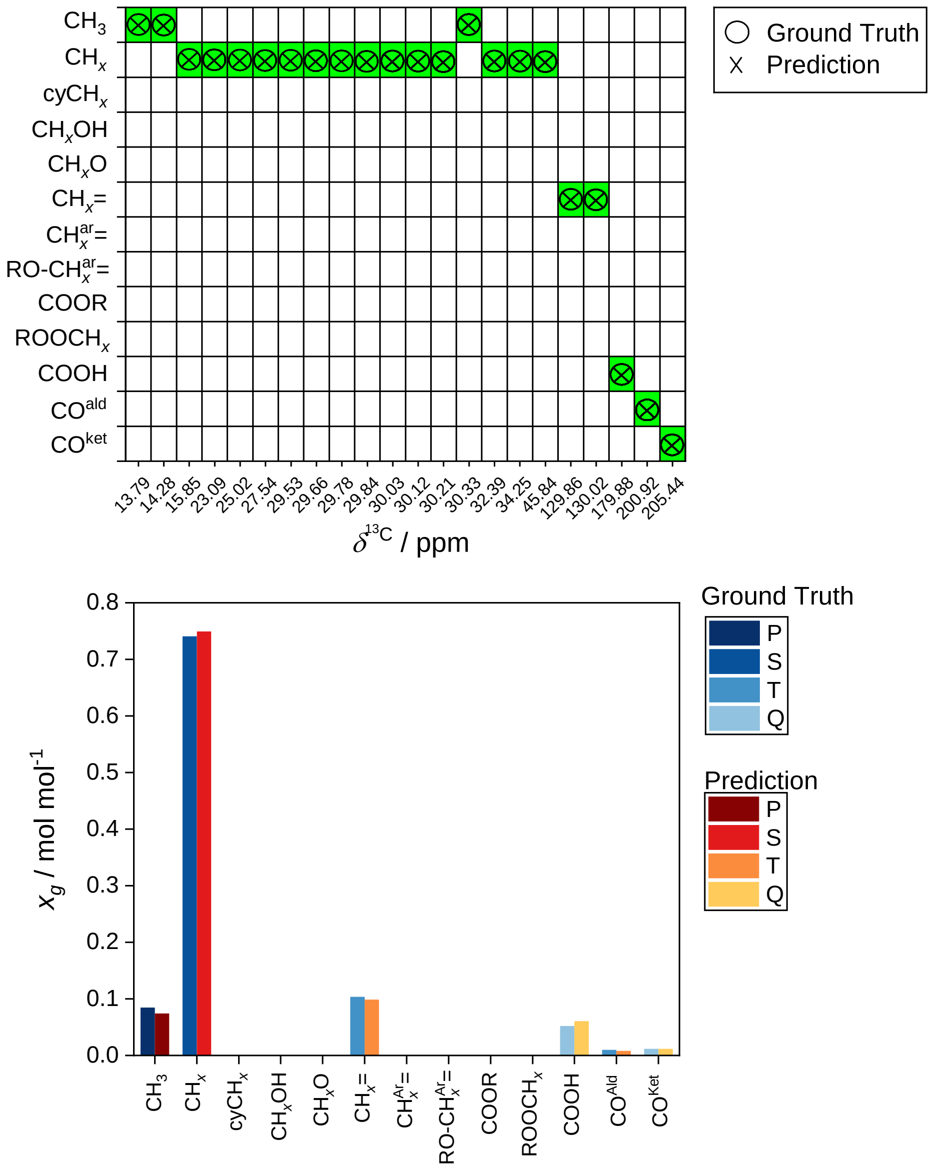


Figure 14: Results of the application of NMR fingerprinting to mixture IV (cf. Table 5). (a): prediction of structural groups and assignment to peaks in the ^{13}C NMR spectrum. Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated. (b): prediction of mole fractions of structural groups in mixture IV by integration of the peaks in the ^{13}C NMR spectrum; the results are shown individually for the different substitution degrees of the carbon atoms (P, S, T, Q).

3.5 Conclusions

In this chapter, a method for the group-specific qualitative and quantitative analysis of unknown samples based on standard NMR experiments (^{1}H NMR, ^{13}C NMR, and ^{13}C DEPT NMR) is presented (if only ^{1}H NMR and ^{13}C NMR spectra are available, the developed method from Chapter 2 can be used). The method is fully automated and requires no prior information on the samples and practically no expert knowledge, apart from that to carry out the experiments, which could also be automated. From the spectra, only the chemical shifts of the peaks, which can usually be picked in a semi-automatic way, are needed as input for the identification of the groups. If also peak areas are supplied, quantitative results on the group composition are provided. Furthermore, no expensive high-field NMR devices are needed; benchtop NMR devices are sufficient. In future work, it would be interesting to combine the NMR fingerprinting with an automated NMR acquisition: the experiments are automatically conducted, the spectra are automatically processed, and the interpretation of the data is presented.

The method is particularly interesting for applications in which information on the complete speciation of the sample is difficult to obtain, e.g., in biotechnology or refinery technology. In such applications, the method opens up new routes for process monitoring and process and quality control. Furthermore, the results from the method also provide a basis for quantitative physical modeling of mixtures with group-contribution methods – without having to know the complete speciation. The developed NMR fingerprinting method is made freely available for testing and application via an interactive website (`https://nmr-fingerprinting.de`) with a graphical user interface and a tutorial.

4 Definition of Pseudo-components

4.1 Introduction

The elucidation and quantification of unknown components in poorly specified mixtures is often infeasible in practical applications. To enable the thermodynamic modeling of such mixtures, pseudo-components are often introduced. The most prominent field in which poorly specified mixtures are modeled by defining pseudo-components is petrochemical engineering [2, 3, 9, 13, 73–75]. 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 [76]; 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 proposed 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. The 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 that is proposed here is based on the elucidation and quantification of *structural groups* in the mixtures by 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 [31]. 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.

The applicability of the method is demonstrated 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.

4.2 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 chapter, predictive thermodynamic modeling of the properties of the mixture can be carried out. Figure 15 visualizes the workflow.

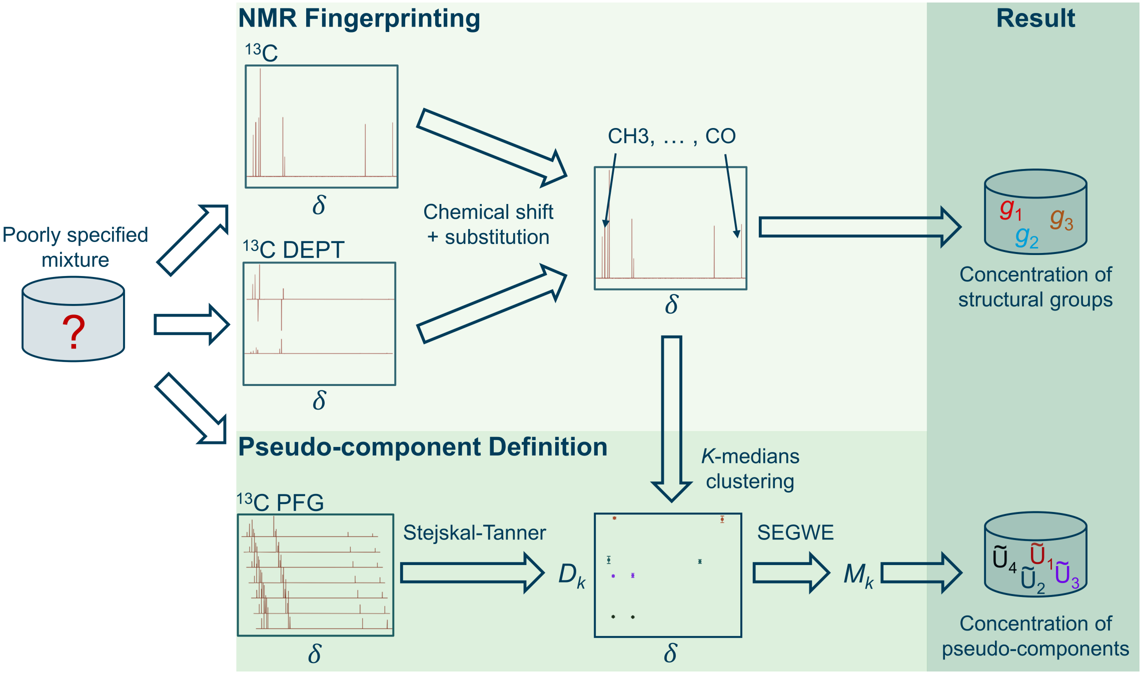


Figure 15: Scheme of the proposed method for the rational definition of pseudo-components. δ denotes the chemical shift. D_k and M_k denote the self-diffusion coefficient and the molar mass, as estimated with the SEGWE model [77, 78] of pseudo-component k, respectively.

In the first step, the 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 15 (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, 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, is used. The simplest way for assigning signals to structural groups is using chemical shift tables [25], namely assigning distinct structural groups to fixed regions of chemical shift of the NMR spectrum. In the present chapter, such a simple approach was applied, based on the shift table of the ^{13}C NMR spectrum used in Refs. [59, 60, 62, 63], which was, however, refined here as described in the following.

In contrast to Refs. [59, 60, 62, 63], the method described here 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. It is noted that also more sophisticated approaches like, e.g., the machine-learning model from Chapter 3, will be interesting to combine with the method of this chapter in the future.

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.

In the second step of the method, the structural groups are clustered to multiple pseudo-components, cf. Figure 15 (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 [79–82]. PFG NMR spectroscopy can, of course, support the elucidation of components, and has been applied for this purpose, see, e.g., Refs. [83–88]. However, to the best of the authors' 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 chapter. 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 [85, 86, 89, 90]. 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. [90]; however, since the signals of quaternary carbons would thereby be suppressed, this approach was not used here.

The ratio behind using PFG NMR spectroscopy in this chapter 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. Therefore, an unsupervised machine-learning technique is used for this purpose.

Specifically, K-medians clustering is used, which is a variant of the K-means clustering algorithm [47], 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 the thesis here, is a priori unknown. This is solved by using the so-called silhouette score [91], 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 15), 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 [77, 78], which was used in the present chapter.

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, the concept of relative diffusion coefficients [92–96] is used, 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 the 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 the 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 chapter and are discussed in Chapter 6.

4.2.1 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 chapter, 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 4.2. In Appendix C, Figures C.1-C.3, ^{13}C NMR spectra of the mixtures are shown. Details are given in Appendix C in Section C.1.

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. It is noted that, in the present chapter, the focus was not on time efficiency of the experiments, which will, however, be addressed in future work, e.g., by using fewer gradient steps in PFG NMR [95] or by exploiting polarization transfer techniques like polarization enhancement nurtured during attached nucleus testing (PENDANT) [97] 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 [98, 99]. 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.

4.2.2 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 the thermodynamic background of the present thesis, a set of organic structural groups as they are used in a widely applied thermodynamic group-contribution method, the UNIFAC-method [31], is used for the regions in the chemical shift tables. This choice can be considered as an example for the application of the method. Furthermore, only groups containing C, H, and O atoms were considered in the present chapter. Extensions and adaptions of the 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 Appendix C.

Table 6 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 6: Assignment of structural groups from the UNIFAC [31] 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. Appendix C.

^{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 [68]. After the identification of the structural groups, they can be quantified, which was done here based on group mole fractions x_g:

$$x_g = \frac{\frac{A_g}{z_g}}{\sum_{g=1}^{G} \frac{A_g}{z_g}} \tag{8}$$

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 6. 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 [100–102]. 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 Appendix C.

4.2.3 Clustering of Structural Groups to Pseudo-components

4.2.3.1 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. (9), which is a modified version of the Stejskal-Tanner equation [81, 103]:

$$\ln\left(\frac{I_p}{I_{0,p}}\right) = -\sum_{n=1}^{2} c_n \left(D_p \gamma^2 \delta^2 \left(\Delta - \frac{\delta}{3} - \frac{\tau}{2}\right) G^2\right)^n \tag{9}$$

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 Ref. [81], it was empirically found that it is sufficient to consider the first two terms of the series in Eq. (9); 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 Ref. [81]. 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. (9) to the experimental data using MATLAB 2021 b [104], where $I_{0,p}$ was fitted simultaneously [69, 105].

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.

4.2.3.2 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 $\mathcal{X}^{\text{mix}} = \{(\boldsymbol{x}_p)\}_{p=1}^{P}$, 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 thesis, K-medians clustering was used, which is a variant of the K-means algorithm but which is more robust towards outliers [106, 107]. 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 [104]. 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 [91] $\overline{s}(K)$, which is a common metric for automatically selecting the most suitable number of clusters for a given clustering problem. Intuitively, $\overline{s}(K)$ measures 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 Appendix C.

Based on the clustering, the relative number of the structural groups in the pseudo-components can be determined in analogy to Eq. (8); this was again done here based on group mole fractions $x_{g,k}$, now for each pseudo-component k.

4.2.4 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} [92–96] was applied. The diffusion coefficient of a pseudo-component $D_{\tilde{\mathrm{U}}}$ was thereby related to the diffusion coefficient of a known reference component D_{ref} in the same sample:

$$D_{\mathrm{rel}} = \frac{D_{\tilde{\mathrm{U}}}}{D_{\mathrm{ref}}} \tag{10}$$

In the literature [95, 96, 108], 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, the reader is referred to Appendix C, where this observation was verified based on own experiments with aqueous solutions up to mass fractions of the diffusing species of 0.28 g/g. Hence, the following assumption was used:

$$\frac{D_{\tilde{\mathrm{U}}}^{\infty}}{D_{\mathrm{ref}}^{\infty}} = \frac{D_{\tilde{\mathrm{U}}}}{D_{\mathrm{ref}}} = \mathrm{const.} \tag{11}$$

From Eq. (11), the number of $D_{\tilde{\mathrm{U}}}^{\infty}$ can be calculated from the experimental data for $D_{\tilde{\mathrm{U}}}$ and D_{ref} from the PFG NMR experiments if $D_{\mathrm{ref}}^{\infty}$ is known. It is recommended to select the reference component in a way that $D_{\mathrm{ref}}^{\infty}$ can be adopted from the literature. As an alternative, it can also be determined experimentally, which is, however, usually tedious [109], or estimated using a prediction method [77, 78, 110–112].

From $D_{\tilde{\mathrm{U}}}^{\infty}$, in turn, the molar mass $M_{\tilde{\mathrm{U}}}$ of component $\tilde{\mathrm{U}}$ can be calculated using basically any predictive model for self-diffusion coefficients at infinite dilution. The SEGWE model [77, 78] is used in the present chapter, which is a semi-empirical extension of the Stokes-Einstein equation [113] and was found to be the best available semi-empirical model for predicting self-diffusion coefficients in a recent study [109]:

$$D_{\tilde{\mathrm{U}}}^{\infty} = \frac{k_{\mathrm{B}} T \left(\frac{3\alpha}{2} + \frac{1}{1+\alpha} \right)}{6 \pi \eta_{\mathrm{S}} \sqrt[3]{\frac{3 M_{\tilde{\mathrm{U}}}}{4 \pi \rho_{\mathrm{eff}} N_{\mathrm{A}}}}} \tag{12a}$$

$$\alpha = \sqrt[3]{\frac{M_{\mathrm{S}}}{M_{\tilde{\mathrm{U}}}}} \tag{12b}$$

where $D_{\tilde{\mathrm{U}}}^{\infty}$ is the self-diffusion coefficient of pseudo-component $\tilde{\mathrm{U}}$ at infinite dilution, $M_{\tilde{\mathrm{U}}}$ is the molar mass of $\tilde{\mathrm{U}}$, 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 [78] $\rho_{\mathrm{eff}} = 627$ kg m^{-3} was used here. Calculating the molar mass $M_{\tilde{\mathrm{U}}}$ from Eq. (12) requires solving a cubic equation and choosing the appropriate solution [77, 78].

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. Appendix C 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.

4.3 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 7. 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 $\cdot$ s as reported for $T = 298.15$ K in Ref. [114]. 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 7: Overview of the test mixtures considered in this chapter. 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.

4.4 Results and Discussion

4.4.1 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 7, are shown. For better clarity, in this chapter, a distinction is made between the group-specific characterization (NMR fingerprinting) together with the clustering step in Section 4.4.1, yielding *relative* amounts for the structural groups in each pseudo-component, and the *quantitative* definition of pseudo-components that also involves predicting the respective molar masses in Section 4.4.2, 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 Appendix C.

4.4.1.1 Mixture I

Figures 16 and 17 show the results of the qualitative definition of pseudo-components for test mixture I. In the left part of Figure 16, 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 16 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 16 (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 16. 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.

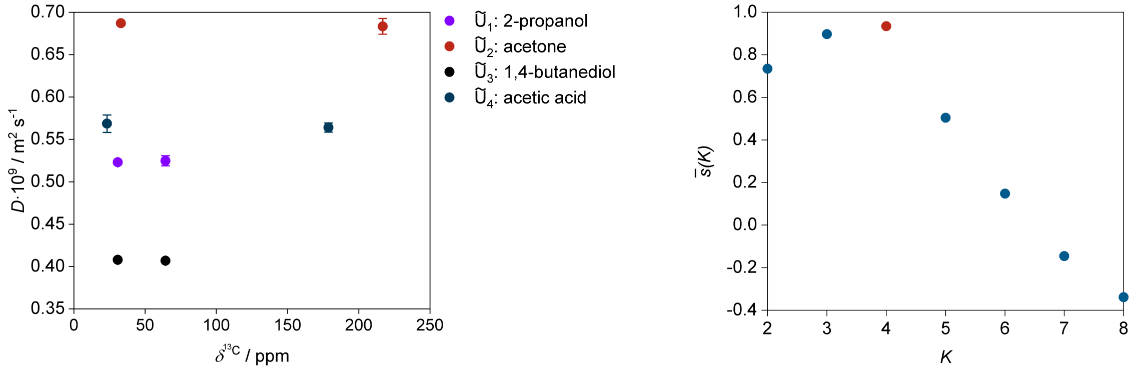


Figure 16: Left: DOSY map of mixture I showing the clustering of the structural groups into four pseudo-components $\tilde{U}$ ($K = 4$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars indicate 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 17 shows the relative composition of the four pseudo-components $\tilde{U}_1$-$\tilde{U}_4$ 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.

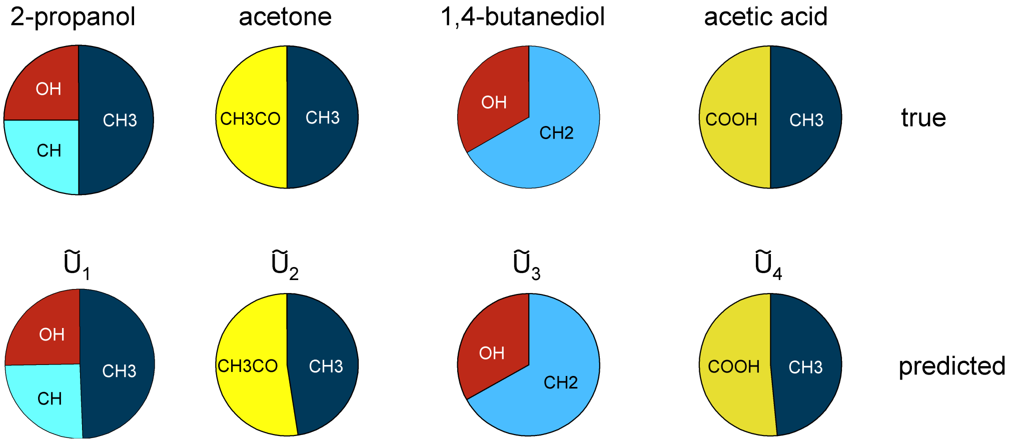


Figure 17: Relative composition of the true components (top row) and the predicted pseudo-components (bottom row) in mixture I in terms of group mole fractions $x_{g,k}$.

4.4.1.2 Mixture II

In Figures 18 and 19, 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 18 left. In Figure C.7 in Appendix C, it is demonstrated 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.

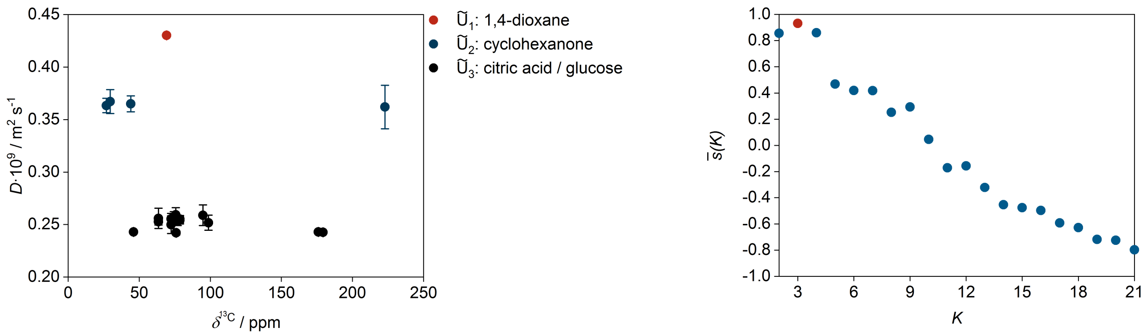


Figure 18: Left: DOSY map of mixture II showing the clustering of the structural groups into three pseudo-components $\tilde{\text{U}}$ ($K = 3$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars 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 19 shows the comparison of the predicted and the true group mole fractions in mixture II. For the component 1,4-dioxane ($\tilde{\text{U}}_1$), 'OH' groups are incorrectly identified in contrast to cyclic ether groups, called 'cy-CH2O' here and which are simply not included in the list of groups considered here, cf. Table 6. 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 demonstrated in Refs. [60, 62]. This can be attributed to the fact that the 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, it can be expected that a polar group is falsely interpreted as another polar group, as it is also the case with the example studied here.

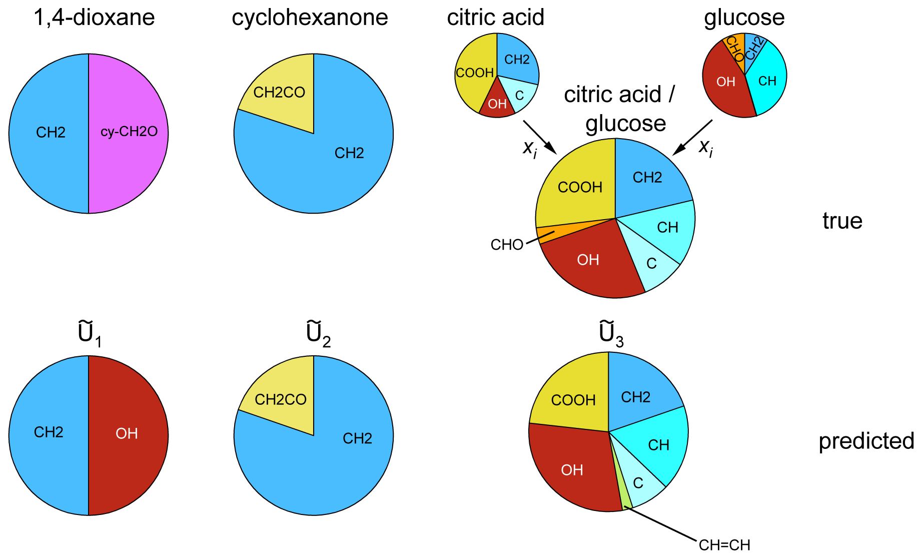


Figure 19: Relative composition of the true components (top row) and the predicted pseudo-components (bottom row) in mixture II in terms of group mole fractions $x_{g,k}$. Arrows indicate that a pseudo-component represents a "mixture" of multiple true components.

The relative composition of pseudo-component $\tilde{\mathrm{U}}_2$ shows an excellent agreement with that of the true component cyclohexanone, whereas the composition of pseudo-component $\tilde{\mathrm{U}}_3$ 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 $\tilde{\mathrm{U}}_3$, 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 [68].

4.4.1.3 Mixture III

Figures 20 and 21 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 20 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 20 left. Given the complexity of the ^{13}C NMR spectrum, cf. Figure C.3 in Appendix C, and the DOSY map for this mixture, cf. Figure 20 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 demonstrated in Figure C.7 in Appendix C.

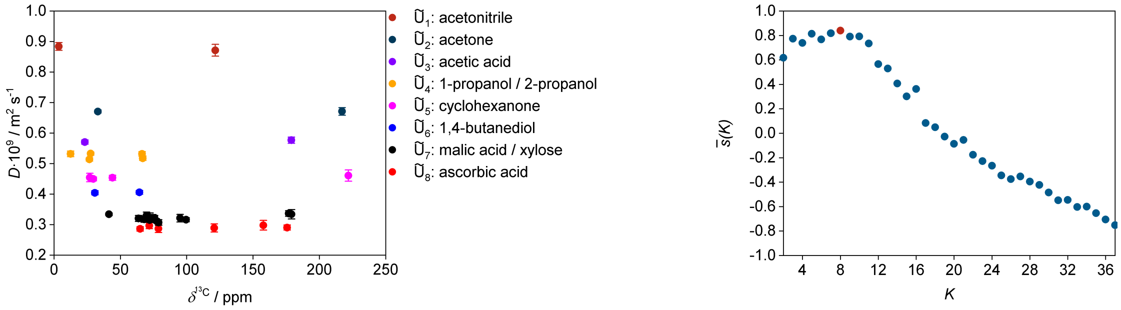


Figure 20: Left: DOSY map of mixture III showing the clustering of the structural groups into eight pseudo-components $\tilde{\mathrm{U}}$ ($K = 8$) by the K-medians algorithm. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error 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 21, 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-butanediol) or represent a “mixture” of the respective true components ($\tilde{\mathrm{U}}_4$, $\tilde{\mathrm{U}}_7$). Similar to the results observed for mixture II, cf. Figure 19, a small number of ’CH=CH’ groups is incorrectly predicted for $\tilde{\mathrm{U}}_7$. Furthermore, in pseudo-component $\tilde{\mathrm{U}}_8$, 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 the list of groups, cf. Table 6. Acetonitrile ($\tilde{\mathrm{U}}_1$) is made up of a single group in UNIFAC (’CH3CN’) that is not included in the list and therefore misinterpreted as a combination of a ’C=C’ and a ’CH3’ group.

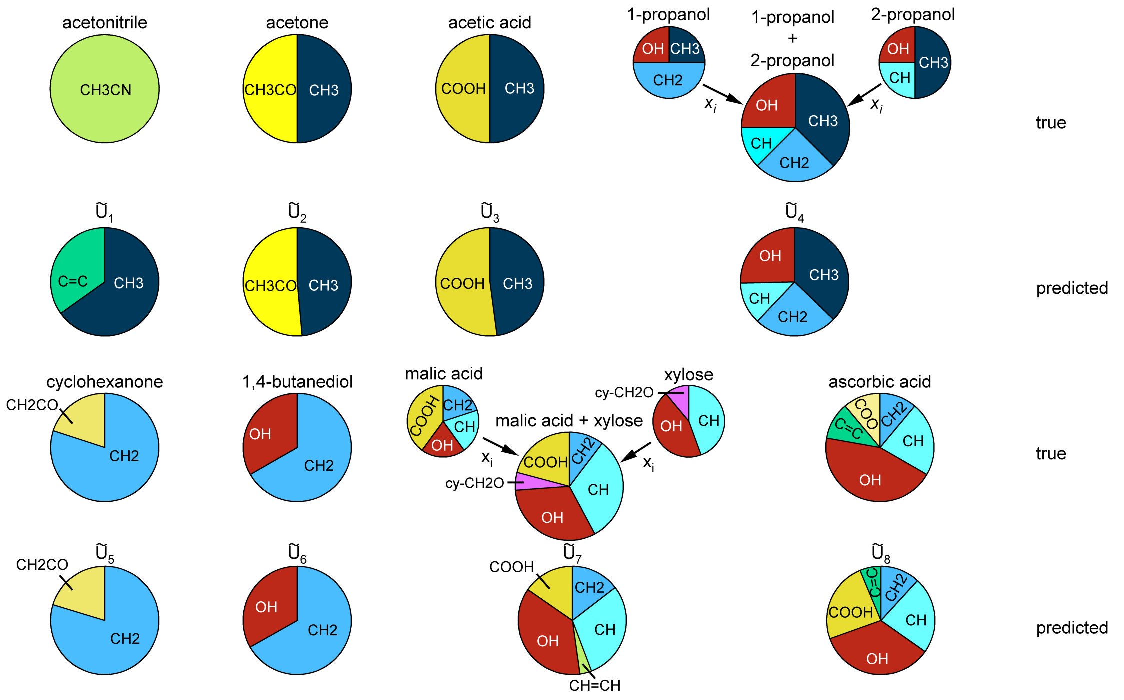


Figure 21: Relative composition of the true components (first and third row) and the predicted pseudo-components (second and fourth row) in mixture III in terms of group mole fractions $x_{g,k}$. Arrows indicate that a pseudo-component represents a "mixture" of multiple true components.

4.4.2 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 17, 19, and 21. For determining the *absolute* number of structural groups in each pseudo-component, information on the molar mass of each pseudo-component is required. In the present thesis, it is proposed to predict the molar mass based on the measured self-diffusion coefficients of each pseudo-component, and have used the SEGWE model [77, 78] 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 4.2.4 for details. For convenience, one of the components of each mixture is simply designated as reference component, but it is noted 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 22 (left) shows the results for the prediction of the molar mass of the pseudo-components $\tilde{U}$ 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. [115] ($D_{\mathrm{ref}}^{\infty} = 0.99 \cdot 10^{-9}$ 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.

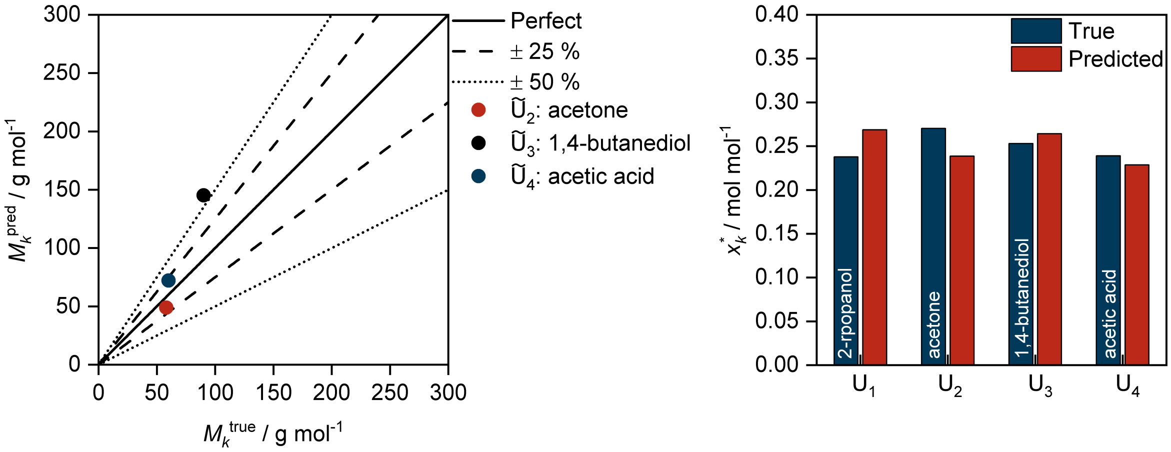


Figure 22: Left: prediction of the molar mass of the pseudo-components in mixture I based on the measured self-diffusion coefficients using the SEGWE model [77, 78]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture I and comparison to the true composition.

Figure 22 (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.

In Figure 23 (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. [116] ($D_{\mathrm{ref}}^{\infty} = 1.110 \cdot 10^{-9}$ 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 23 (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.

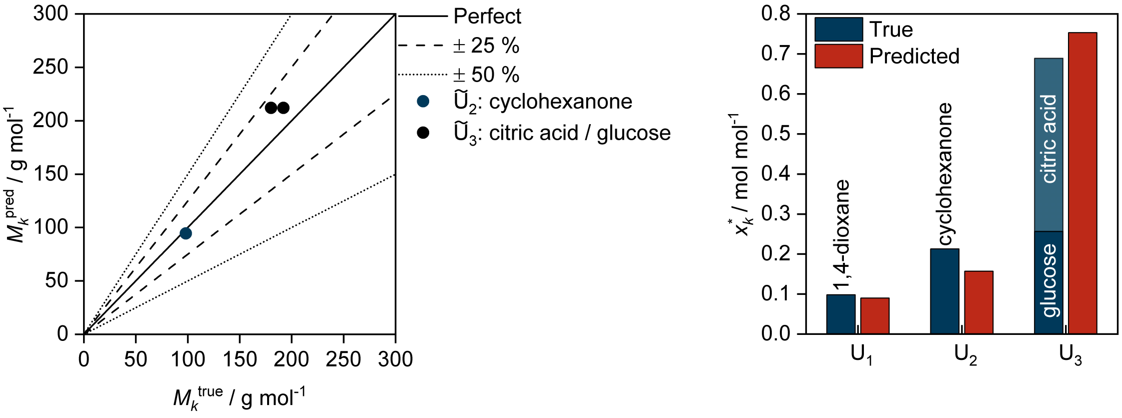


Figure 23: Left: prediction of the molar mass of the pseudo-components in mixture II based on the measured self-diffusion coefficients using the SEGWE model [77, 78]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{\mathrm{U}}$ in mixture II and comparison to the true composition.

Figure 24 (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. [117] ($D_\mathrm{ref}^\infty = 1.649 \cdot 10^{-9}\ \mathrm{m^2\ s^{-1}}$ 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 C.9 in Appendix C. 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 24 (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.

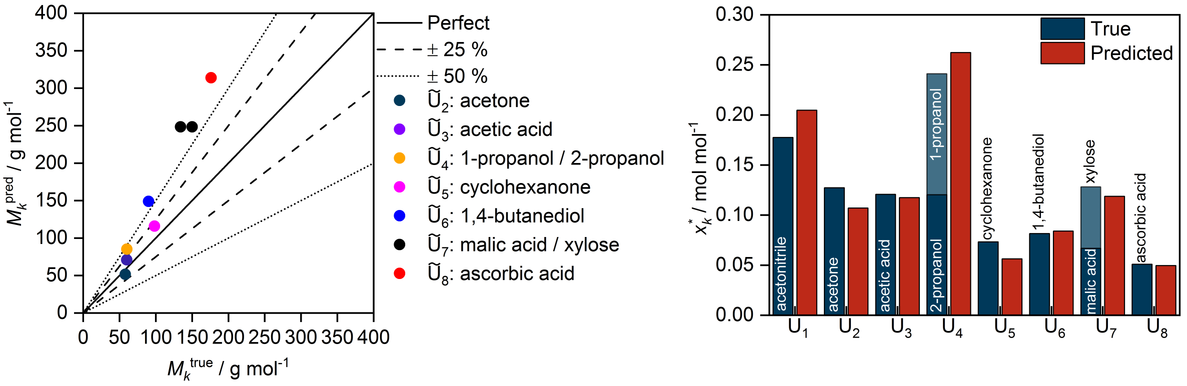


Figure 24: Left: prediction of the molar mass of the pseudo-components in mixture III based on the measured self-diffusion coefficients using the SEGWE model [77, 78]. Right: prediction of the water-free mole fractions x_k^* of the pseudo-components $\tilde{U}$ in mixture III and comparison to the true composition.

4.5 Conclusions

In this chapter, a new method for the representation of poorly specified mixtures by pseudo-components is introduced. Standard NMR experiments are carried out, yielding the basis for the *NMR fingerprinting*. 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.

In the present chapter, the applicability of the method was demonstrated 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, cf. Chapter 6.

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 thesis 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 chapter, 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 [68], cf. also Chapters 2 and 3.

5 Estimation of σ-Profiles and Activity Coefficients

5.1 Introduction

The separation of target components from complex liquid mixtures is a ubiquitous task in process engineering. Fluid separation processes, such as distillation, extraction, and crystallization, are based on concentration differences in coexisting phases and are often modeled using the equilibrium stage concept. Hereby information on the activity coefficients of the target components in the liquid phase is needed for the calculation of the phase equilibrium. Since the experimental determination of activity coefficients is tedious, prediction methods have been developed, of which the most widely used are UNIFAC [32, 50, 118] and COSMO-RS [119, 120]. UNIFAC is a group-contribution version of the G^{E} model UNIQUAC [121]. COSMO-RS is basically also a G^{E} model, in which the energetic part is described based on quantum-chemical calculations. The key element in this is the determination of the σ-profile that represents the surface charges on a cavity in a dielectric medium, in which the target molecule is embedded [122].

In the following, it is assumed that each poorly specified mixture contains at least one component of which the nature and the concentration are known, which is called target component here. In this chapter, a method is presented, that enables the prediction of the activity coefficients of such target components in poorly specified mixtures without having to know anything about the unknown components.

For solving this task, the NEAT method (NMR spectroscopy for the estimation of activity coefficients of target components in poorly specified mixtures) was developed in Refs. [59, 60]. NEAT is based on a combination of NMR spectroscopy and a thermodynamic group-contribution method and enables the prediction of the activity coefficient γ_{T} of a target component T in a poorly specified mixture based only on information on the target component T and a single NMR spectrum of the mixture. In NEAT, information on the type and concentration of the chemical groups of the unknown components in the poorly specified mixture is obtained from the NMR spectrum of the mixture.

The information on the groups is then used in a thermodynamic group-contribution method for calculating the activity coefficient of the target component T. It has been shown that the results obtained with NEAT for the poorly specified mixture often match those obtained for the fully specified mixture remarkably well [59, 60]. This has been demonstrated for aqueous, non-aqueous, and also for reactive mixtures. Furthermore, it has been shown that, using only a single NMR spectrum, NEAT can be successfully applied for predicting activity coefficients for any mixtures, as long as the composition of the unknown components remains constant [61]. This is, for instance, practically relevant for describing the selective removal of the target component T or the selective removal of a known solvent from the mixture. For a complete description of phase equilibria, reliable information on the activity coefficients of all components in the considered mixture is needed. However, the activity coefficient γ_T of the target component is directly related to the affinity of the target component to a second phase. In a recent study, it was demonstrated how NEAT can thus be applied for the conceptual design of liquid-liquid extraction processes [62].

However, in all previous works on NEAT [59–62], the same thermodynamic group-contribution method was used: UNIFAC. In this chapter, it is demonstrated that NEAT is a generic framework that is not restricted to the use of UNIFAC, but can, in principle, be combined with any group-contribution method. In all cases, the quality of the NEAT results is inherently limited by the quality of the underlying thermodynamic group-contribution method. It would only be by chance that the predictions of experimental data obtained from NEAT would be better than those obtained from the group-contribution method (assuming the composition of the mixture would be known). It is, therefore, interesting to study the application of different thermodynamic group-contribution methods in NEAT.

The previous studies on NEAT are extended here and it is investigated how COSMO-RS, which is one of the most important models for predicting activity coefficients besides UNIFAC, can be used within the NEAT framework. In its original version, COSMO-RS [119, 120, 122] is not a group-contribution method. However, group-contribution methods have been developed, that enable an estimation of σ-profiles, cavity surface areas A, and cavity volumes V without having to perform the quantum-chemical calculation. Such a group-contribution version of COSMO-RS is used in the present chapter. Furthermore, there are different versions of the COSMO-RS G^E model, some of which have not been fully disclosed. In the present chapter, the group-contribution version of COSMO-RS (OL) [123] as it was presented by the group of Gmehling in Oldenburg [124] is used. It is designated in the following as GC-COSMO-RS (OL). It is combined here with results from ^{13}C NMR spectroscopy from which the group composition of the studied mixtures was obtained in the same way as in Ref. [60]. The new method is tested using several poorly specified mixtures of the type known target component T + unknown components U + solvent water W as test cases. In all cases, information on the unknown components in the mixtures was not used for the predictions with NEAT but only for comparison of the predictions with results for the fully specified mixtures. New NMR measurements were carried out only for some of these mixtures. The other cases were already studied in Ref. [59] and the NMR spectroscopic results were simply re-evaluated to obtain information on the group speciation that could be used together with GC-COSMO-RS (OL). For all systems, the predictions obtained with the present version of NEAT (NMR + GC-COSMO-RS (OL)) were compared to results that were obtained using the full knowledge of the speciation of the mixture. Additionally, in one case, the predictions of the newly introduced version of NEAT are compared to predictions of the recently described version of NEAT (NMR + modified UNIFAC (Dortmund) [50, 59–62] as well as with results that were obtained with the respective G^{E} models using the full knowledge of the speciation of the mixture.

5.2 Materials and Methods

5.2.1 Materials

For all mixtures prepared in this chapter, ultra-pure water, which was produced with a water purification system of Merck Millipore (Elix Essential 5), was used. Information on all other chemicals that were used in the experiments in this chapter is given in Table 8.

Table 8: Suppliers and purities of chemicals used in this chapter. Purities according to supplier's specification.

Chemical	Formula	Supplier	Purity
Acetic acid	$C_2H_4O_2$	Carl Roth	≥99.80%
Acetone	C_3H_6O	Fisher Chemical	≥99.97%
Cyclohexanone	$C_6H_{10}O$	Sigma Aldrich	≥99.80%
Ethanol	C_2H_6O	Merck	≥99.90%
2-Propanol	C_3H_8O	Sigma Aldrich	≥99.90%
TMSP-d4	$NaC_6H_9D_4O_2Si$	Deutero	≥98.00%

5.2.2 Methods

5.2.2.1 Sample Preparation and NMR Spectroscopy

All sample mixtures were prepared gravimetrically in 20 ml glass vessels. The mass of each sample was approximately 10 g. To each sample, a small amount of sodium 3-(trimethylsilyl)tetradeuteriopropionate (TMSP-d4) was added as NMR standard. After thoroughly shaking the samples to obtain homogenous solutions, 1 ml of each sample was transferred to a 5 mm NMR vial. Quantitative inverse-gated ^{13}C NMR spectra were recorded with a 400 MHz NMR spectrometer from Bruker (Avance) with a cryogenic probe (pulse angle: 60°, acquisition time: 1.5 - 5.6 s, number of scans: 64 or 128, Relaxation Delay: 45 - 65 s). Software from MestReLabs (MNova) was used for baseline correction, phase correction, and quantitative evaluation. Direct integration turned out to be sufficient in all cases, as basically only spectral areas, which were assigned to groups, had to be evaluated. The chemical shift refers to the TMSP-d4 peak.

5.2.2.2 Assignment and Quantification of Chemical Groups

The composition of all studied mixtures was known from sample preparation. However, this information was not used for the predictions with NEAT, but only for testing the quality of predictions. For the predictions with NEAT, each mixture was considered as poorly specified and of the type known target component T + unknown components U + water W. Only the nature and the mass fraction of the target component were assumed to be known. Information on the unknown components U was obtained from the NMR spectrum of the mixture. As comprehensively described in Ref. [60], this includes the assignment of chemical groups to regions of chemical shift in the NMR spectrum. The group assignment used for the evaluation of the NMR spectra will in general not be completely identical with the group list of the thermodynamic group-contribution method. Hence, a mapping is needed. As the group lists of modified UNIFAC (Dortmund), in the following abbreviated as UNIFAC (DO), and GC-COSMO-RS (OL) are not identical, the group assignment cannot simply be adopted from Ref. [60] but has to be reconsidered. The assignment used in this chapter for the predictions with NEAT based on GC-COSMO-RS (OL) is given in Table 9. The group assignment used in this chapter for the predictions with NEAT based on UNIFAC (DO), which is shown in Table 10, is similar, but not identical with the one from Ref. [60]. The only difference is that 'CH3' and 'CH2' groups were distinguished here as it was observed that these groups can easily be distinguished in most spectra that were evaluated. The mass fraction of the groups in the mixture was obtained from the integration over the chemical shift regions. All identified groups were lumped to a pseudo-component $\tilde{\mathrm{U}}$ to which a molar mass of 150 g mol^{-1} was assigned, as suggested in Refs. [59, 60]. The mass fraction of water W in the poorly specified mixture was calculated from the mass balance. In another version of NEAT, for which results are provided in Appendix D, information on the mass fraction of water from the gravimetric sample preparation was used. The stoichiometry of the pseudo-component Ũ and the estimated composition of all studied mixtures are given in Appendix D.

Table 9: Assignment of chemical groups from GC-COSMO-RS (OL) to ^{13}C NMR chemical shift regions used for the predictions with NEAT in this chapter. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

^{13}C NMR chemical shift region / ppm	chemical group name	GC-COSMO-RS (OL) label
0 - 30	methyl group	'CH3' (1)
30 - 60	methylene group	'CH2' (4)
60 - 90	alcohol group	'CH2' (7)[a]+'OH(P)' (35)
90 - 150	alkenyl group	'CH=CH' (58)
150 - 180	carboxyl group	'COOH' (44)
>180	carbonyl group	'CO' (51)
n.a.	water	n.a.

[a] In GC-COSMO-RS (OL), a 'CH2' group bound to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table 10: Assignment of chemical groups from the UNIFAC (DO) table to ^{13}C NMR chemical shift regions used for the predictions with NEAT in this chapter. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

^{13}C NMR chemical shift region / ppm	chemical group name	UNIFAC (DO) label
0 - 30	methyl group	'CH3' (1,1)
30 - 60	methylene group	'CH2' (2,1)
60 - 90	alcohol group	'CH2' (2,1)+'OH(P)' (14,5)
90 - 150	alkenyl group	'CH=CH' (6,2)
150 - 180	carboxyl group	'COOH' (42,20)
>180	carbonyl group	'CH2CO' (19,9)
n.a.	water	'H2O' (16,7)

5.2.2.3 Calculation of Activity Coefficients

After estimating the groups of the pseudo-component $\tilde{\text{U}}$, each poorly specified mixture is considered as a pseudo-ternary mixture of the components target component T + pseudo-component $\tilde{\text{U}}$ + water W. In the following, the procedure for the calculation of the activity coefficient γ_T of the target component T in the poorly specified mixture based on a combination of NMR spectroscopy and GC-COSMO-RS (OL) is described. The procedure for the application of NEAT based on UNIFAC (DO) is the same as in Ref. [60] and, therefore, not discussed here. In COSMO-RS (OL), σ-profiles and cavity surface areas for the hydrogen bonding (hb) part, and the non-hydrogen bonding (nhb) part are considered [123, 124], which are denoted here as $p^\text{hb}(\sigma)$, $p^\text{nhb}(\sigma)$ and A^hb, A^nhb respectively. Furthermore, in COSMO-RS (OL), activity coefficients are modeled as the sum of two contributions, a residual and a combinatorial part. The residual part of the activity coefficient of a component is calculated from its cavity surface areas A^hb and A^nhb and the sigma-profiles $p^\text{hb}(\sigma)$ and $p^\text{nhb}(\sigma)$ of all pure components that make up the mixture. The combinatorial part is calculated from the cavity volumes and the cavity surface areas of the pure components. For the target component T and water W, σ-profiles, cavity surface areas A, and cavity volumes V from quantum-chemical calculations were used, which were taken from the Dortmund Data Bank, 2018, www.ddbst.com (DDB) [125]. For the pseudo-component $\tilde{\text{U}}$, the corresponding numbers were obtained from a summation of the contributions of the estimated groups of $\tilde{\text{U}}$ using the available group parameters of GC-COSMO-RS (OL) reported in the DDB. Corrections that were introduced in GC-COSMO-RS (OL) to consider intramolecular interactions, such as intramolecular hydrogen bonds or steric effects, were neglected in the present version of NEAT. For the results based on the fully specified mixture, which were only used for comparison here, the σ-profiles, cavity surface areas A, and cavity volumes V of the unknown components U were calculated as described above, whereas, for the target component T and water W, the numbers were taken from the DDB, if not stated otherwise. The calculations of all activity coefficients were carried out using the COSMO-RS (OL) tool supplied with the installation of the DDB. Computational details are given in Appendix D.

5.3 Results and Discussion

For simplicity, only activity coefficients at $T = 298$ K and $p = 1$ bar are considered here. The pressure is not specified in the following as it has only a weak influence on the activity coefficient and is not considered in the studied G^E models. In the following, the term “system” refers to a set of components without specifying the composition whereas the term “mixture” is used if additionally the composition is specified. For each system, several mixtures were studied. Table 11 gives an overview of the systems that were investigated in the present chapter. In all following figures, lines denote the results that were obtained using the full knowledge on the speciation of the mixture, whereas the symbols denote the predictions with NEAT, for which only the information on the target component T was used, if not stated otherwise. The predictions with NEAT are based on GC-COSMO-RS (OL), if not stated otherwise. In Figure 25, results for the activity coefficient of two target components (T = acetone or T = acetic acid) in ternary aqueous mixtures of system I and II are shown. In all cases, 2-propanol is considered to be the unknown component U. The results for the fully specified mixtures show a strong influence of U on γ_T for T = acetone and a rather weak influence for T = acetic acid. For all mixtures, excellent agreement of the prediction with NEAT with the results for the fully specified mixtures is observed. In Figure D.4 in Appendix D, it is shown that the assumed molar mass $M_{\tilde{\mathrm{U}}}$ of the pseudo-component Ũ, which was set to $M_{\tilde{\mathrm{U}}}$ = 150 g mol^{-1} for the results in Figure 25, has only a minor influence on the predictions with NEAT, if no unreasonably small values of $M_{\tilde{\mathrm{U}}}$ are chosen. This was found for all studied mixtures as well as in Refs. [59, 60]. Therefore, only predictions based on $M_{\tilde{\mathrm{U}}}$ = 150 g mol^{-1} are shown in the following. In Figure D.18 in Appendix D, it is demonstrated that NEAT is not restricted to a prediction at T = 298 K, i.e., the temperature at which the NMR

Table 11: Overview of the studied systems. Besides the target component T and the unknown components U, water W is always present in the mixtures. In all mixtures of a given system, the molar ratio of the target component to water $n_\mathrm{T}/n_\mathrm{W}$ is constant. For all five-component systems, the mass ratio of the three unknown components is 1:1:1 in all mixtures.

System	Target component T	Unknown component(s) U	$n_\mathrm{T}/n_\mathrm{W}$	NMR data
I	acetone	2-propanol	0.040	This chapter
II	acetic acid	2-propanol	0.050	This chapter
III	ethanol	acetic acid	0.043	Ref. [59]
IV	ethanol	methyl acetate	0.044	Ref. [59]
V	ethanol	2-butanone	0.044	Ref. [59]
VI	ethanol	cyclohexanone	0.044	This chapter
VII	1,4-butanediol	cyclohexanone acetonitrile methyl acetate	0.022	Ref. [59]
VIII	acetone	D-xylose acetic acid methyl acetate	0.035	Ref. [59]

analysis is carried out, but also gives good results for other temperatures. This is not surprising, since GC-COSMO-RS (OL), which is the basis of the NEAT version here, takes the temperature dependence of activity coefficients into account.

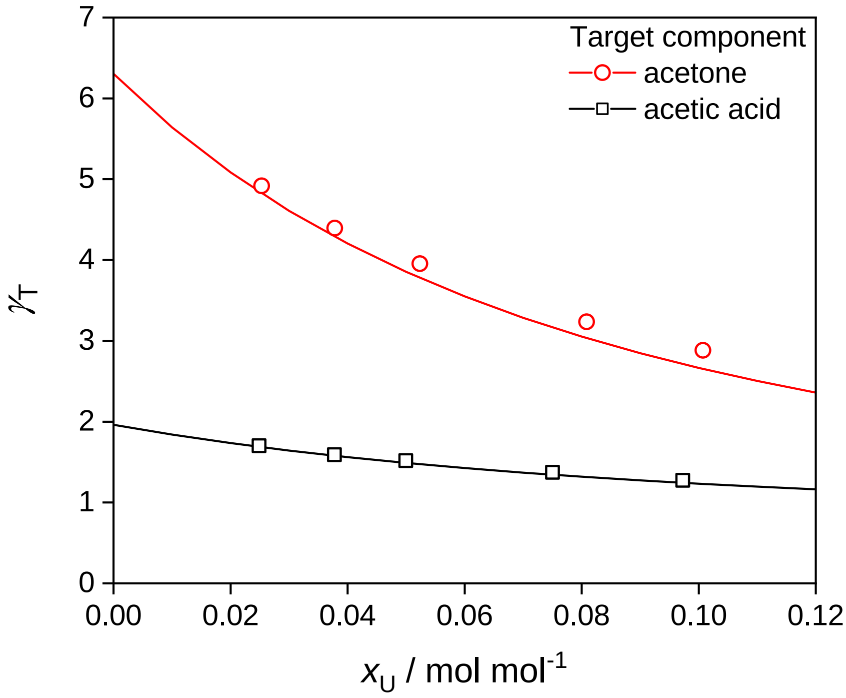


Figure 25: Activity coefficient γ_{T} of target components (T = acetone or T = acetic acid) in ternary mixtures of system I and II, cf. Table 11, at 298 K. Lines: results from GC-COSMO-RS (OL) for the fully specified mixtures. Symbols: predictions with NEAT. No information on the unknown component U was used in NEAT.

Figure 26 shows the σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system I (x_{U} = 0.025 mol mol^{-1}) as predicted by NEAT based on the NMR analysis of the mixture. For comparison, also the σ-profile of U = 2-propanol is shown. The agreement of the σ-profiles is very good, which explains the excellent prediction of γ_{T} in Figure 25. The respective plot for a mixture of system II is given in Figure D.5 in Appendix D and shows a similar agreement.

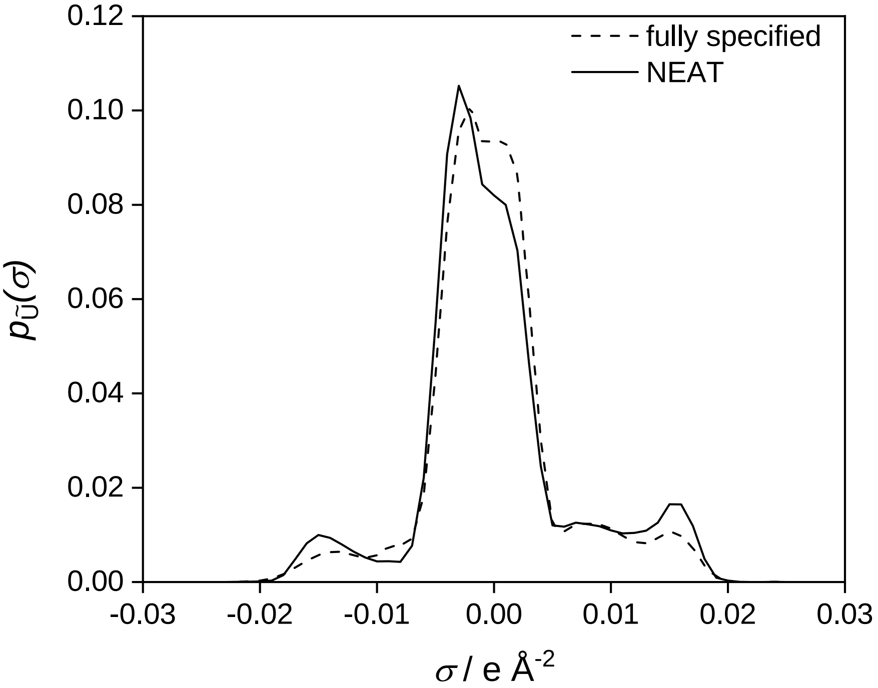


Figure 26: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system I as predicted with NEAT and the σ-profile of U = 2-propanol. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Appendix D. Dashed line: result from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.037 mol mol^{-1}, x_{U} = 0.025 mol mol^{-1}.

In Figure 27, the influence of different unknown components U on the activity coefficient γ_{T} of the target component T = ethanol is studied. The unknown components are U = acetic acid, U = methyl acetate, U = 2-butanone, and U = cyclohexanone (systems III - VI, cf. Table 11). The studied compositions were limited by the solubilities in the respective systems. The results show that acetic acid has a weak influence on γ_{T}, whereas cyclohexanone has a strong influence. Methyl acetate and 2-butanone show a medium influence. For all unknown components U, the agreement of the predictions with NEAT with the results for the fully specified mixture is very good. In particular, the differences in the strength of the influence of the different unknown components U on γ_{T} are predicted well with NEAT.

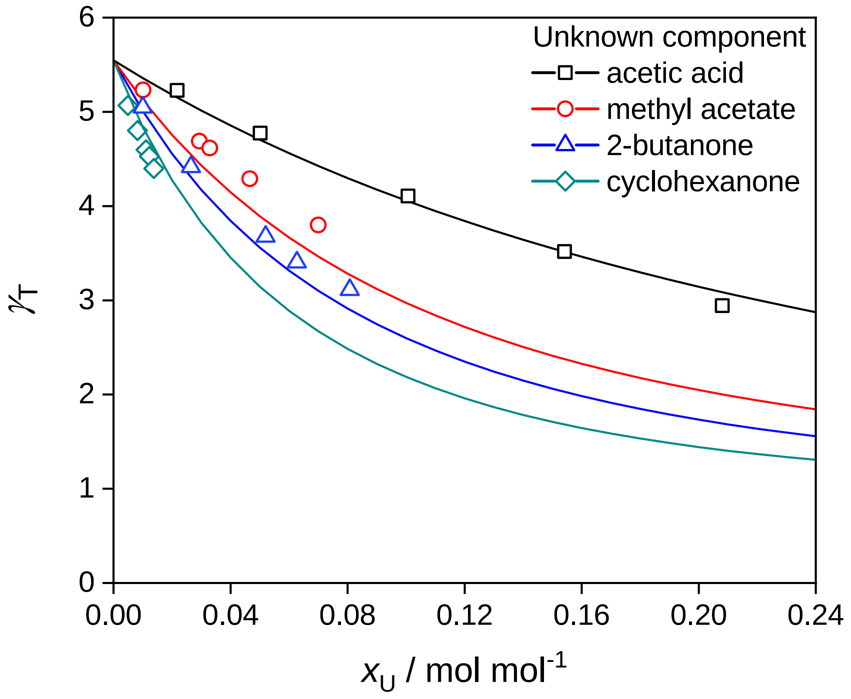


Figure 27: Activity coefficient γ_T of target component T = ethanol in ternary mixtures of systems III - VI, cf. Table 11, at 298 K. Lines: results from GC-COSMO-RS (OL) for the fully specified mixtures. Symbols: predictions with NEAT. No information on the unknown component U was used in NEAT.

Figure 28 shows the σ-profile of the pseudo-component $\tilde{U}$ in a mixture of system III ($x_U = 0.022$ mol mol^{-1}) as predicted with NEAT based on an NMR analysis of the mixture. For comparison, also the σ-profile of U = acetic acid is shown. The agreement of the σ-profiles is excellent, which is in line with the excellent prediction of γ_T in Figure 27. The respective plots for systems IV - VI are given in the Figures D.6-D.8 in Appendix D and show a similar agreement.

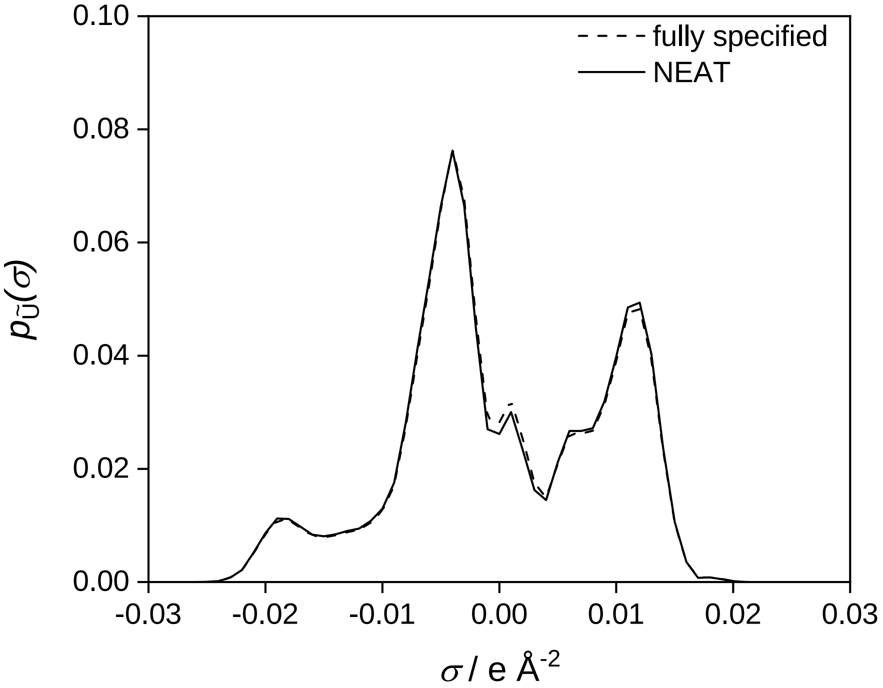


Figure 28: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system III as predicted with NEAT and the σ-profile of U = acetic acid. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Appendix D. Dashed line: result from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.041 mol mol^{-1}, x_{U} = 0.022 mol mol^{-1}.

In Figure 29, results for the activity coefficient of two target components (T = 1,4-butanediol or T = acetone) for aqueous five-component mixtures containing the unknown components U = cyclohexanone, acetonitrile, methyl acetate or U = D-xylose, acetic acid, methyl acetate are shown. For D-xylose, GC-COSMO-RS (OL) cannot be used in its current version since the required parameters are not available. Therefore, for obtaining the results for system VIII, quantum-chemical obtained σ-profiles, cavity surface areas A, and the cavity volume V from the DDB were used for D-xylose for the fully specified mixture. In Figure 29, results for two versions of NEAT are shown, one that is based on GC-COSMO-RS (OL) and one that is based on UNIFAC (DO). Both predictions with NEAT are compared to the results for the fully specified mixture that were obtained with the respective thermodynamic model. For system VIII, the results from GC-COSMO-RS (OL) and UNIFAC (DO) are similar, whereas there is a large deviation for system VII. However, it is not the intention of the present chapter to evaluate and compare the suitability of GC-COSMO-RS (OL) and UNIFAC (DO) for the studied mixtures. For both models, the agreement of all predictions with NEAT with the results for the fully specified mixture is excellent. The results demonstrate the generic nature of the idea of NEAT that allows to select the most suitable thermodynamic model for the studied system.

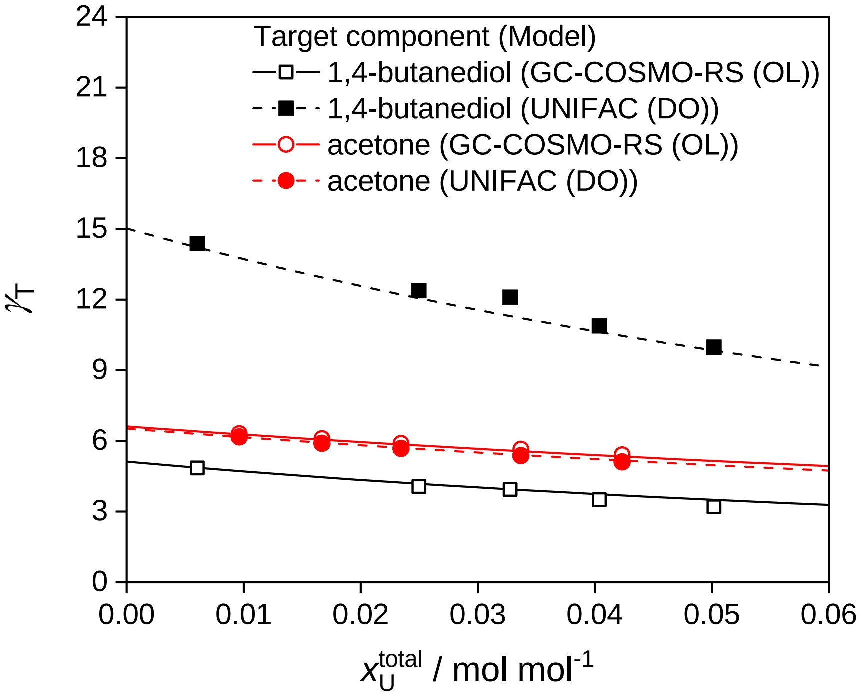


Figure 29: Activity coefficient γ_T of target components (T = 1,4-butanediol or T = acetone) in five-component mixtures of system VII and VIII, cf. Table 11, at 298 K. Lines: results for fully specified mixtures. Symbols: predictions with NEAT. No information on the unknown components U was used in NEAT.

Figure 30 shows the σ-profiles of the pseudo-component $\tilde{U}$ calculated with information from NEAT and for the fully specified mixture in a five-component aqueous mixture of system VII ($x_U^{total} = 0.006$ mol mol^{-1}). Compared to the results in Figures 26 and 28, the deviations are larger, which can be attributed to the more complex nature of the mixture. The respective plot for system VIII is given in Figure D.9 in Appendix D and shows a similar agreement.

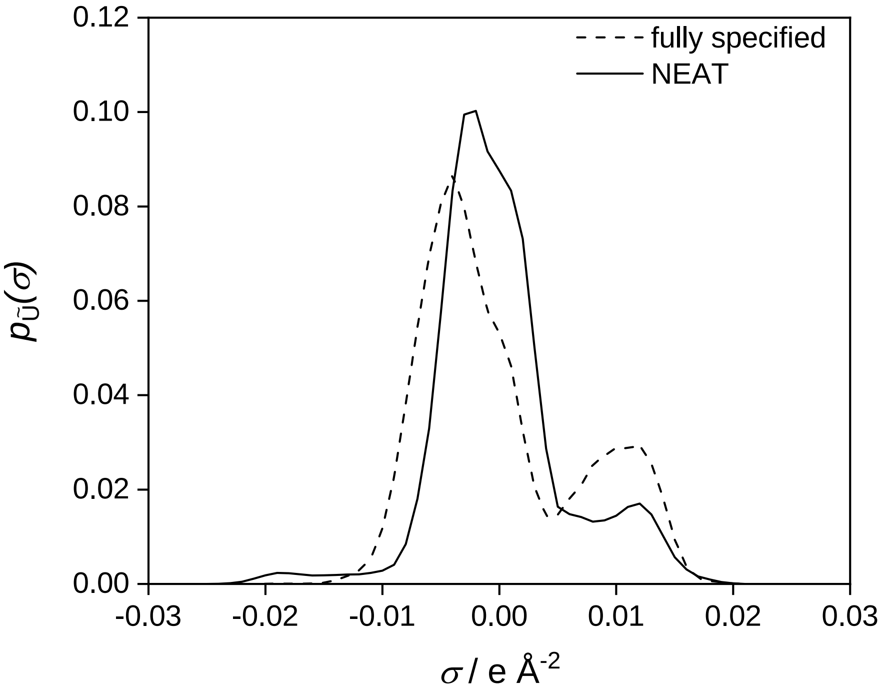


Figure 30: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system VII as predicted with NEAT and the mixed σ-profile of the unknown components U_1 = cyclohexanone, U_2 = acetonitrile, U_3 = methyl acetate. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Appendix D. Dashed line: results from GC-COSMO-RS (OL) for the mixture of all U_i in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.021 mol mol^{-1}, $x_{\mathrm{U}}^{\mathrm{total}}$ = 0.006 mol mol^{-1}.

5.4 Conclusions

In previous works, the method NEAT was introduced, which enables the prediction of the activity coefficients of target components in poorly specified mixtures based on a combination of NMR spectroscopy and thermodynamic group-contribution methods. In all previous studies with NEAT, that group-contribution method was UNIFAC. In the present chapter, it is demonstrated that NEAT can also be used with other group-contribution methods. This is important as the quality of the predictions of NEAT is determined by the quality of the underlying thermodynamic group-contribution method. The method that was used in the present chapter is GC-COSMO-RS (OL), which is a group-contribution version of COSMO-RS (OL). The new version of NEAT is tested with several poorly specified mixtures. The predictions show excellent agreement with results for the fully specified mixtures. The results of this chapter demonstrate that the concept of NEAT is generic. It can be used together with different thermodynamic group-contribution methods. Only the mapping of the group table that is used in the NMR evaluation and that of the thermodynamic group-contribution method has to be adjusted.

6 Conceptual Fluid Separation Process Design

6.1 Introduction

Modeling and simulation are essential for process development and optimization. In the present chapter, the focus is on fluid separation processes, namely liquid-liquid extraction and distillation. For modeling such processes, information on the phase equilibria is needed, which is typically supplied in the form of models of the Gibbs excess energy G^{E}, such as NRTL [126], UNIQUAC [121], or the group-contribution method UNIFAC [31, 127]. These models, as well as all other thermodynamic models, require a complete specification of the composition of the studied mixture.

These problems are circumvented by using the methodology described in Chapter 4 that results in a quantitative representation of the unknown components in a mixture by K pseudo-components $\tilde{\mathrm{U}}_k$, described by their structural group composition. The obtained numbers for the structural groups of type g in pseudo-component k, denoted by $\nu_{g,k}$, may have non-integer values as several true components with different concentrations can be lumped into a single pseudo-component. However, this does not hinder the application of the results for predicting many thermodynamic properties, such as normal-boiling points [27], vapor pressures [29], critical properties [28], or even quantum-chemical descriptors [124] and activity coefficients [31, 32] using respective group-contribution methods. Hence, this procedure allows the prediction of thermodynamic properties of mixtures without the need to elucidate the unknown components' molecular structure. Such modeling generally requires mapping the group distribution scheme used in defining the pseudo-components to that used in the thermodynamic group-contribution method, which is a task that usually can be solved.

In summary, the quantitative analysis of group compositions via NMR fingerprinting together with a rational approach for defining pseudo-components yields, in principle, all pieces of information that are required for modeling processes with poorly specified mixtures – without the need to elucidate and quantify all true species, which can become highly elaborate in practice.

In the present chapter, the applicability of this approach is demonstrated by the thermodynamic modeling of poorly specified mixtures and using the resulting models for fluid separation process design. For this purpose, the phase behavior of various poorly specified mixtures was predicted. The results were used for solvent screening for liquid-liquid extraction processes and for simulating open evaporation processes by predicting residue curves. All this is done without knowing the exact composition of the mixture. The group-contribution method UNIFAC [31, 127] was used to predict the activity coefficients for the phase equilibrium calculations. Furthermore, a method for the description of the vapor pressure [27, 29] of the pseudo-components, which is needed for the description of vapor-liquid equilibria, was employed. Note that although only liquid-liquid equilibria and vapor-liquid equilibria are considered in this chapter, the methodology can also be extended to other types of phase equilibria, like solid-liquid equilibria or vapor-liquid-liquid equilibria, straightforwardly; the only prerequisite for doing so is that suitable group-contribution methods for the respective relevant fluid properties are available.

6.2 Methods

6.2.1 Overview of the Workflow

Figure 31 summarizes all steps of the procedure applied in the present chapter. The methodology for NMR fingerprinting and the pseudo-component method, cf. upper part of Figure 31, were adopted from Ref. [128], cf. also Chapter 4; the main contribution of this chapter is the application of the respective results together with thermodynamic group-contribution methods for predicting phase equilibria and simulating thermal separation processes, cf. lower part of Figure 31.

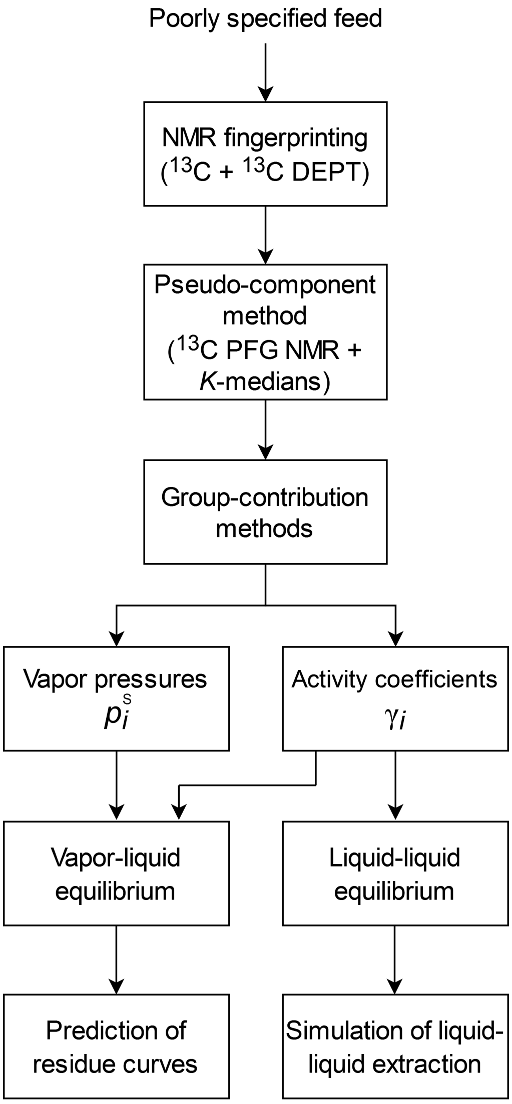

Poorly specified feed
NMR fingerprinting (^{13}C + ^{13}C DEPT)
Pseudo-component method (^{13}C PFG NMR + *K*-medians)
Group-contribution methods
Vapor pressures p_i^s
Activity coefficients γ_i
Vapor-liquid equilibrium
Liquid-liquid equilibrium
Prediction of residue curves
Simulation of liquid-liquid extraction

Figure 31: Schematic overview of the workflow applied in the present chapter. The NMR fingerprinting method and the pseudo-component method were adopted from Chapter 4. Their results were used to simulate liquid-liquid extraction processes and predict residue curves for open evaporation processes with poorly specified feeds.

6.2.2 NMR Fingerprinting and Pseudo-component Method

No experiments were carried out in the present chapter. The NMR-spectroscopic information of the studied mixtures required for the NMR fingerprinting, namely the ^{13}C NMR and ^{13}C DEPT NMR spectra, was taken from Chapter 4.

The NMR fingerprinting was based on the approach of Chapter 4, in which the groups identified in the fingerprinting were already mapped to groups of UNIFAC [31, 127]. Additionally, in this chapter, also a mapping of the groups determined in the fingerprinting to the groups of the method of Refs. [27, 29] for the prediction of vapor pressures was carried out. For technical details, cf. Appendix E.

Table 12 summarizes the structural groups distinguished in this chapter together with the mapping to the groups of UNIFAC [31, 127] and the method of Refs. [27, 29]. Furthermore, Table 12 also indicates the assignment of each group to regions of chemical shift taking into account the substitution degree of the carbon atoms, which was obtained by considering ^{13}C DEPT NMR spectra.

Table 12: Mapping of groups identified and quantified by NMR fingerprinting based on ^{13}C and ^{13}C DEPT NMR to those from Refs. [27, 29] and UNIFAC [31, 127, 129]. Carbon atom classification: P primary, S secondary, T tertiary, Q quaternary. The numbers in parentheses correspond to the group identifiers in the publications of the original methods [27, 29, 31, 127, 129].

^{13}C NMR chemical shift region	Carbon	Nannoolal [27, 29] label	UNIFAC [31, 127, 129] label
0-60 ppm	P	CH3 (1)	CH3 (1)
	S	CH2 (4)	CH2 (2)
	T	CH (5)	CH (3)
	Q	C (6)	C (4)
60-90 ppm	S	CH2 (7) + OH(P) (35/36)[a]	CH2 (2) + OH (14)
	T	CH (7) + OH(S) (34)	CH (3) + OH (14)
	Q	C (7) + OH(T) (33)	C (4) + OH (14)
90-150 ppm	S/T	CH=CH (58)	CH=CH (6)
	Q	C=C (58)	C=C (70)
150-180 ppm	Q	COOH (44)	COOH (42)
>180 ppm	T	CHO (52)	CHO (20)
	Q	CO (51)	CH3CO (18) / CH2CO (19)

[a] If the number of carbon atoms in the pseudo-component is less than five, group 36 was used, otherwise group 35 was used, as recommended in Refs. [27, 29].

The quantification of the groups was done by a manual integration of the respective peaks in the quantitative ^{13}C NMR spectrum, as described in Chapter 4.

The method used for the quantitative definition of pseudo-components based on the

NMR fingerprinting and the diffusion coefficients measured by ^{13}C PFG NMR experiments was also adopted from Chapter 4. From this, the composition of each poorly specified mixture was obtained, cf. Appendix E for details. Three mixtures were considered in the present chapter as examples (details are given below). Their composition was known from gravimetric preparation, but that information was not used in the study except for a comparison of the results. Hence, they are labeled as poorly specified mixtures in the following.

6.2.3 Simulation of Liquid-liquid Extraction Processes

The studied poorly specified mixtures were considered as feeds of a single-stage liquid-liquid extraction described with an equilibrium-stage model. The task was to compare the performance of different extracting agents for a constant feed/solvent mass ratio. The process can be considered either a batch process or a continuous process. Eight common solvents from different chemical classes were considered as extracting agents; the essential requirement was that they exhibit a miscibility gap with water, which was the solvent in all studied mixtures, at the considered temperature, which was T = 298.15 K. The partitioning of all (pseudo-)components in thermodynamic equilibrium was calculated based on the isoactivity criterion:

$$x_i{}' \ \gamma_i{}'(T, \boldsymbol{x}') = x_i{}'' \ \gamma_i{}''(T, \boldsymbol{x}''); \quad i = 1...N \tag{13}$$

where $x_i{}'$ and $x_i{}''$ are the mole fraction of (pseudo-)component i in the raffinate (water-rich phase$'$) and the extract (extracting agent-rich phase$''$). $\gamma_i{}'$ and $\gamma_i{}''$ are the activity coefficients in the raffinate and extract phase, respectively, which were calculated using UNIFAC [31, 127] based on the estimated composition for the poorly specified mixtures. For comparison only, all calculations were also carried out using the true composition of the studied mixtures. Numerical details for the calculations are given in Appendix E.

6.2.4 Prediction of Residue Curves

Residue curves are often used to model single-stage open evaporations, where the vapor phase is continuously removed. The residue curve thereby describes the composition of the liquid phase over time, which is considered to be always in thermodynamic equilibrium with the removed vapor phase [130]. This process is described by the Rayleigh-equation [131, 132]. The pressure was 1 bar in all calculations in the present study.

In all cases, the vapor phase was treated as a mixture of ideal gases. Furthermore, the pressure dependence of the chemical potential in the liquid phase was always neglected.

The vapor-liquid equilibrium was therefore modeled by extended Raoult's law:

$$p_i^{\mathrm{S}}(T)\ x_i\ \gamma_i(T,\boldsymbol{x}) = p\ y_i = p_i; \quad i = 1 \ldots N \tag{14}$$

where p_i^{S} is the vapor pressure of the pure component i, x_i and y_i are the mole fraction of i in the liquid and vapor phase in equilibrium, respectively, and p_i is the partial pressure of component i in the vapor phase. The activity coefficients γ_i in the liquid phase were again calculated with UNIFAC [31, 127].

For the fully specified mixtures, which were modeled as a reference here only, the vapor pressure of the pure components p_i^{S} was calculated using the Antoine equation with parameters taken from the Dortmund Data Bank (DDB) [133]. Some of the considered components have very low vapor pressures, so that no data were available. In these cases, the vapor pressure of the component was set to zero, cf. Appendix E for details.

For the poorly specified mixtures, the vapor pressure p_k^{S} of the pseudo-component k was estimated using the group-contribution method of Refs. [27, 29]. If the estimated vapor pressure was $p_k^{\mathrm{S}} < 10^{-10}$ bar at $T = 373$ K, it was set to zero. A detailed description of the calculation of the vapor pressures is given in Appendix E.

6.3 Overview of Studied Mixtures

Table 13 gives an overview of the three mixtures that were considered as feed mixtures in the application examples studied in the present chapter. The information on the feed mixtures was adopted from Ref. [128] - no NMR experiments were carried out in the present chapter. All mixtures are aqueous solutions that contain a cocktail of diluted substances (the mole fractions of all other components were below 0.038 mol/mol) - a situation that is often encountered, e.g., in biotechnological processes. It is emphasized that the general methodology proposed here is not limited to such mixtures, and it is desirable to demonstrate the applicability for other types of poorly specified mixtures in future work.

The fact that diluted aqueous mixtures were studied has several consequences for the application of the methodology. Firstly, it explains why the study was carried out with ^{13}C NMR spectroscopy. Secondly, the high dilution poses challenges for spectroscopy, which can, however, be overcome by sufficiently long measuring times. On the other hand, the high dilution in the solvent water facilitates the application of the SEGWE model for predicting diffusion coefficients at infinite dilution in a known solvent. As the concentration dependence of diffusion coefficients can be very strong, it was not neglected, despite the low concentrations of the solutes. To account for it, a reference component was used. For details, see Ref. [128]. The reference components were randomly chosen in the three studied mixtures.

Table 13: Overview of the mixtures studied in this chapter. All mixtures are diluted aqueous solutions, the solvent water is not explicitly included in the table. x_i and M_i are the mole fraction and the molar mass of the true components i, respectively, which were known from the preparation of the samples; x_i^{pred} and M_i^{pred} are corresponding numbers for the pseudo-components $\tilde{\text{U}}$ predicted based on NMR fingerprinting. The dashed lines indicate which true components were lumped into the pseudo-components by the approach described in Chapter 4. Target components are labeled with (T), see text.

Mixture	Components i	M_i / g mol^{-1}	M_i^{pred} / g mol^{-1}	x_i / mol mol^{-1}	x_i^{pred} / mol mol^{-1}
I	2-propanol (T)	60.10	-	0.033	0.034
	acetone ($\tilde{\text{U}}_1$)	58.08	48.98	0.038	0.046
	1,4-butanediol ($\tilde{\text{U}}_2$)	90.12	145.21	0.035	0.022
	acetic acid ($\tilde{\text{U}}_3$)	60.05	72.11	0.033	0.029
II	1,4-dioxane (T)	88.11	-	0.006	0.006
	cyclohexanone ($\tilde{\text{U}}_1$)	98.15	94.47	0.012	0.012
	glucose ($\tilde{\text{U}}_2$)	180.16	212.08	0.015	0.034
	citric acid ($\tilde{\text{U}}_2$)	192.12		0.025	
III	acetonitrile (T)	41.05	-	0.022	0.022
	acetone ($\tilde{\text{U}}_1$)	58.08	51.71	0.016	0.017
	acetic acid ($\tilde{\text{U}}_2$)	60.05	71.06	0.015	0.013
	1-propanol ($\tilde{\text{U}}_3$)	60.10	85.26	0.015	0.021
	2-propanol ($\tilde{\text{U}}_3$)	60.10		0.015	
	cyclohexanone ($\tilde{\text{U}}_4$)	98.15	115.96	0.009	0.007
	1,4-butanediol ($\tilde{\text{U}}_5$)	90.12	148.90	0.010	0.006
	malic acid ($\tilde{\text{U}}_6$)	134.09	248.36	0.008	0.009
	xylose ($\tilde{\text{U}}_6$)	150.13		0.007	
	ascorbic acid ($\tilde{\text{U}}_7$)	176.12	313.85	0.006	0.004

In the application examples that were studied here, it was simply assumed that the reference component is the component to be separated from the unknown mixture. This is motivated by the fact that in separation processes with poorly specified mixtures, the target component to be separated is always known - and also its concentration in the feed mixture is usually known. As the focus of the present study is on the application side, the reference component is labeled with T (target) in the following, see Table 13.

The fact that a priori information on the solvent and a reference component was used does not indicate a general limitation of the methodology: if this information is not directly available, samples of poorly specified mixtures can, in general, be diluted with a known inert solvent and also a reference component can be added gravimetrically.

Additional results for the quantification of the structural groups based on NMR fingerprinting are presented in Appendix E.

The results presented in Table 13 show that the NMR fingerprinting in combination with the pseudo-component method works well. The true compositions of the mixtures and the respective predicted compositions agree well in most cases. There are considerable discrepancies in the predicted molar masses of some components that are, however, mainly caused by deficiencies of the SEGWE [77, 78] model, as shown in Ref. [128]. Better methods for predicting diffusion coefficients would, hence, be highly desirable for the present methodology. Still, the good agreement of the predicted and the true compositions shows that the method works well also with the SEGWE model. It is noted that for the components labeled with ($\tilde{\mathrm{U}}$) in Table 13, no information on the true nature and concentration was used to determine the predicted values, whereas, as explained above, for the target components (T), the mass fraction and the nature was assumed to be known. Still, as the predicted molar masses are used for the prediction of the mole fractions, also the mole fraction of the target component T is a prediction, albeit a much simpler one than for the other components $\tilde{\mathrm{U}}$. Since water and the target component were assumed to be known here, they were treated as true components in the calculations (e.g., using Antoine parameters from the DDB [133] for calculating the vapor pressure).

6.4 Results and Discussion

6.4.1 Solvent Screening for Liquid-liquid Extraction Processes

The simulation study of liquid-liquid extraction processes with poorly specified feeds that was carried out in the present chapter represents the scenario of a solvent screening. The aim of the process is the removal of a known target component T from a poorly specified aqueous mixture by an extracting agent E, which has to be selected from a list of candidates. For simplicity, both the temperature and the mass ratio of the extracting agent and the feed were kept constant and set to T = 298.15 K and $m_\mathrm{E}/m_\mathrm{F} = 5$, respectively. However, the approach easily generalizes to other temperatures and mass ratios. Eight common solvents that have a wide miscibility gap with water were considered as extracting agents: 1-octanol, 1-decanol, toluene, hexane, decane, dipropyl ether, ethyl propionate, and 3-octanone. The extension to other extracting agents is straightforward.

The selectivity is quantified by the separation factor α_{T} of the target component, which compares the distribution of the target component on the coexisting phases in thermodynamic equilibrium with the respective distributions of all identified pseudo-components $\tilde{\mathrm{U}}_k$:

$$\alpha_{\mathrm{T}} = \frac{\left(\frac{x''_{\mathrm{T}}}{x'_{\mathrm{T}}}\right)}{\left(\frac{\sum_{k=1}^{K} x''_{\tilde{\mathrm{U}}_k}}{\sum_{k=1}^{K} x'_{\tilde{\mathrm{U}}_k}}\right)} \tag{15}$$

where the double prime indicates the extract phase and the prime denotes the raffinate phase. The numerator is the partition coefficient of the target component and the denominator is the partition coefficient of all other components (except water) lumped together. The yield is:

$$Y_{\mathrm{T}} = \frac{n''_{\mathrm{T}}}{n'_{\mathrm{T}} + n''_{\mathrm{T}}} \tag{16}$$

where n_{T} refers to the mole number of the target component. For comparison, these quantities were also calculated for the fully specified mixtures, i.e., using the complete knowledge of the composition of the mixtures, cf. Table 13. These results are called ground truth in the following.

In Figure 32, the results for the extraction of the target component T=2-propanol from mixture I (cf. Table 13), are shown. A very good agreement between the predictions (open symbols) and the ground truth (closed symbols) is found for all extracting agents; this holds for both the separation factor α_{T} (Figure 32 left) and the extraction yield Y_{T} (Figure 32 right).

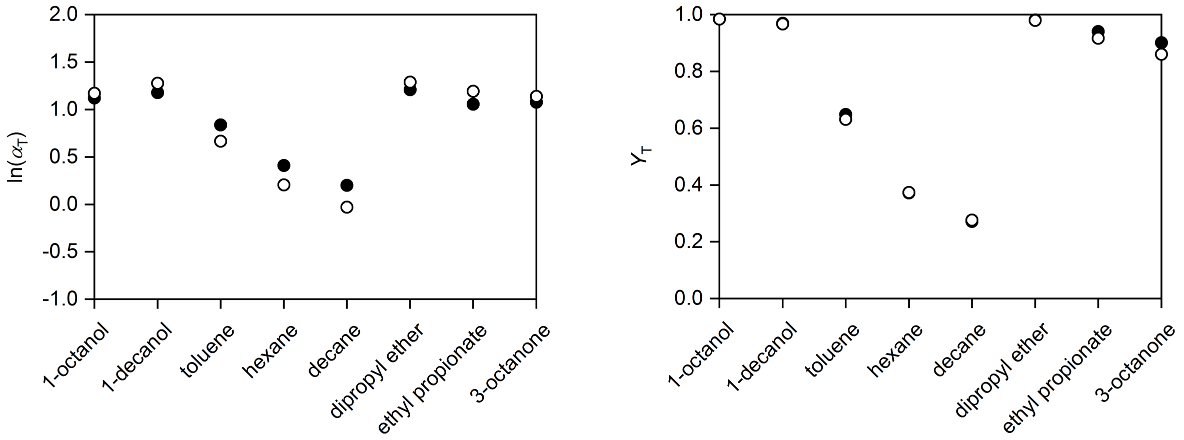


Figure 32: Separation factor α_T (left) and extraction yield Y_T (right) of target component T=2-propanol in a single-stage liquid-liquid extraction process with mixture I (cf. Table 13), as feed F for different extracting agents E at T = 298.15 K and $m_\mathrm{E}/m_\mathrm{F}$ = 5. Closed symbols: results for the fully specified feed. Open symbols: predictions for the poorly specified feed using NMR fingerprinting and the pseudo-component method. In some cases, the prediction is so good that the two symbols become indistinguishable.

Figure 33 shows the corresponding results for extracting the target component T=1,4-dioxane from mixture II. Again, for both the separation factor α_T (Figure 33 left) and the extraction yield Y_T (Figure 33 right), excellent agreement between the predictions (open symbols) and the ground truth (closed symbols) is found for all extracting agents. Note that the pseudo-component method could not distinguish glucose and citric acid here, which were lumped into a single pseudo-component ($\tilde{\mathrm{U}}_2$), cf. Table 13. The fact that, nevertheless, good results were obtained demonstrates that the approach is robust with respect to the definition of the pseudo-components. It is assumed that this is because the clustering of the identified structural groups is based on physics, namely on information on self-diffusion coefficients. While glucose and citric acid are chemically different, they differ not so much regarding their interactions with water and their molar mass.

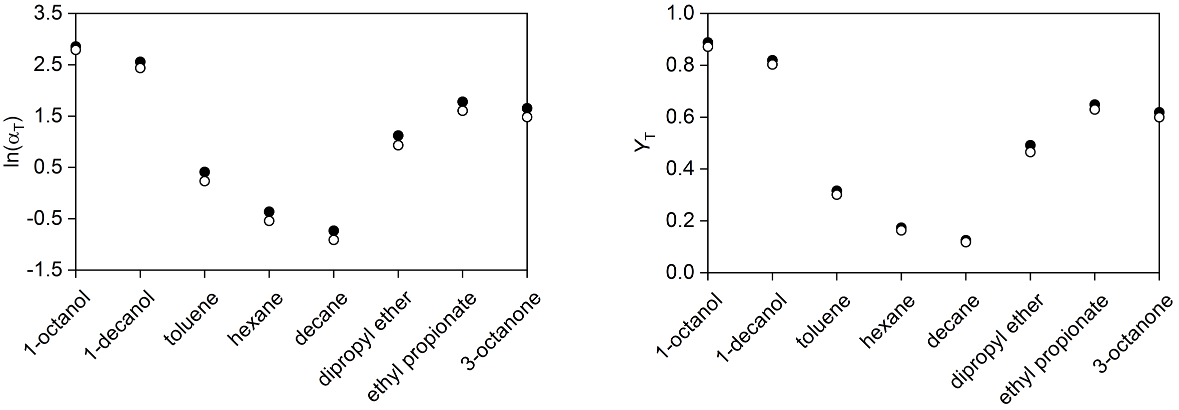


Figure 33: Separation factor α_T (left) and extraction yield Y_T (right) of target component T=1,4-dioxane in a single-stage liquid-liquid extraction process with mixture II (cf. Table 13), as feed F for different extracting agents E at $T = 298.15$ K and $m_E/m_F = 5$. Closed symbols: results for the fully specified feed. Open symbols: predictions for the poorly specified feed using NMR fingerprinting and the pseudo-component method.

In Figure 34, the results for mixture III with the target component T=acetonitrile are shown. For both the separation factor α_T (Figure 34 left) and the extraction yield of the target component Y_T (Figure 34 right), excellent agreement between the ground truth and the predictions is found for all studied extracting agents. In some cases, the differences are even indistinguishable in Figure 34. This is particularly interesting since the mixture contains nine components that were assumed to be unknown, which were lumped into a total of seven pseudo-components by the algorithms, cf. Table 13. Again, this shows the robustness of the approach.

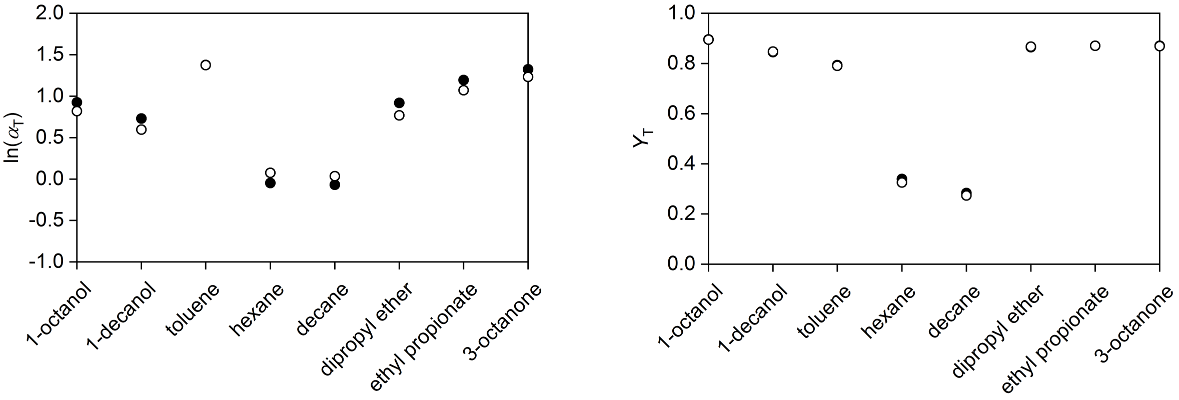


Figure 34: Separation factor α_T (left) and extraction yield Y_T (right) of target component T=acetonitrile in a single-stage liquid-liquid extraction process with mixture III (cf. Table 13), as feed F for different extracting agents E at $T = 298.15$ K and $m_E/m_F = 5$. Closed symbols: results for the fully specified feed. Open symbols: predictions for the poorly specified feed using NMR fingerprinting and the pseudo-component method. In some cases, the prediction is so good that the two symbols become indistinguishable.

6.4.2 Simulation of Open Evaporation Processes

In the following, results for an open evaporation process are shown. The results are presented over the evaporation ratio β:

$$\beta = 1 - \frac{n^{\mathrm{L}}}{n^{\mathrm{L},0}} \tag{17}$$

where n^{L} is the number of moles in the residue and $n^{\mathrm{L},0}$ the initial value of that property. The simulations were performed until the number of moles n^{L} in the residue approached the total amount of non-volatile components; see Appendix E for details.

Figure 35 shows the boiling temperature T of the three test mixtures during the open evaporation process at $p = 1$ bar. The overall agreement of the calculations for the fully specified mixtures (solid lines) with the predictions based on NMR fingerprinting and the pseudo-component method (dashed lines) is good in all cases. However, for mixture I, the boiling temperature is underestimated for low evaporation ratios, which is presumably because the vapor pressure of $\tilde{\mathrm{U}}_1$ (representing acetone, cf. Table 13) is overestimated. The influence of $\tilde{\mathrm{U}}_1$ on the boiling temperature is then reduced due to its fast evaporation, cf. also Figure 36. In Appendix E, the predicted vapor pressures of the pseudo-components based on the group-contribution method used here [27, 29] are compared to the vapor pressures as calculated by the Antoine equation. In general, considerable differences were obtained between these two calculations, which is presumably a two-fold problem: first, the prediction accuracy of the molar masses is not sufficient in some cases, and second, the group-contribution method for the prediction of the vapor pressure is not accurate enough. However, the impact of these false predictions is moderate as the order in which the components evaporate is correctly described in most cases, cf. Figures E.1-E.3 in Appendix E, and the qualitative behavior of the vapor pressure curves are well predicted.

Furthermore, in mixtures I and III, relatively high temperatures are predicted for high evaporation ratios; in mixture I, this is mainly caused by pseudo-component $\tilde{\mathrm{U}}_2$ (representing 1,4-butanediol, cf. Table 13), for which the vapor pressure is underestimated, cf. Figure E.1 in Appendix E. In mixture III, for high evaporation ratios, the non-volatile components malic acid, xylose, and ascorbic acid accumulate in the liquid phase.

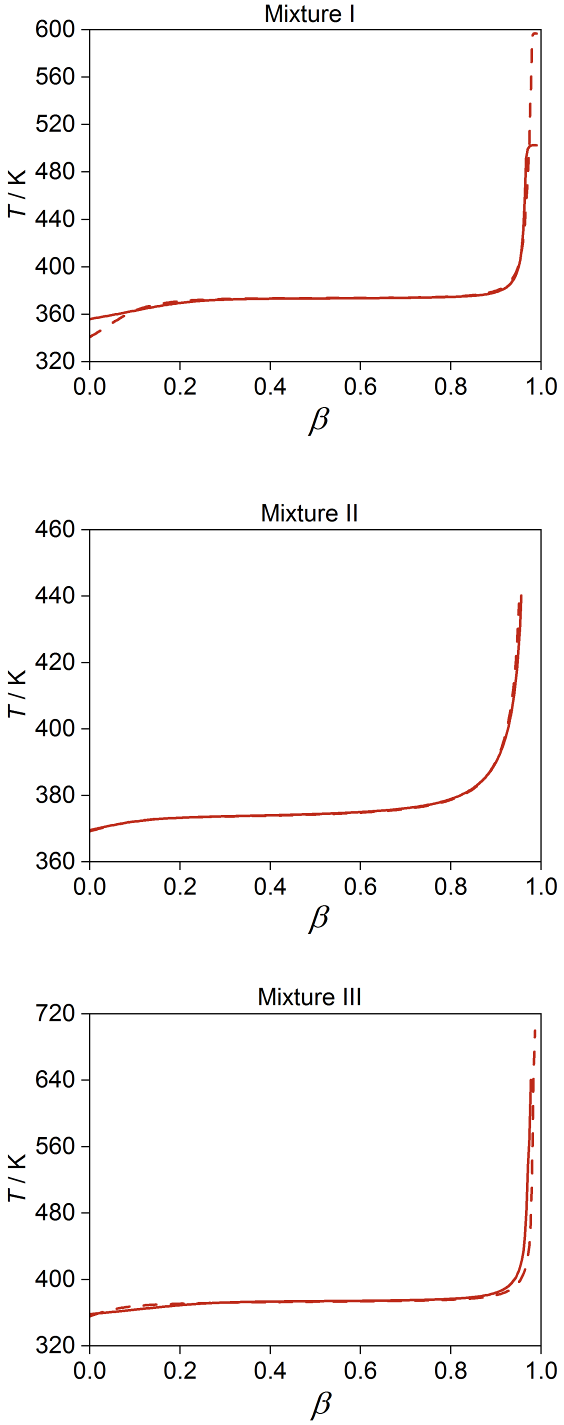


Figure 35: Results from simulations of open evaporation processes at 1 bar for mixtures I - III (cf. Table 13). The boiling temperature T is shown as a function of the evaporation ratio β. Solid lines: results for the fully specified mixtures. Dashed lines: predictions for the poorly specified mixtures based on NMR fingerprinting and the pseudo-component method.

Figure 36 shows the residue curves, i.e., the mole fractions of the components in the liquid phase over the course of the process, for mixture I at p = 1 bar (cf. Table 13). In the left panel, results for the fully specified mixture are shown, the right panel shows the corresponding results obtained from NMR fingerprinting and the pseudo-component method. Overall, the predictions agree well with the results obtained using the complete speciation of the feed. The imperfect prediction of the composition of the feed (at β = 0) is mainly caused by errors in the prediction of the molar masses. For instance, the underestimation of the molar mass of acetone (represented by $\tilde{\mathrm{U}}_1$) leads to an overestimation of its mole fraction. Similarly, the mole fraction of acetic acid is underestimated because its molar mass is overestimated.

The maximum in the residue curve of acetic acid is also found for the pseudo-component $\tilde{\mathrm{U}}_3$, which represents acetic acid, but the maximum is by far overpredicted. This is presumably a consequence of the overprediction of the molar mass of acetic acid and problems of the group-contribution method used to predict the vapor pressure of that component (see Figure E.1 in Appendix E).

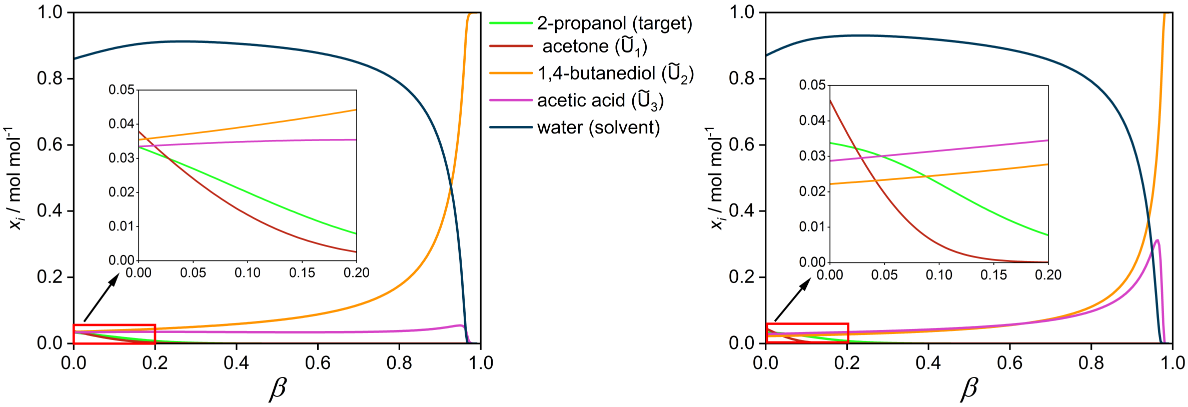


Figure 36: Residue curves showing the liquid-phase mole fractions x_i as a function of the evaporation ratio β for mixture I (cf. Table 13), at p = 1 bar. Left: results obtained using the full speciation. Right: predictions based on NMR fingerprinting and the pseudo-component method.

Figure 37 shows the respective results for the residue curves for mixture II. The agreement between the predicted residue curves and the ones calculated using the complete information on the composition of the mixture is excellent for all components. Citric acid and glucose were lumped into a single pseudo-component ($\tilde{\mathrm{U}}_2$), cf. Table 13; therefore, only the sum of the mole fractions of both components is indicated in both panels of Figure 37. A representation with individual mole fractions is shown in Figure E.4 in Appendix E.

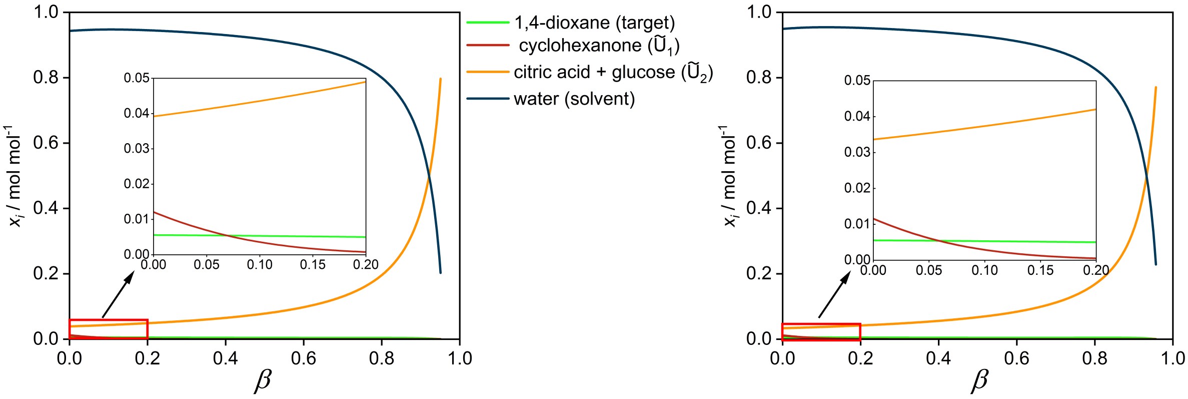


Figure 37: Residue curves showing the liquid-phase mole fractions x_i as a function of the evaporation ratio β for mixture II (cf. Table 13), at p = 1 bar. Left: results obtained using the full speciation. Right: predictions based on NMR fingerprinting and the pseudo-component method.

It is noted that the concentrations of citric acid and glucose in the liquid phase for high evaporation ratios are so high that, in practice, they would precipitate, i.e., an additional solid-liquid equilibrium (SLE) would occur, which, however, was not considered here.

Figure 38 shows the respective residue curves for mixture III. A good agreement between the predicted residue curves and those obtained using the full speciation is found for most components. While the maximum in the concentration curve of 1,4-butanediol ($\tilde{U}_5$) is well predicted, as for mixture I, poor results are obtained for acetic acid ($\tilde{U}_2$), for which a strong maximum is predicted, which is not found when the full speciation is used. The reasons are the same as for mixture I: the poor prediction of the molar mass and the vapor pressure. 1-propanol and 2-propanol were lumped into a single pseudo-component ($\tilde{U}_3$); the same holds for malic acid and xylose ($\tilde{U}_6$), cf. Table 13. Therefore, only the sums of the mole fractions of the respective true components are plotted in these cases, cf. Figure E.5 in Appendix E for a plot of the individual concentrations. The lumping of 1-propanol and 2-propanol is a good example for a lumping that is uncritical for the prediction of thermophysical properties. Also, the lumping of the two high-boilers, malic acid and xylose, does not substantially affect the prediction of residue curves. As with mixture II, the possible occurrence of an SLE was not considered here, although it might occur in practice for high evaporation ratios.

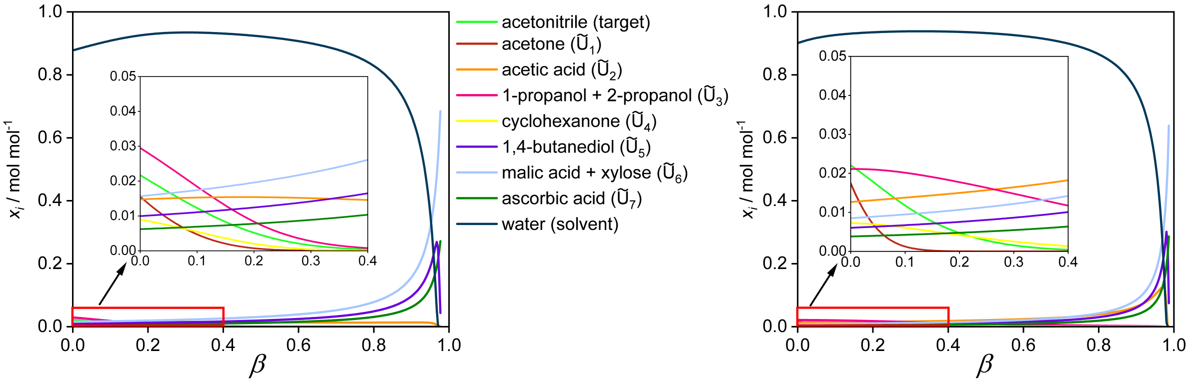


Figure 38: Residue curves showing the liquid-phase mole fractions x_i as a function of the evaporation ratio β for mixture III (cf. Table 13), at p = 1 bar. Left: results obtained using the full speciation. Right: predictions based on NMR fingerprinting and the pseudo-component method.

6.5 Conclusions

In this chapter, it was demonstrated that predictive thermodynamic modeling can be achieved without knowing the full speciation of the mixtures based on the characterization of the mixture described in Chapter 4.

The approach was only applied to aqueous mixtures, but there is no reason why it could not be applied to non-aqueous solutions. On the contrary: for non-aqueous solutions, it could be attractive to combine the information from ^{13}C NMR spectroscopy that was used here with that from ^{1}H NMR spectroscopy, which could further improve the results. Also, variants based solely on ^{1}H NMR spectroscopy could be developed.

Furthermore, only mixtures were studied in which the components were highly diluted in the solvent (always water here). This is no prerequisite for applying the NMR fingerprinting and the definition of pseudo-components, but the absence of a component that is present in large excess troubles the determination of the molar mass of the pseudo-components. If such an excess component is not present, it can be added. The prerequisite for this would be that the added component is miscible with the poorly specified mixture and non-reactive. Furthermore, it is desirable to have one component of which the concentration is known. This will be the case in most practical problems with poorly specified mixtures, e.g., the concentration of a product in an otherwise poorly specified mixture. If this was not the case, such a component could be added to the mixture. In the present chapter, this reference component was designated.

The method depends on the quality of prediction methods that are needed at two points: firstly, the SEGWE model for predicting diffusion coefficients is used to get the information on the molar mass of the pseudo-components, and secondly, group-contribution methods are used to predict the thermodynamic properties of interest. The quality of the latter methods directly limits what can be achieved with the present approach. Luckily, many suitable thermodynamic group-contribution methods are available for a large variety of properties. Applying these methods in the present framework requires mapping the groups that can be identified by NMR spectroscopy to groups considered in the thermodynamic method. This may require some case-specific adaptions but will hardly pose fundamental problems. It is emphasized that also the flexibility on the NMR spectroscopic side can be used for such adaptions and that it can be expected to see progress in the group assignment in NMR spectroscopy by using machine learning [68].

The results from the present examples also show that even the best available semi-empirical methods for predicting diffusion coefficients [109] (the SEGWE model [77, 78]) and a well-developed method for predicting pure-component vapor pressures (the method of Refs. [27, 29]) have critical deficiencies. Some of the most critical deviations observed in comparing the predictions to those obtained using the full speciation resulted from deficiencies of these models and not directly from the methodology presented here, which will profit from any future improvement of the group-contribution methods.

The present results were obtained based on NMR experiments with cryogenic high-field instruments. In future work, the methodology should be adapted to use results from much simpler and cheaper benchtop NMR spectrometers. Furthermore, the present chapter demonstrates that the approach developed in Chapter 4 to solve the problem of modeling poorly specified mixtures is broadly applicable. It is also flexible and can be tuned to specific needs. More studies should follow to elucidate and demonstrate the full potential of the new approach.

7 Conclusions

Complex mixtures with unknown composition are ubiquitous in fields such as biotechnology, refinery technology, waste-water treatment, or in renewable feedstocks. The design of efficient processes involving such poorly specified mixtures is hindered by the fact that, basically, all thermodynamic models require information about the composition of the mixture. In the present thesis, a framework for addressing this challenge is introduced, including methods for group-specific characterization of poorly specified mixtures, the rational definition of pseudo-components, and the application for thermodynamic modeling and simulation.

Two methods for automatic identification and quantification of structural groups were developed based on the machine-learning concept of support vector classification. The first method enables the assignment of thirteen structural groups to spectral regions in ^{13}C NMR using additional knowledge on ^{1}H NMR. While the method was only trained on pure-component data, it was successfully applied to the analysis of mixtures. The second method, which is also provided in the form of an interactive website (`https://www.nmr-fingerprinting.de`) enables incorporating knowledge from ^{13}C DEPT NMR spectra, which yield information about the substitution degree of the carbon atoms. Furthermore, knowledge about the presence of labile protons is included, further improving the results. Additionally, an automatic workflow for the fully automatic training of the method was introduced by using so-called SMILES arbitrary target specification (SMARTS) strings that are based on the simplified molecular-input line-entry system (SMILES) that encodes components using simple text strings. SMARTS not only enables a rigorous definition of structural groups, it furthermore delivers great flexibility for adaptation of the methods by introducing new groups or changing existing ones.

Furthermore, an approach for the rational definition of pseudo-components based on the group-specific characterization of a mixture was developed. The approach is based on the measurement of self-diffusion coefficients using ^{13}C PFG NMR. The basic idea is simple: structural groups that are part of the same molecule should show the same diffusion behavior. Due to uncertainties in the data, determining the number of pseudo-components based on such data is still tedious, which is why the unsupervised K-medians algorithm was used to cluster the structural groups into multiple pseudo-components. To evaluate the quality of the clustering and find the beforehand unknown number of pseudo-components, the silhouette score was used, which is a common metric for such a task and measures the overall consistency of the definition of the pseudo-components. Information on the self-diffusion coefficients was also used to determine the molar mass of the pseudo-components.

The developed methods in the thesis enable the thermodynamic modeling of poorly specified mixtures, which was demonstrated by calculating quantum-chemical descriptors of the unknown part of the mixture as well as activity coefficients. Furthermore, two processes were simulated here, namely the solvent screening for a poorly specified mixture that requires modeling liquid-liquid equilibria and open evaporation processes, for which vapor-liquid equilibria were modeled. For both types of processes, excellent results were obtained.

The methods developed in the present thesis can be extended in many ways. The structural groups considered here cover only the elements C, H, and O. Incorporating additional elements would be highly interesting, but presumably requires additional data, e.g., from 2D NMR methods like the heteronuclear multiple bond correlation (HMBC) experiment or information from elemental analysis. However, since there is no comprehensive database for 2D NMR spectra, synthetic data generation would be highly needed, e.g., by using DFT calculations of NMR spectra or using machine-learning methods. Furthermore, the NMR fingerprinting approach and the pseudo-component method should be unified in future work, to share information among the approaches.

The ^{13}C PFG NMR experiments were performed here with high-field spectrometers. In future work, mobile benchtop spectrometers could be considered, which are easier to handle and would facilitate the application of the new framework directly at process plants. To compensate for the reduced sensitivity, new PFG NMR methods, e.g., based on polarization enhancement nurtured during attached nucleus testing (PENDANT) [97], could be developed and applied. As an alternative, also so-called “pure shift” NMR methods [134, 135], or more formal “homonuclear broadband decoupling” methods could be used, which simplify the ^{1}H NMR spectrum by suppressing the multiplet structure and are therefore particularly interesting to be combined with benchtop spectrometers in the proposed framework.

Literature

[1] D. T. Allen, M. R. Gray, T. T. Le: Structural characterization and thermodynamic property estimation for wood tars: a functional group approach, Liq. Fuels Technol. 2 (1984) 327–353. DOI:`10.1080/07377268408915356`.

[2] G. L. Alexander, B. J. Schwarz, J. M. Prausnitz: Phase equilibriums for high-boiling fossil fuel distillates. 2. Correlation of equation-of-state constants with characterization data for phase equilibrium calculations, Ind. Eng. Chem. Fundam. 24 (1985) 311–315. DOI:`10.1021/i100019a006`.

[3] B. Carreón-Calderón, V. Uribe-Vargas, E. Ramírez-Jaramillo, M. Ramírez-de Santiago: Thermodynamic characterization of undefined petroleum fractions using group contribution methods, Ind. Eng. Chem. Res. 51 (2012) 14188–14198. DOI:`10.1021/ie3016076`.

[4] C. A. Jackson, W. J. Simonsick: Application of mass spectrometry to the characterization of polymers, Curr. Opin. Solid State Mater. Sci. 2 (1997) 661–667. DOI:`10.1016/S1359-0286(97)80006-X`.

[5] M. Malik, J. Mays, M. R. Shah (Eds.): Molecular characterization of polymers, Elsevier, 2021. DOI:`10.1016/c2019-0-00391-3`.

[6] M. C. Cuellar, A. J. Straathof: Downstream of the bioreactor: advancements in recovering fuels and commodity chemicals, Curr. Opin. Biotechnol. 62 (2020) 189–195. DOI:`10.1016/j.copbio.2019.11.012`.

[7] A. Guellil, F. Thomas, J.-C. Block, J.-L. Bersillon, P. Ginestet: Transfer of organic matter between wastewater and activated sludge flocs, Water Res. 35 (2001) 143–150. DOI:`10.1016/S0043-1354(00)00240-2`.

[8] W. J. Sim, T. E. Daubert: Prediction of vapor-liquid equilibria of undefined mixtures, Ind. Eng. Chem. Process Des. Dev. 19 (1980) 386–393. DOI:`10.1021/i260075a010`.

[9] G. L. Alexander, A. L. Creagh, J. M. Prausnitz: Phase equilibriums for high-boiling fossil fuel distillates. 1. Characterization, Ind. Eng. Chem. Fundam. 24 (1985) 301–310. DOI:`10.1021/i100019a005`.

[10] C. F. Leibovici: A consistent procedure for the estimation of properties associated to lumped systems, Fluid Phase Equilib. 87 (1993) 189–197. DOI:`10.1016/0378-3812(93)85026-I`.

[11] M. A. Fahim, A. S. Elkilani: Prediction of solubility of hydrogen in petroleum cuts using modified UNIFAC, Can. J. Chem. Eng. 70 (1992) 335–340. DOI:`10.1002/cjce.5450700218`.

[12] V. J. Pereira, V. B. Regueira, G. M. N. Costa, S. A. B. Vieira de Melo: Modeling the saturation pressure of systems containing crude oils and CO_2 using the SRK equation of state, J. Chem. Eng. Data 64 (2019) 2134–2142. DOI:`10.1021/acs.jced.8b01077`.

[13] Z. Gao, Z. Xu, S. Zhao, L. Zhang: Heavy petroleum supercritical fluid deasphalting process simulation based on the saturate, aromatic, resin, and asphaltene composition, Energy Fuels 36 (2022) 8818–8827. DOI:`10.1021/acs.energyfuels.2c00891`.

[14] M. T. Rätzsch, H. Kehlen: Continuous thermodynamics of polymer solutions: the effect of polydispersity on the liquid-liquid equilibrium, J. Macromol. Sci. Chem. 22 (1985) 323–334. DOI:`10.1080/00222338508056606`.

[15] M. T. Rätzsch: Continuous thermodynamics, Pure Appl. Chem. 61 (1989) 1105–1114. DOI:`10.1351/pac198961061105`.

[16] M. Spraul, B. Schütz, E. Humpfer, M. Mörtter, H. Schäfer, S. Koswig, P. Rinke: Mixture analysis by NMR as applied to fruit juice quality control, Magn. Reson. Chem. 47 (2009) 130–137. DOI:`10.1002/mrc.2528`.

[17] J. S. McKenzie, J. A. Donarski, J. C. Wilson, A. J. Charlton: Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics, Prog. Nucl. Magn. Reson. Spectrosc. 59 (2011) 336–359. DOI:`10.1016/j.pnmrs.2011.04.003`.

[18] J. van Duynhoven, E. van Velzen, D. M. Jacobs: Chapter three-quantification of complex mixtures by NMR, in: Annual reports on NMR spectroscopy, Academic Press, 2013, pp. 181–236. DOI:`10.1016/b978-0-12-408097-3.00003-2`.

[19] E. Schievano, M. Tonoli, F. Rastrelli: NMR quantification of carbohydrates in complex mixtures. a challenge on honey, Anal. Chem. 89 (2017) 13405–13414. DOI:`10.1021/acs.analchem.7b03656`.

[20] R. Behrens, E. Kessler, K. Münnemann, H. Hasse, E. von Harbou: Monoalkylcarbonate formation in the system monoethanolamine–water–carbon dioxide, Fluid Phase Equilib. 486 (2019) 98–105. DOI:`10.1016/j.fluid.2018.12.031`.

[21] A. Fröscher, K. Langenbach, E. von Harbou, W. R. Thiel, H. Hasse: NMR spectroscopic study of chemical reactions in mixtures containing oleic acid, formic acid, and formoxystearic acid, Ind. Eng. Chem. Res. 58 (2019) 5622–5630. DOI:`10.1021/acs.iecr.8b05715`.

[22] D. C. Burns, E. P. Mazzola, W. F. Reynolds: The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products, Nat. Prod. Rep. 36 (2019) 919–933. DOI:`10.1039/c9np00007k`.

[23] M. Elyashberg: Identification and structure elucidation by NMR spectroscopy, TrAC, Trends Anal. Chem. 69 (2015) 88–97. DOI:`10.1016/j.trac.2015.02.014`.

[24] A. Bakiri, J. Hubert, R. Reynaud, S. Lanthony, D. Harakat, J.-H. Renault, J.-M. Nuzillard: Computer-aided ^{13}C NMR chemical profiling of crude natural extracts without fractionation, J. Nat. Prod. 80 (2017) 1387–1396. DOI:`10.1021/acs.jnatprod.6b01063`.

[25] K. P. C. Vollhardt, N. E. Schore: Organic chemistry: structure and function, 8 ed., Macmillan Learning: New York, 2018.

[26] K. Joback, R. Reid: Estimation of pure component properties from group-contributions, Chem. Eng. Commun. 57 (1987) 233–243. DOI:`10.1080/00986448708960487`.

[27] Y. Nannoolal, J. Rarey, D. Ramjugernath, W. Cordes: Estimation of pure component properties: Part 1. Estimation of the normal boiling point of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilib. 226 (2004) 45–63. DOI:`10.1016/j.fluid.2004.09.001`.

[28] Y. Nannoolal, J. Rarey, D. Ramjugernath: Estimation of pure component properties: Part 2. Estimation of critical property data by group contribution, Fluid Phase Equilib. 252 (2007) 1–27. DOI:`10.1016/j.fluid.2006.11.014`.

[29] Y. Nannoolal, J. Rarey, D. Ramjugernath: Estimation of pure component properties: Part 3. Estimation of the vapor pressure of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilib. 269 (2008) 117–133. DOI:`10.1016/j.fluid.2008.04.020`.

[30] Y. Nannoolal, J. Rarey, D. Ramjugernath: Estimation of pure component properties. Part 4: estimation of the saturated liquid viscosity of non-electrolyte organic compounds via group contributions and group interactions, Fluid Phase Equilib. 281 (2009) 97–119. DOI:`10.1016/j.fluid.2009.02.016`.

[31] A. Fredenslund, R. L. Jones, J. M. Prausnitz: Group-contribution estimation of activity coefficients in nonideal liquid mixtures, AIChE J. 21 (1975) 1086–1099. DOI:`10.1002/aic.690210607`.

[32] D. Constantinescu, J. Gmehling: Further development of modified UNIFAC (Dortmund): revision and extension 6, J. Chem. Eng. Data 61 (2016) 2738–2748. DOI:`10.1021/acs.jced.6b00136`.

[33] M. Badertscher, P. Bühlmann, E. Pretsch: Structure determination of organic compounds, 4 ed., Springer: Berlin Heidelberg, 2009. DOI:`10.1007/978-3-540-93810-1`.

[34] A. Bagno, F. Rastrelli, G. Saielli: Toward the complete prediction of the ^{1}H and ^{13}C NMR spectra of complex organic molecules by DFT methods: application to natural substances, Chem. - Eur. J. 12 (2006) 5514–5525. DOI:`10.1002/chem.200501583`.

[35] W. Bremser: HOSE — a novel substructure code, Anal. Chim. Acta 103 (1978) 355–365. DOI:`10.1016/s0003-2670(01)83100-7`.

[36] Y.-Y. Du, G.-Y. Bai, X. Zhang, M.-L. Liu: Classification of wines based on combination of ^{1}H NMR spectroscopy and principal component analysis, Chin. J. Chem. 25 (2007) 930–936. DOI:`10.1002/cjoc.200790181`.

[37] S. Masoum, C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi, D. N. Rutledge: Application of support vector machines to ^{1}H NMR Data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins, Anal. Bioanal. Chem. 387 (2007) 1499–1510. DOI:`10.1007/s00216-006-1025-x`.

[38] J. L. Ward, C. Harris, J. Lewis, M. H. Beale: Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of Arabidopsis Thaliana, Phytochemistry 62 (2003) 949–957. DOI:`10.1016/s0031-9422(02)00705-7`.

[39] S. L. Howells, R. J. Maxwell, J. R. Griffiths: Classification of tumour ^{1}H NMR spectra by pattern recognition, NMR Biomed. 5 (1992) 59–64. DOI:`10.1002/nbm.1940050203`.

[40] C. Cobas: NMR signal processing, prediction, and structure verification with machine learning techniques, Magn. Reson. Chem. 58 (2020) 512–519. DOI:`10.1002/mrc.4989`.

[41] P. R. Filgueiras, N. A. Portela, S. R. C. Silva, E. V. R. Castro, L. M. S. L. Oliveira, J. C. M. Dias, A. C. Neto, W. Romão, R. J. Poppi: Determination of saturates, aromatics, and polars in crude oil by ^{13}C NMR and support vector regression with variable selection by genetic algorithm, Energy Fuels 30 (2016) 1972–1978. DOI:`10.1021/acs.energyfuels.5b02377`.

[42] M. K. Moro, Á. C. Neto, V. Lacerda, W. Romão, L. S. Chinelatto, E. V. Castro, P. R. Filgueiras: FTIR, ^{1}H and ^{13}C NMR data fusion to predict crude oils properties, Fuel 263 (2020) 116721. DOI:`10.1016/j.fuel.2019.116721`.

[43] S. H. Martínez-Treviño, V. Uc-Cetina, M. A. Fernández-Herrera, G. Merino: Prediction of natural product classes using machine learning and ^{13}C NMR spectroscopic data, J. Chem. Inf. Model. 60 (2020) 3376–3386. DOI:`10.1021/acs.jcim.0c00293`.

[44] C. L. Wilkins, T. L. Isenhour: Multiple discriminant function analysis of carbon-13 nuclear magnetic resonance spectra. Functional group identification by pattern recognition, Anal. Chem. 47 (1975) 1849–1851. DOI:`10.1021/ac60361a029`.

[45] C. L. Wilkins, T. R. Brunner: Classification of binary carbon-13 nuclear magnetic resonance spectra, Anal. Chem. 49 (1977) 2136–2141. DOI:`10.1021/ac50022a011`.

[46] J. A. Fine, A. A. Rajasekar, K. P. Jethava, G. Chopra: Spectral deep learning for prediction and prospective validation of functional groups, Chem. Sci. 11 (2020) 4618–4630. DOI:`10.1039/c9sc06240h`.

[47] C. M. Bishop: Pattern recogniction and machine learning, 1 ed., Springer: Berlin, 2006.

[48] K. Murphy: Machine learning : a probabilistic perspective, 1 ed., MIT Press: Cambridge, Mass., 2012.

[49] Spectral database of organic compounds (National Institute of Advanced Industrial Science and Technology), https://sdbs.db.aist.go.jp/sdbs/, Last accessed: 24.04.2023.

[50] U. Weidlich, J. Gmehling: A modified UNIFAC model. 1. Prediction of VLE, h^{E}, and γ^{∞}., Ind. Eng. Chem. Res. 26 (1987) 1372–1381. DOI:`10.1021/ie00067a018`.

[51] O. Luaces, J. Díez, J. Barranquero, J. J. del Coz, A. Bahamonde: Binary relevance efficacy for multilabel classification, Prog. Artif. Intell. 1 (2012) 303–313. DOI:`10.1007/s13748-012-0030-x`.

[52] J. Opitz, S. Burst: Macro F_1 and macro F_1, arXiv:1911.03347 (2019). DOI:`10.48550/arXiv.1911.03347`.

[53] G. C. Cawley, N. L. Talbot: On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107.

[54] M. Stone: Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Series B Stat. Methodol. 36 (1974) 111–133. DOI:`10.1111/j.2517-6161.1974.tb00994.x`.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay: Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[56] C.-C. Chang, C.-J. Lin: LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (2011) 1–27. DOI:`10.1145/1961189.1961199`.

[57] P. Szymański, T. Kajdanowicz: A scikit-based Python environment for performing multi-label classification, arXiv:1702.01460 (2017). DOI:`10.48550/arXiv.1702.01460`.

[58] K. Sechidis, G. Tsoumakas, I. Vlahavas: On the stratification of multi-label data, in: machine learning and knowledge discovery in databases, Springer Berlin Heidelberg, 2011, pp. 145–158. DOI:`10.1007/978-3-642-23808-6_10`.

[59] F. Jirasek, J. Burger, H. Hasse: Method for estimating activity coefficients of target components in poorly specified mixtures, Ind. Eng. Chem. Res. 57 (2018) 7310–7313. DOI:`10.1021/acs.iecr.8b00917`.

[60] F. Jirasek, J. Burger, H. Hasse: NEAT–NMR spectroscopy for the estimation of activity coefficients of target components in poorly specified mixtures, Ind. Eng. Chem. Res. 58 (2019) 9155–9165. DOI:`10.1021/acs.iecr.9b01269`.

[61] F. Jirasek, J. Burger, H. Hasse: Application of NEAT for determining the composition dependence of activity coefficients in poorly specified mixtures, Chem. Eng. Sci. 208 (2019) 115161. DOI:`10.1016/j.ces.2019.115161`.

[62] F. Jirasek, J. Burger, H. Hasse: Application of NEAT for the simulation of liquid–liquid extraction processes with poorly specified feeds, AIChE J. 66 (2020) e16826. DOI:`10.1002/aic.16826`.

[63] T. Specht, K. Münnemann, F. Jirasek, H. Hasse: Estimating activity coefficients of target components in poorly specified mixtures with NMR spectroscopy and COSMO-RS, Fluid Phase Equilib. 516 (2020) 112604. DOI:`10.1016/j.fluid.2020.112604`.

[64] Daylight Chemical Information Systems theory manual, version 4.9, https://www.daylight.com/dayhtml/doc/theory/index.html, Last accessed: 12.12.2022.

[65] D. Weininger: SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36. DOI:`10.1021/ci00057a005`.

[66] S. Kuhn, N. E. Schlörer: Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 - a free in-house NMR database with integrated LIMS for academic service laboratories, Magn. Reson. Chem. 53 (2015) 582–589. DOI:`10.1002/mrc.4263`.

[67] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, J. L. Markley: BioMagResBank, Nucleic Acids Res. 36 (2007) D402–D408. DOI:`10.1093/nar/gkm957`.

[68] T. Specht, K. Münnemann, H. Hasse, F. Jirasek: Automated methods for identification and quantification of structural groups from nuclear magnetic resonance spectra using support vector classification, J. Chem. Inf. Model. 61 (2021) 143–155. DOI:`10.1021/acs.jcim.0c01186`.

[69] T. D. Claridge (Ed.): High-resolution NMR techniques in organic chemistry, third ed., Elsevier, 2016. DOI:`10.1016/c2015-0-04654-8`.

[70] RDKit: Open-source cheminformatics., https://www.rdkit.org, Last accessed: 13.12.2022.

[71] Y. Guan, S. V. S. Sowndarya, L. C. Gallegos, P. C. S. John, R. S. Paton: Real-time prediction of ^{1}H and ^{13}C chemical shifts with DFT accuracy using a 3D graph neural network, Chem. Sci. 12 (2021) 12012–12026. DOI:`10.1039/d1sc03343c`.

[72] T. Head, M. Kumar, H. Nahrstaedt, G. Louppe, I. Shcherbatyi: scikit-optimize v0.9.0, https://scikit-optimize.github.io, Last accessed: 20.04.2023. DOI:`10.5281/ZENODO.5565057`.

[73] B. Carreón-Calderón, V. Uribe-Vargas, M. Ramírez-de Santiago, E. Ramírez-Jaramillo: Thermodynamic characterization of heavy petroleum fluids using group

contribution methods, Ind. Eng. Chem. Res. 53 (2014) 5598–5607. DOI:`10.1021/ie403967z`.

[74] M. Mahmudi, M. T. Sadeghi: A novel three pseudo-component approach (ThPCA) for thermodynamic description of hydrocarbon-water systems, J. Pet. Explor. Prod. Technol. 4 (2013) 281–289. DOI:`10.1007/s13202-013-0072-z`.

[75] A. G. Abdul Jameel, A. M. Elbaz, A.-H. Emwas, W. L. Roberts, S. M. 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 (2016) 3894–3905. DOI:`10.1021/acs.energyfuels.6b00303`.

[76] A. Raimondi, A. Favela-Contreras, F. Beltrán-Carbajal, A. Piñón-Rubio, J. Luis de la Peña-Elizondo: Design of an adaptive predictive control strategy for crude oil atmospheric distillation process, Control Eng. Pract. 34 (2015) 39–48. DOI:`10.1016/j.conengprac.2014.09.014`.

[77] R. Evans, Z. Deng, A. K. Rogerson, A. S. McLachlan, J. J. Richards, M. Nilsson, G. A. Morris: Quantitative interpretation of diffusion-ordered NMR spectra: can we rationalize small molecule diffusion coefficients?, Angew. Chem., Int. Ed. 52 (2013) 3199–3202. DOI:`10.1002/anie.201207403`.

[78] R. Evans, G. D. Poggetto, M. Nilsson, G. A. Morris: Improving the interpretation of small molecule diffusion coefficients, Anal. Chem. 90 (2018) 3987–3994. DOI:`10.1021/acs.analchem.7b05032`.

[79] C. D'Agostino, M. Mantle, L. Gladden, G. Moggridge: Prediction of mutual diffusion coefficients in non-ideal mixtures from pulsed field gradient NMR data: Triethylamine–water near its consolute point, Chem. Eng. Sci. 74 (2012) 105–113. DOI:`10.1016/j.ces.2012.02.025`.

[80] C. D'Agostino, J. Stephens, J. Parkinson, M. Mantle, L. Gladden, G. Moggridge: Prediction of the mutual diffusivity in acetone–chloroform liquid mixtures from the tracer diffusion coefficients, Chem. Eng. Sci. 95 (2013) 43–47. DOI:`10.1016/j.ces.2013.03.033`.

[81] D. Bellaire, H. Kiepfer, K. Münnemann, H. Hasse: PFG-NMR and MD simulation study of self-diffusion coefficients of binary and ternary mixtures containing cyclohexane, ethanol, acetone, and toluene, J. Chem. Eng. Data 65 (2020) 793–803. DOI:`10.1021/acs.jced.9b01016`.

[82] D. Bellaire, O. Großmann, K. Münnemann, H. 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, J. Chem. Thermodyn. 166 (2022) 106691. DOI:`10.1016/j.jct.2021.106691`.

[83] K. F. Morris, C. S. Johnson: Diffusion-ordered two-dimensional nuclear magnetic resonance spectroscopy, J. Am. Chem. Soc. 114 (1992) 3139–3141. DOI:`10.1021/ja00034a071`.

[84] K. F. Morris, P. Stilbs, C. S. Johnson: Analysis of mixtures based on molecular size and hydrophobicity by means of diffusion-ordered 2D NMR, Anal. Chem. 66 (1994) 211–215. DOI:`10.1021/ac00074a006`.

[85] G. Kapur, M. Findeisen, S. Berger: Analysis of hydrocarbon mixtures by diffusion-ordered NMR spectroscopy, Fuel 79 (2000) 1347–1351. DOI:`10.1016/S0016-2361(99)00271-9`.

[86] D. Li, R. Hopson, W. Li, J. Liu, P. G. Williard: ^{13}C INEPT diffusion-ordered NMR spectroscopy (DOSY) with internal references, Org. Lett. 10 (2008) 909–911. DOI:`10.1021/ol703039v`.

[87] S. Balayssac, S. Trefi, V. Gilard, M. Malet-Martino, R. Martino, M.-A. 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, J. Pharm. Biomed. Anal. 50 (2009) 602–612. DOI:`10.1016/j.jpba.2008.10.034`.

[88] G. Pagès, V. Gilard, R. Martino, M. Malet-Martino: Pulsed-field gradient nuclear magnetic resonance measurements (PFG NMR) for diffusion ordered spectroscopy (DOSY) mapping, Analyst 142 (2017) 3771–3796. DOI:`10.1039/c7an01031a`.

[89] D. Wu, A. Chen, C. S. Johnson, Jr.: Heteronuclear-detected diffusion-ordered NMR spectroscopy through coherence transfer, J. Magn. Reson., Ser. A 123 (1996) 215–218. DOI:`10.1006/jmra.1996.0239`.

[90] A. Botana, P. W. Howe, V. Caër, G. A. Morris, M. Nilsson: High resolution ^{13}C DOSY: the DEPTSE experiment, J. Magn. Reson. 211 (2011) 25–29. DOI:`10.1016/j.jmr.2011.03.016`.

[91] P. J. Rousseeuw: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. DOI:`10.1016/0377-0427(87)90125-7`.

[92] J. A. Jones, D. K. Wilkins, L. J. Smith, C. M. Dobson: Characterisation of protein unfolding by NMR diffusion measurements, J. Biomol. NMR 10 (1997) 199–203. DOI:`10.1023/a:1018304117895`.

[93] S. Yao, G. J. Howlett, R. S. Norton: Peptide self-association in aqueous trifluoroethanol monitored by pulsed field gradient NMR diffusion measurements, J. Biomol. NMR 16 (2000) 109–119. DOI:`10.1023/A:1008382624724`.

[94] E. J. Cabrita, S. Berger: DOSY studies of hydrogen bond association: tetramethylsilane as a reference compound for diffusion studies, Magn. Reson. Chem. 39 (2001) S142–S148. DOI:`10.1002/mrc.917`.

[95] C. A. Crutchfield, D. J. Harris: Molecular mass estimation by PFG NMR spectroscopy, J. Magn. Reson. 185 (2007) 179–182. DOI:`10.1016/j.jmr.2006.12.004`.

[96] E. Durand, M. Clemancey, J.-M. Lancelin, J. Verstraete, D. Espinat, A.-A. Quoineaud: Aggregation states of asphaltenes: Evidence of two chemical behaviors by ^{1}H diffusion-ordered spectroscopy nuclear magnetic resonance, J. Phys. Chem. C 113 (2009) 16266–16276. DOI:`10.1021/jp901954b`.

[97] J. Homer, M. C. Perry: Enhancement of the NMR spectra of insensitive nuclei using PENDANT with long-range coupling constants, J. Chem. Soc., Perkin Trans. 2 (1995) 533. DOI:`10.1039/p29950000533`.

[98] G. C. Levy, J. D. Cargioli: Spin-lattice relaxation in solutions containing Cr(III) paramagnetic relaxation agents, J. Magn. Reson. (1969-1992) 10 (1973) 231–234. DOI:`10.1016/0022-2364(73)90221-7`.

[99] Z. Zhou, Y. He, X. Qiu, D. Redwine, J. Potter, R. Cong, M. Miller: Optimum Cr(acac)$_3$ concentration for NMR quantitative analysis of polyolefins, Macromol. Symp. 330 (2013) 115–122. DOI:`10.1002/masy.201300034`.

[100] A. A. Smith: INFOS: spectrum fitting software for NMR analysis, J. Biomol. NMR 67 (2017) 77–94. DOI:`10.1007/s10858-016-0085-2`.

[101] S. Sokolenko, T. Jézéquel, G. Hajjar, J. Farjon, S. Akoka, P. Giraudeau: Robust 1D NMR lineshape fitting using real and imaginary data in the frequency domain, J. Magn. Reson. 298 (2019) 91–100. DOI:`10.1016/j.jmr.2018.11.004`.

[102] Y. Matviychuk, E. Steimers, E. von Harbou, D. J. Holland: Improving the accuracy of model-based quantitative nuclear magnetic resonance, Magn. Reson. 1 (2020) 141–153. DOI:`10.5194/mr-1-141-2020`.

[103] M. A. Connell, P. J. Bowyer, P. Adam Bone, A. L. Davis, A. G. Swanson, M. Nilsson, G. A. Morris: Improving the accuracy of pulsed field gradient NMR diffusion experiments: correction for gradient non-uniformity, J. Magn. Reson. 198 (2009) 121–131. DOI:`10.1016/j.jmr.2009.01.025`.

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

[105] G. A. Morris: Diffusion-ordered spectroscopy, John Wiley & Sons, Ltd, 2009. DOI:`10.1002/9780470034590.emrstm0119.pub2`.

[106] C. Whelan, G. Harrell, J. Wang: Understanding the k-medians problem, in: Proceedings of the International Conference on Scientific Computing, 2015, pp. 219–222.

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

[108] R. Neufeld, D. Stalke: Accurate molecular weight determination of small molecules via DOSY-NMR by using external calibration curves with normalized diffusion coefficients, Chem. Sci. 6 (2015) 3354–3364. DOI:`10.1039/c5sc00670h`.

[109] O. Großmann, D. Bellaire, N. Hayer, F. Jirasek, H. Hasse: Database for liquid phase diffusion coefficients at infinite dilution at 298 K and matrix completion methods for their prediction, Digital Discovery 1 (2022) 886–897. DOI:`10.1039/d2dd00073c`.

[110] C. R. Wilke, P. Chang: Correlation of diffusion coefficients in dilute solutions, AIChE J. 1 (1955) 264–270. DOI:`10.1002/aic.690010222`.

[111] K. A. Reddy, L. K. Doraiswamy: Estimating liquid diffusivity, Ind. Eng. Chem. Fundam. 6 (1967) 77–79. DOI:`10.1021/i160021a012`.

[112] M. T. Tyn, W. F. Calus: Diffusion coefficients in dilute binary liquid mixtures, J. Chem. Eng. Data 20 (1975) 106–109. DOI:`10.1021/je60064a006`.

[113] A. Einstein: Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen, Ann. Phys. 322 (1905) 549–560. DOI:`10.1002/andp.19053220806`.

[114] VDI-Gesellschaft Verfahrenstechnik und Chemieingenieurwesen, Düsseldorf, Germany (Ed.): VDI-Wärmeatlas, 11 ed., Springer Berlin Heidelberg, 2013. DOI:`10.1007/978-3-642-19981-3`.

[115] K. C. Pratt, W. A. Wakeham, A. R. J. P. Ubbelohde: The mutual diffusion coefficient for binary mixtures of water and the isomers of propanol, Proc. R. Soc. Lond. A Math. Phys. Sci. 342 (1975) 401–419. DOI:`10.1098/rspa.1975.0031`.

[116] D. G. Leaist, K. MacEwan, A. Stefan, M. 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, J. Chem. Eng. Data 45 (2000) 815–818. DOI:`10.1021/je000079n`.

[117] A. J. Easteal, L. A. Woolf, R. Mills: Velocity cross-correlation coefficients for the system acetonitrile-water at 278 K and 298 K, Z. Phys. Chem. 155 (1987) 69–78. DOI:`10.1524/zpch.1987.155.Part_1_2.069`.

[118] A. Fredenslund, J. Gmehling, P. Rasmussen: Vapor-liquid equilibria using UNIFAC, a group-contribution method, Elsevier: Amsterdam, The Netherlands, 1977.

[119] A. Klamt: Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena, J. Phys. Chem. 99 (1995) 2224–2235. DOI:`10.1021/j100007a062`.

[120] A. Klamt, V. Jonas, T. Bürger, J. C. W. Lohrenz: Refinement and parametrization of COSMO-RS, J. Phys. Chem. A 102 (1998) 5074–5085. DOI:`10.1021/jp980017s`.

[121] D. S. Abrams, J. M. Prausnitz: Statistical thermodynamics of liquid mixtures: a new expression for the excess gibbs energy of partly or completely miscible systems, AIChE J. 21 (1975) 116–128. DOI:`10.1002/aic.690210115`.

[122] A. Klamt, G. Schüürmann: COSMO: A new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient, J. Chem. Soc., Perkin Trans. 2 (1993) 799–805. DOI:`10.1039/P29930000799`.

[123] H. Grensemann, J. Gmehling: Performance of a conductor-like screening model for real solvents model in comparison to classical group contribution methods, Ind. Eng. Chem. Res. 44 (2005) 1610–1624. DOI:`10.1021/ie049139z`.

[124] T. Mu, J. Rarey, J. Gmehling: Group contribution prediction of surface charge density profiles for COSMO-RS(Ol), AIChE J. 53 (2007) 3231–3240. DOI:`10.1002/aic.11338`.

[125] Dortmund data bank, www.ddbst.com, 2018.

[126] H. Renon, J. M. Prausnitz: Local compositions in thermodynamic excess functions for liquid mixtures, AIChE J. 14 (1968) 135–144. DOI:`10.1002/aic.690140124`.

[127] R. Wittig, J. Lohmann, J. Gmehling: Vapor-liquid equilibria by UNIFAC group contribution. 6. revision and extension, Ind. Eng. Chem. Res. 42 (2003) 183–188. DOI:`10.1021/ie020506l`.

[128] T. Specht, K. Münnemann, H. Hasse, F. Jirasek: Rational method for defining and quantifying pseudo-components based on NMR spectroscopy, Phys. Chem. Chem. Phys. 25 (2023) 10288–10300. DOI:`10.1039/d3cp00509g`.

[129] Online group assignment for UNIFAC, http://www.ddbst.com/unifacga.html, Last accessed: 23.03.2023.

[130] V. Kiva, E. Hilmen, S. Skogestad: Azeotropic phase equilibrium diagrams: a survey, Chem. Eng. Sci. 58 (2003) 1903–1953. DOI:`10.1016/S0009-2509(03)00018-6`.

[131] L. Rayleigh: LIX. On the distillation of binary mixtures, Lond. Edinb. Dublin philos. Mag. J. Sci. 4 (1902) 521–537. DOI:`10.1080/14786440209462876`.

[132] J. Stichlmair, H. Klein, S. Rehfeldt: Distillation, Wiley, 2021. DOI:`10.1002/9781119414674`.

[133] Dortmund data bank, www.ddbst.com, 2021.

[134] M. Nilsson, G. A. Morris: Pure shift proton DOSY: diffusion-ordered ^{1}H spectra without multiplet structure, Chem. Commun. (2007) 933. DOI:`10.1039/b617761a`.

[135] K. Zangger: Pure shift NMR, Progress in Nuclear Magnetic Resonance Spectroscopy 86-87 (2015) 1–20. DOI:`10.1016/j.pnmrs.2015.02.002`.

[136] Download of NMRShiftDB data (file: nmrshiftdb2.nmredata.sd, modified 25.01.2023), Last accessed: 09.02.2023. URL: `https://sourceforge.net/projects/nmrshiftdb2/files/data/`.

[137] Download of BMRB data, Last accessed: 09.02.2023. URL: `https://bmrb.io/ftp/pub/bmrb/metabolomics/entry_directories/`.

[138] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. Gushurst, D. L. Grier, B. A. Leland, J. Laufer: Description of several chemical structure file formats used by computer programs developed at molecular design limited, J. Chem. Inf. Comput. Sci. 32 (1992) 244–255. DOI:`10.1021/ci00007a012`.

[139] M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn, M. Erdelyi, C. Steinbeck, A. J. Williams, C. Butts, T. D. Claridge, B. Mikhova, W. Robien, H. Dashti, H. R. Eghbalnia, C. Farès, C. Adam, P. Kessler, F. Moriaud, M. Elyashberg, D. Argyropoulos, M. Pérez, P. Giraudeau, R. R. Gil, P. Trevorrow, D. Jeannerat: NMReDATA, a standard to report the NMR assignment and parameters of organic compounds, Magn. Reson. Chem. 56 (2018) 703–715. DOI:`10.1002/mrc.4737`.

[140] E. L. Ulrich, K. Baskaran, H. Dashti, Y. E. Ioannidis, M. Livny, P. R. Romero, D. Maziuk, J. R. Wedell, H. Yao, H. R. Eghbalnia, J. C. Hoch, J. L. Markley: NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments, J. Biomol. NMR 73 (2018) 5–9. DOI:`10.1007/s10858-018-0220-3`.

[141] NMRPyStar package., https://github.com/mattfenwick/NMRPyStar, Last accessed: 13.12.2022.

[142] H. Dashti, W. M. Westler, J. L. Markley, H. R. Eghbalnia: Unique identifiers for small molecules enable rigorous labeling of their atoms, Sci. Data 4 (2017). DOI:`10.1038/sdata.2017.73`.

[143] S. R. Heller, A. McNaught, I. Pletnev, S. Stein, D. Tchekhovskoi: InChI, the IUPAC international chemical identifier, J. Cheminf. 7 (2015). DOI:`10.1186/s13321-015-0068-4`.

[144] Theory of support vector classification in scikit-learn, https://scikit-learn.org/stable/modules/svm.html#svc, Last accessed: 23.01.2023.

[145] G. R. Fulmer, A. J. M. Miller, N. H. Sherden, H. E. Gottlieb, A. Nudelman, B. M. Stoltz, J. E. Bercaw, K. I. Goldberg: NMR chemical shifts of trace impurities: Common laboratory solvents, organics, and gases in deuterated solvents relevant to the organometallic chemist, Organometallics 29 (2010) 2176–2179. DOI:`10.1021/om100106e`.

[146] M. Yemloul, V. Castola, S. Leclerc, D. Canet: Self-diffusion coefficients obtained from proton-decoupled carbon-13 spectra for analyzing a mixture of terpenes, Magn. Reson. Chem. 47 (2009) 635–640. DOI:`10.1002/mrc.2442`.

[147] E. O. Stejskal, J. E. Tanner: Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient, J. Chem. Phys. 42 (1965) 288–292. DOI:`10.1063/1.1695690`.

[148] R. Evans: The interpretation of small molecule diffusion coefficients: Quantitative use of diffusion-ordered nmr spectroscopy, Prog. Nucl. Magn. Reson. Spec-

trosc. 117 (2020) 33–69. DOI:`https://doi.org/10.1016/j.pnmrs.2019.11.002`.

[149] D. Li, I. Keresztes, R. Hopson, P. G. Williard: Characterization of reactive intermediates by multinuclear diffusion-ordered NMR spectroscopy (DOSY), Acc. Chem. Res. 42 (2009) 270–280. DOI:`10.1021/ar800127e`.

[150] M. T. Tyn, W. F. Calus: Temperature and concentration dependence of mutual diffusion coefficients of some binary liquid systems, J. Chem. Eng. Data 20 (1975) 310–316. DOI:`10.1021/je60066a009`.

[151] H. Ueadaira, H. Uedaira: Diffusion coefficients of xylose and maltose in aqueous solution, Bull. Chem. Soc. Jpn. 42 (1969) 2140–2142. DOI:`10.1246/bcsj.42.2140`.

[152] C.-M. Hsieh, S. I. Sandler, S.-T. Lin: Improvements of COSMO-SAC for vapor-liquid and liquid-liquid equilibrium predictions, Fluid Phase Equilib. 297 (2010) 90–97. DOI:`10.1016/j.fluid.2010.06.011`.

[153] J. Rachford, H.H., J. Rice: Procedure for use of electronic digital computers in calculating flash vaporization hydrocarbon equilibrium, J. Pet. Technol. 4 (1952) 19–3. DOI:`10.2118/952327-G`.

[154] Y. Teh, G. Rangaiah: A study of equation-solving and gibbs free energy minimization methods for phase equilibrium calculations, Chem. Eng. Res. Des. 80 (2002) 745–759. DOI:`10.1205/026387602320776821`.

[155] M. Ohanomah, D. Thompson: Computation of multicomponent phase equilibria-part I. Vapour-liquid equilibria, Comput. Chem. Eng. 8 (1984) 147–156. DOI:`10.1016/0098-1354(84)87001-5`.

[156] D. V. Nichita, C. F. Leibovici: A rapid and robust method for solving the rachford–rice equation using convex transformations, Fluid Phase Equilib. 353 (2013) 38–49. DOI:`10.1016/j.fluid.2013.05.030`.

Appendix

A Supporting Information for Chapter 2

A.1 Experimental Methods

In Table A.1, information on the chemicals for the preparation of the mixtures I-III that were studied in Chapter 2 (cf. Table 3) is summarized.

Table A.1: Suppliers and purities of chemicals used in Chapter 2. Purities are indicated as specified by the suppliers.

Chemical	Formula	Supplier	Purity
Acetone	C_3H_6O	Fisher Scientific	≥99.98%
Ethyl acetate	$C_4H_8O_2$	Fisher Scientific	≥99.92%
4-Hydroxybenzoic acid	$C_7H_6O_3$	Sigma Aldrich	≥99.00%
Ibuprofen	$C_{13}H_{18}O_2$	Sigma Aldrich	≥98.00%
1-Propanol	C_3H_8O	Honeywell	≥99.50%
tert-Butylhydroquinone	$C_{10}H_{14}O_2$	Merck	≥97.00%
Tetramethylsilane (TMS)	$C_4H_{12}Si$	Carl Roth	≥99.90%

A.1.1 Preparation of Samples and Acquisition of NMR Spectra

In the following section, the experimental procedure for the experimental examination of mixtures I-III are described. Mixture samples were prepared gravimetrically in 20 ml glass vessels using a balance of Mettler Toledo. The mass of each sample was at about 5 to 11 g. A small amount of tetramethylsilane (TMS) was added to each sample and used as reference in ^{13}C and ^{1}H NMR spectroscopy. 1 ml of each sample was transferred to a 5 mm NMR vial. NMR spectra were recorded with a 400 MHz Avance NMR spectrometer from Bruker utilizing a double resonant probe head. All measurements were performed at 30°C, since most considered pure component spectra were recorded at this temperature [49]. ^{1}H NMR spectra were recorded using a flip angle of 10°, a relaxation delay of 80 s, 8 scans and a bandwidth of about 25 ppm. MNova was used for automatic baseline and phase correction. Inverse-gated ^{13}C NMR spectra were recorded with a flip angle of 90°, a relaxation delay of 300 s, 16 scans and a bandwidth of about 250 ppm. Additionally, on the ^{13}C NMR spectra exponential line broadening from MNova was applied (0.2 Hz). Table 3 gives an overview of the components that make up the mixtures I-III, their composition, and the structural groups that are present in the mixtures.

A.1.2 Quantitative Evaluation and Comparison to Ground Truth

For the quantitative evaluation for the studied mixtures I-III, cf. Table 3, all picked peaks in the ^{13}C NMR spectrum of each mixture were integrated. By summing up the areas of all peaks in a section, the area A_s for each section s was obtained. For sections in which no peak was picked, A_s was set to zero. If a single structural group g was assigned to section s by the group-assignment method, the area A_s was completely attributed to the respective group g. Multiple areas A_s can be attributed to a single group g, if this group was assigned to multiple sections of the ^{13}C NMR spectrum by the group-assignment method. By summing up all areas attributed to group g, a total area A_g for each structural group g was obtained. As each of the considered structural groups contains exactly one carbon nucleus, the mole fraction x_g^{pred} of each group g in the mixture as predicted based on the results of the group-assignment method was calculated by:

$$x_g^{\text{pred}} = \frac{A_g}{\sum_{g=1}^{G} A_g} \tag{A.1}$$

where $G = 13$ is the number of considered structural groups here.

The predicted group mole fractions x_g^{pred} were compared to the true group mole fractions x_g, which were calculated from the known mole fractions x_i of the components i of each mixture and the known stoichiometry of each component i:

$$x_g = \frac{\sum_{i=1}^{N_\mathrm{mix}} x_i \cdot \nu_g^i}{\sum_{i=1}^{N_\mathrm{mix}} \sum_{g=1}^{G} x_i \cdot \nu_g^i} \tag{A.2}$$

where $N_\mathrm{mix} = 3$ is the number of components in the studied (ternary) mixtures and ν_g^i denotes the stoichiometric coefficient of group g in component i, cf. Table A.2.

Table A.2: Stoichiometric coefficients ν_g^i of structural groups g in components i.

Component i	Structural group g	Stoichiometric coefficient ν_g^i
Acetone	CH_3	2
	CO^ket	1
Ethyl acetate	CH_3	2
	COOR	1
	$ROOCH_x$	1
4-Hydroxy benzoic acid	$CH_x^\mathrm{ar}{=}$	5
	$RO{-}CH_x^\mathrm{ar}{=}$	1
	COOH	1
Ibuprofen	CH_3	3
	CH_x	3
	$CH_x^\mathrm{Ar}{=}$	6
	COOH	1
1-Propanol	CH_3	1
	CH_x	1
	CH_xOH	1
tert-Butylhydroquinone	CH_3	3
	CH_x	1
	$CH_x^\mathrm{ar}{=}$	4
	$RO{-}CH_x^\mathrm{ar}{=}$	2

A.2 NMR Spectra of Mixtures

In the following, the recorded ^{13}C and ^{1}H NMR spectra of mixtures I-III (cf. Table 3) are shown. In each spectrum, the peak positions that were used for evaluation of the proposed methods are flagged. Small peaks that could not be assigned to the main components of the mixtures (presumably from contaminations of the utilized chemicals) were not considered. If it was considered helpful, enlarged depictions of relevant regions of a spectrum are additionally shown, in particular for improved visibility of very broad peaks.

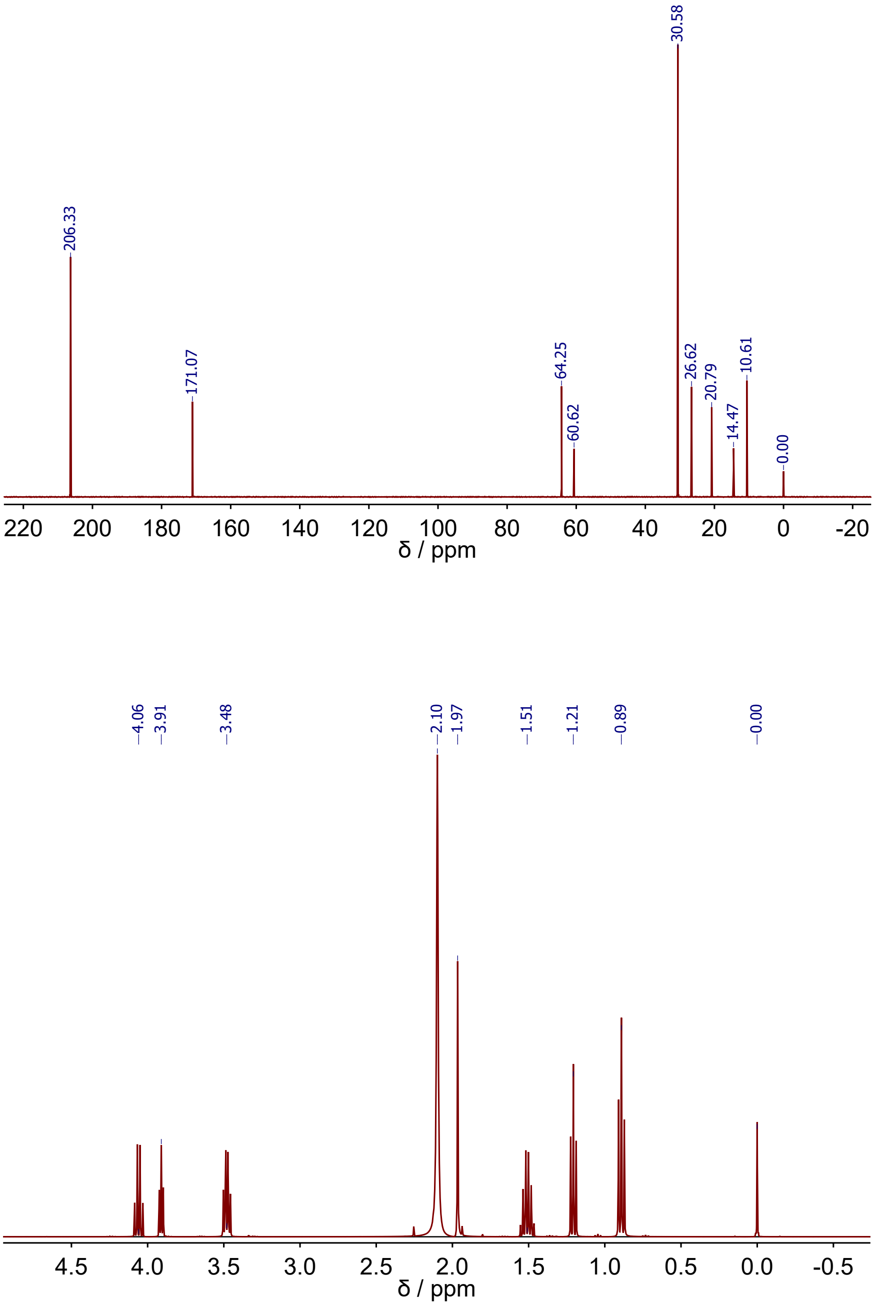


Figure A.1: ^{13}C and ^{1}H NMR spectra of mixture I (1-propanol, acetone, ethyl acetate), cf. Table 3. The black line is the baseline of the NMR spectrum. The peak at 0 ppm belongs to TMS, which was used as reference. Top: ^{13}C NMR spectrum. Bottom: ^{1}H NMR spectrum.

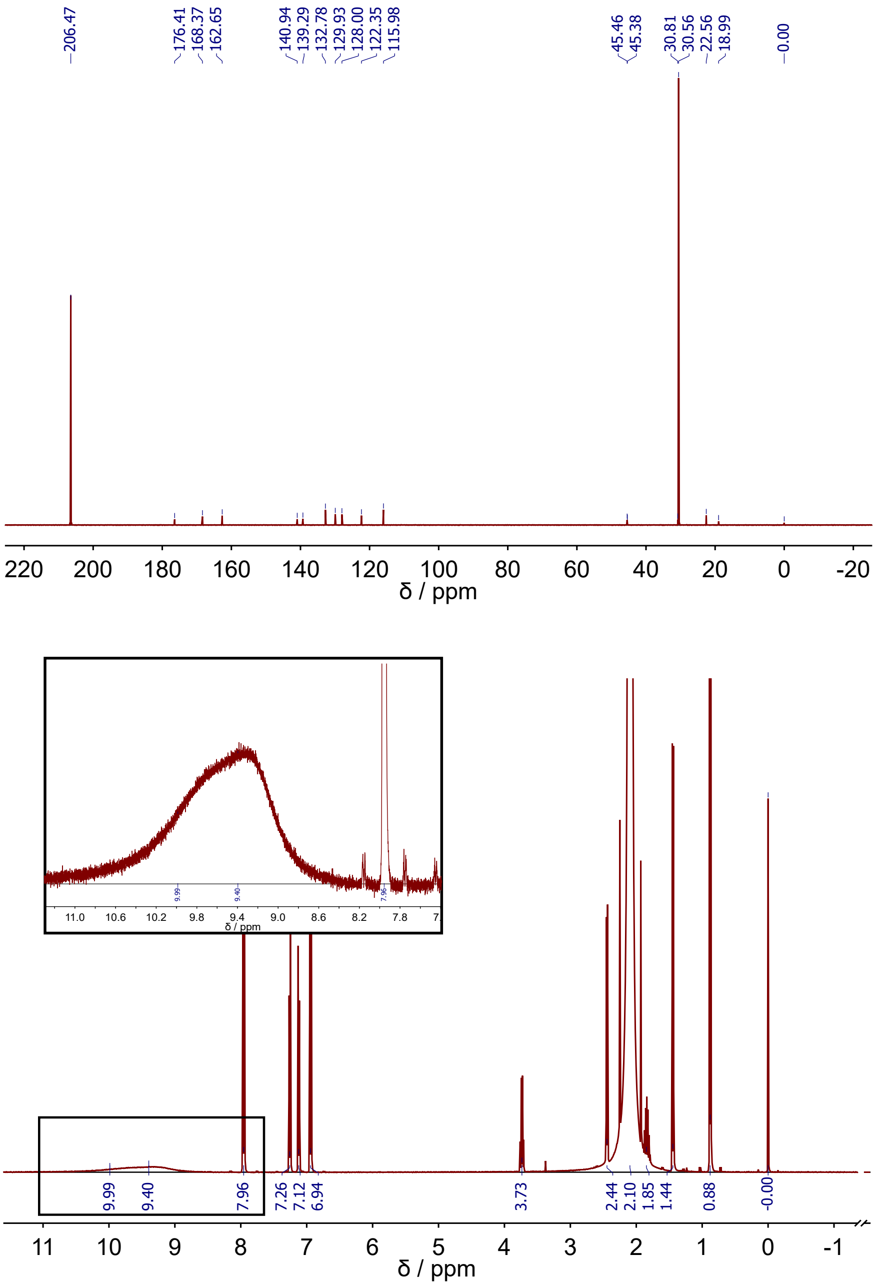


Figure A.2: ^{13}C and ^{1}H NMR spectra of mixture II (ibuprofen, acetone, 4-hydroxy benzoic acid), cf. Table 3. The black line is the baseline of the NMR spectrum. The peak at 0 ppm belongs to TMS, which was used as reference. Top: ^{13}C NMR spectrum. Bottom: ^{1}H NMR spectrum. Inset: enlarged depiction of the region in the ^{1}H NMR spectrum indicated by the black box.

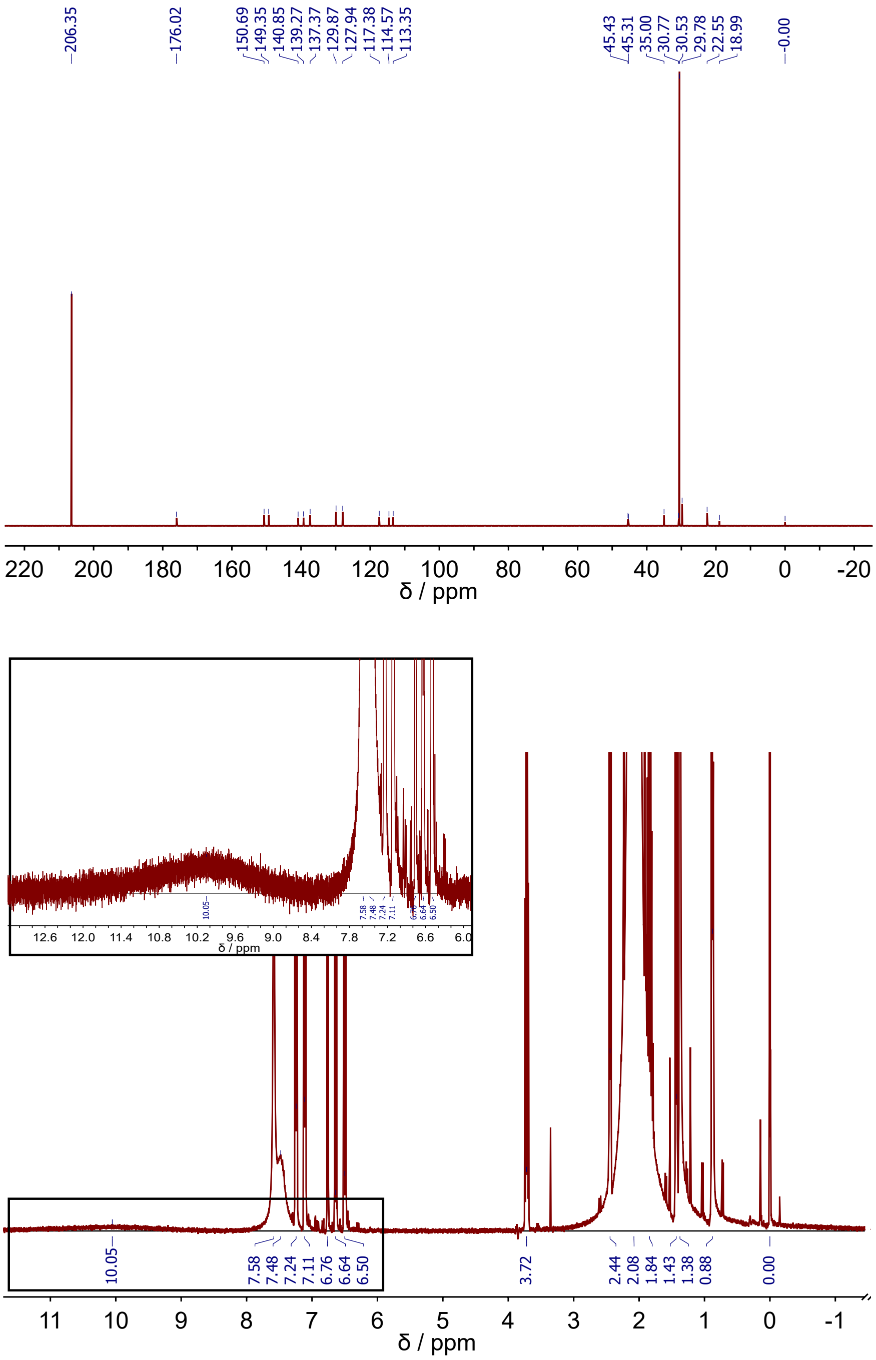


Figure A.3: ^{13}C and ^{1}H NMR spectra of mixture III (ibuprofen, acetone, tert-butylhydroquinone), cf. Table 3. The black line is the baseline of the NMR spectrum. The peak at 0 ppm belongs to TMS, which was used as reference. Top: ^{13}C NMR spectrum. Bottom: ^{1}H NMR spectrum. Inset: enlarged depiction of the region in the ^{1}H NMR spectrum indicated by the black box.

A.3 Influence of NMR Spectra Binning on Validation Score

In the following, the average F_1^{macro} validation score of the proposed group-identification method is compared for different discretizations of the NMR spectra, i.e., for different numbers of sections $S^{^{13}\mathrm{C}}$ and $S^{^{1}\mathrm{H}}$ in which the ^{13}C and ^{1}H NMR spectra, respectively, are divided (cf. Section 2.2.3). Figure A.4 shows that the F_1^{macro} validation score is quite robust and shows similar scores for many combinations of $S^{^{13}\mathrm{C}}$ and $S^{^{1}\mathrm{H}}$ (please note the narrow scale of the F_1^{macro} scores). Only for small numbers for $S^{^{13}\mathrm{C}}$ (<18), slightly lower scores are obtained. The highest average F_1^{macro} score on the validation data is achieved with the combination $S^{^{13}\mathrm{C}} = 23$ and $S^{^{1}\mathrm{H}} = 14$, which, therefore, were used to train and evaluate the methods shown in Chapter 2.

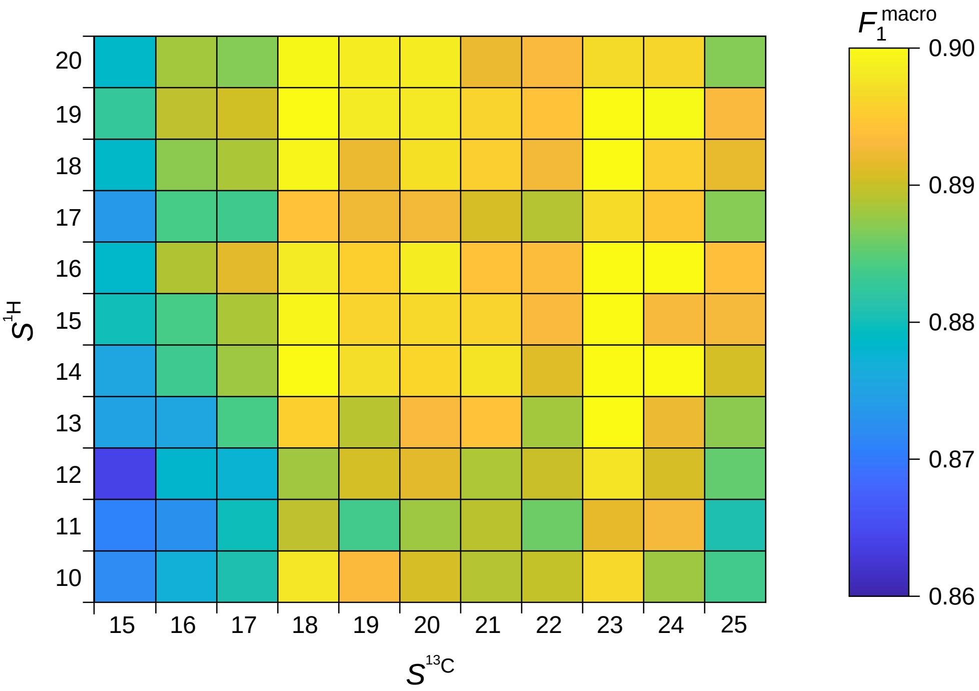


Figure A.4: Dependence of average F_1^{macro} validation score of the proposed group-identification method on the number of sections $S^{^{13}\mathrm{C}}$ and $S^{^{1}\mathrm{H}}$ in which the ^{13}C NMR and ^{1}H NMR spectra, respectively, are divided, cf. Section 2.2.3.

A.4 Handling of Special Cases

A.4.1 Classification Scores

The experimental data to train the proposed methods are scarce, i.e., experimental NMR spectra are only available for a limited number of pure components and, as a consequence, the number of positive examples for individual structural groups (classes) is greatly limited. The data set is furthermore extremely heterogeneous, as shown in Figure 1. The situation is especially challenging for training the group-assignment method, for which a separate Support Vector Classification (SVC) unit is trained to predict whether a peak in a *specific* region of the ^{13}C NMR spectrum indicates the presence of a *specific* structural group in the studied sample. In this situation, an appropriate handling of edge cases is indispensable.

Consider the case that an SVC unit was trained for identification of cyCH$_x$ groups in a specific section S* of the ^{13}C NMR spectrum. It is furthermore considered that at least some positive examples for cyCH$_x$ groups in section S* in the training set such that the considered SVC unit can be reasonably trained, but no positive example is available in the test (validation) set. In this case, the SVC unit cannot correctly predict the presence of a cyCH$_x$ group, i.e., $TP_{\mathrm{cyCH}_x} = 0$. Furthermore, the SVC unit cannot incorrectly predict the absence of a cyCH$_x$ group, i.e., $FN_{\mathrm{cyCH}_x} = 0$. In this case, the recall R is undefined (0/0), cf. Eq. (6). Per default in scikit-learn [55], undefined scores are set to 0. This is, however, not a good choice here as demonstrated in the following. In the considered scenario, the SVC unit has two choices if it is applied to a test (validation) data point: first, it can incorrectly predict the presence of a cyCH$_x$ group, i.e., $FP_{\mathrm{cyCH}_x} = 1$. In this case, the precision $P_{\mathrm{cyCH}_x} = 0$, cf. Eq. (5), and the recall R_{cyCH_x} is undefined, cf. Eq. (6). In combination, this yields $F_{1,\mathrm{cyCH}_x} = 0$, which is fair since the method was wrong. Second, if the SVC unit chooses to correctly predict the absence of cyCH$_x$ groups, i.e., $TN_{\mathrm{cyCH}_x} = 1$, both the precision P_{cyCH_x} and the recall R_{cyCH_x} are undefined, again yielding $F_{1,\mathrm{cyCH}_x} = 0$ if the default settings of scikit-learn [55] are adopted, even though the SVC unit is correct in this case. Therefore the F_1-score was set to 1 if both P and R are undefined.

A.5 Step-by-step Example for the Group-assignment Method

In the following, a detailed description of the application of the group-assignment method to NMR spectra of a mixture is given. This is the most sophisticated of the described scenarios; a transfer to the application of the group-identification method to mixture spectra, or of both methods to pure component spectra is straightforward.

The analysis of structural groups of mixtures is harder than the respective task for a pure component. In mixtures, NMR peaks are frequently subject to influences on their chemical shift due to intermolecular interactions, especially in ^{1}H NMR spectroscopy; this may lead to the shift of a specific peak by a few ppm rendering the correct assignment of this peak extremely difficult. Additionally, a greater number of different structural groups is usually present in mixtures, which may also confuse human NMR experts. It is referred to mixture I from Chapter 2 (1-propanol, acetone and ethyl acetate, cf. Table 3) as example here to describe the application of the group-assignment method step-by-step in the following.

The analysis is based on the ^{13}C and ^{1}H NMR spectra of mixture I, which are shown in Figure A.1. In the first step, automatic phase correction and baseline correction are performed with MNova and a chemical shift of 0 ppm is assigned to the peak of the NMR standard TMS. In the second step, the peaks in the NMR spectra are identified. It is suggested to use a manual peak picking. If an automatic peak picking, e.g., as offered by the MNova software, is used, a subsequent manual revision is recommended to remove wrong peaks, e.g., due to small baseline distortions. Moreover, very broad peaks are often missed by automatic approaches. ^{13}C satellites in the ^{1}H NMR spectrum must not be considered as peaks and can in most cases be identified easily due to multiplicity and symmetry to the main peak. Satellites in ^{13}C NMR are mostly not visible (due to very small intensity) and are also symmetric to the main peak. For singlets, the maxima of the peak were picked. In the case of two singlets (e.g., cf. Figure A.2), the overlapping was taken into account for peak picking of the maxima. For multiplets, the mean of the maxima of the peaks belonging to it was adopted.

In the next step, the input vectors $\boldsymbol{x}^{^{13}\mathrm{C}}$ and $\boldsymbol{x}^{^{1}\mathrm{H}}$, are defined depending on the identified peaks in the ^{13}C and ^{1}H NMR spectrum, respectively, as explained in detail in Section 2.2.3. $\boldsymbol{x}^{^{1}\mathrm{H}}$ is appended to $\boldsymbol{x}^{^{13}\mathrm{C}}$ yielding the final input vector of mixture I, referred to as $\boldsymbol{x}_{\mathrm{mix,I}}$ here. The results for $\boldsymbol{x}^{^{13}\mathrm{C}}$ and $\boldsymbol{x}^{^{1}\mathrm{H}}$ are given in Tables A.3 and A.4, respectively.

Table A.3: Input vector $\boldsymbol{x}^{^{13}\mathrm{C}}$ for mixture I (cf. Table 3). Ones (zeros) represent the presence (absence) of peaks in the respective sections of the NMR spectrum, indicated by the section limits $\delta^{^{13}\mathrm{C}}$. N_s represents the number of peaks in the respective section. All numbers are obtained from the ^{13}C NMR spectrum of the mixture, cf. Figure A.1 with $S^{^{13}\mathrm{C}} = 23$.

$\delta^{^{13}\mathrm{C}}$/ ppm	$\boldsymbol{x}^{^{13}\mathrm{C}}$	N_s
<9	0	0
9-18	1	2
18-27	1	2
27-37	1	1
37-46	0	0
46-55	0	0
55-64	1	1
64-73	1	1
73-82	0	0
82-91	0	0
91-100	0	0
100-110	0	0
110-119	0	0
119-128	0	0
128-137	0	0
137-146	0	0
146-155	0	0
155-164	0	0
164-173	1	1
173-183	0	0
183-192	0	0
192-201	0	0
>201	1	1

Table A.4: Input vector $\boldsymbol{x}^{^1\mathrm{H}}$ for mixture I (cf. Table 3). Ones (zeros) represent the presence (absence) of peaks in the respective sections of the NMR spectrum, indicated by the section limits $\delta^{^1\mathrm{H}}$. All numbers are obtained from the ^{1}H NMR spectrum of the mixture, cf. Figure A.1 with $S^{^1\mathrm{H}} = 14$.

$\delta^{^1\mathrm{H}}$ / ppm	$\boldsymbol{x}^{^1\mathrm{H}}$
<0.7	0
0.7-1.4	1
1.4-2.1	1
2.1-2.9	0
2.9-3.6	1
3.6-4.3	1
4.3-5.0	0
5.0-5.7	0
5.7-6.4	0
6.4-7.1	0
7.1-7.9	0
7.9-8.6	0
8.6-9.3	0
>9.3	0

From the input vector $\boldsymbol{x}_{\mathrm{mix,I}}$, the values of the decision function $d_g^s(\boldsymbol{x}_{\mathrm{mix,I}})$ for all structural group g / ^{13}C NMR section s combinations are calculated with the (previously trained) classifier using scikit-learn. The results are shown in Figure A.5. Only the results for the ^{13}C NMR sections in which at least one peak is observed (cf. Table A.3 and Figure A.1) and only for those structural group/^{13}C NMR section combinations for which a SVC unit was trained, are shown. $d_g^s(\boldsymbol{x}_{\mathrm{mix,I}}) > 0$ indicates that the classifier detects the respective structural group g in the respective section s of the ^{13}C NMR spectrum (these combinations are labeled with an asterisk in Figure A.5). However, as explained in Chapter 2, the decision of the classifier was not simply adopted but post-processing rules were applied to obtain physically consistent results. In the section ranging from 55-64 ppm, a ROOCH$_x$ and a CH$_x$ group are detected by the classifier. However, in the ^{13}C NMR spectrum, only a single peak is observed in this section, i.e., $N_\mathrm{s} = 1$ for this section, cf. Table A.3 and Figure A.1. Since it is rather unrealistic that the peaks of two different groups appear perfectly congruent in ^{13}C NMR spectroscopy, the ROOCH$_x$ group is accepted as the respective decision function exhibits the largest value, whereas the CH$_x$ group is rejected.

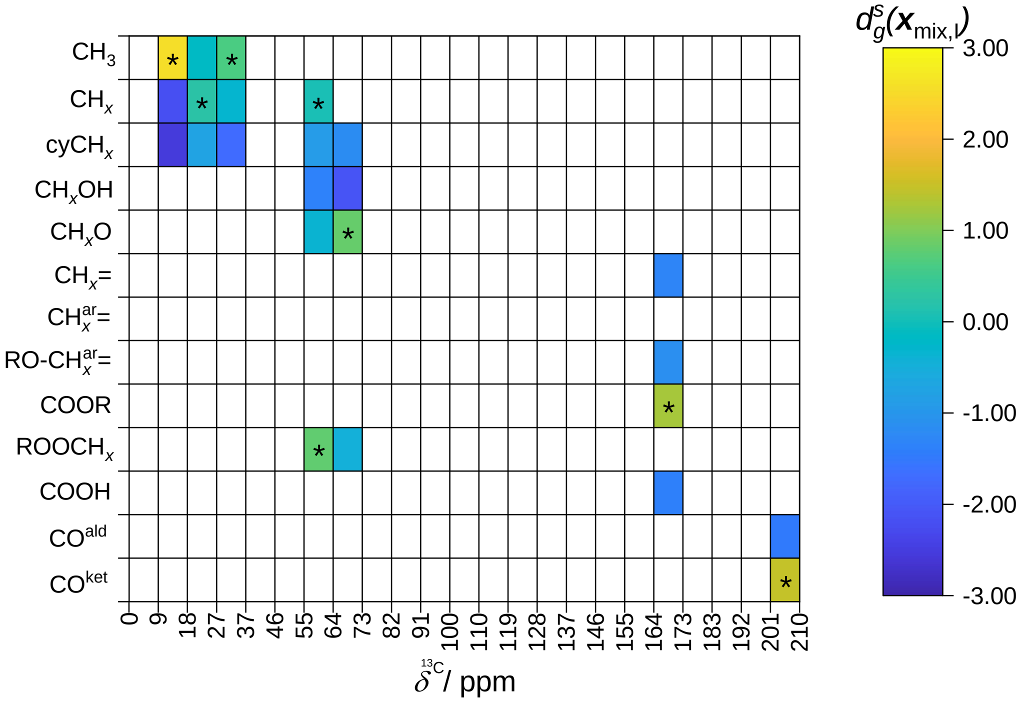


Figure A.5: Decision function values of the group-assignment method applied to mixture I. Asterisks indicate $d_g^s(\boldsymbol{x}_{\mathrm{mix,I}}) > 0$.

A.6 Linearly and Non-linearly Separable Data Sets

In the following, a brief overview of linear and non-linear classification problems is given by considering two different synthetic data sets in two dimensions; multi-dimensional problems can be considered analogously. Figure A.6 (left) shows an example of a data set with two classes that can simply be separated by a linear function. In contrast, Figure A.6 (right) shows a data set that obviously can not be separated reasonably with a linear function; as explained in detail in Ref. [47], after transformation of the data with a suitable transformation function $\phi(\boldsymbol{x})$, also such data can be separated with a linear function (which corresponds to a non-linear function in the original space, i.e. prior to the transformation).

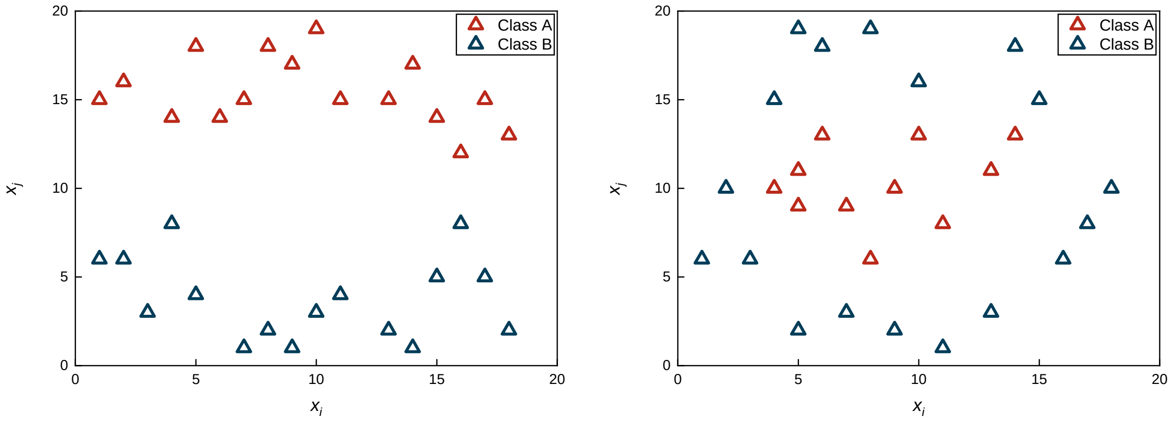


Figure A.6: Synthetic data set with two classes. Left: dataset that is perfectly separable with a linear hyperplane. Right: dataset that is not linearly separable. Adapted from Ref. [47].

A.7 Performance of the Group-identification Method for Different Solvents

In this section, the overall F_1 test scores of the proposed group-identification method for the application to pure component spectra, cf. Figure 2, is compared to the respective F_1 test scores that are obtained if only those components are considered as test data points, of which the NMR spectra were not recorded in $CDCl_3$. The goal here is to examine whether the group-identification method, which was trained to predominantly pure component NMR spectra recorded in $CDCl_3$, cf. Tables A.5, shows a lower performance if the structural groups are predicted from NMR spectra recorded with a different solvent. Since the NMR spectra of only 92 components were measured in other solvents, only scores of those structural groups with at least 10 examples were compared to obtain a reasonable sample size. The results are depicted in Figure A.7 and show that most structural groups show similar performance irrespective of whether the NMR spectra were recorded with $CDCl_3$ or other solvents. Surprisingly, the F_1 score for the COOH group is even higher if only solvents other than $CDCl_3$ are considered, although one could expect some deterioration due to shifting proton peaks depending on the solvents. Overall, the group-identification method is found to be rather robust towards different solvents with which NMR spectra are recorded. However, as the sample size is rather small and solvent effects can, in general, have a strong influence on peaks in NMR spectroscopy, these results should be treated with caution. The additional integration of information on the solvent in future work might significantly improve the reliability of the methods for different solvents.

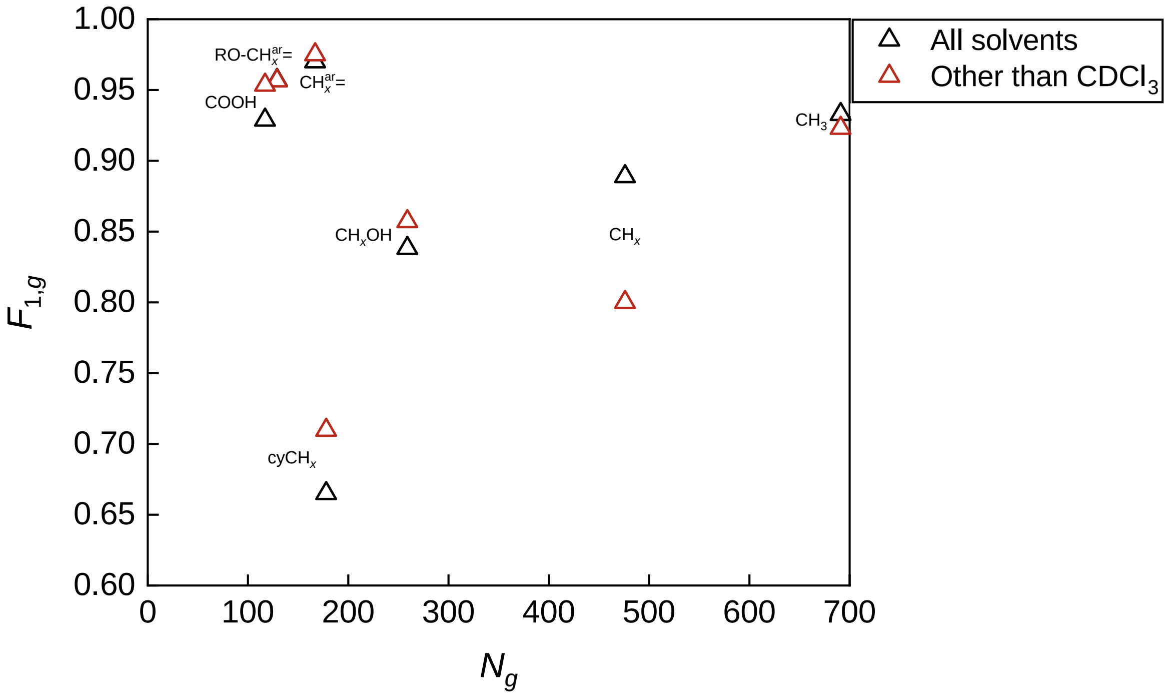


Figure A.7: F_1 test scores of the group-identification method for the considered structural groups g and the NMR spectra of the considered pure components. N_g denotes the number of pure components in the data set in which the respective structural group g is present.

A.8 Additional Results of the Group-assignment Method

In the following, additional results for the application of the group-assignment method using the post-processing rules, cf. Section A.5 and Section 2.2.5.3, to unknown samples are shown. In Figures A.8-A.10, results for pure components are shown. All components considered here were not included in the training set of the method. Furthermore, the method was only trained to components up to 160 g mol^{-1}, cf. Section 2.2, whereas all components considered here have a greater molar mass (up to 306 g mol^{-1}). In Figure A.11, results for three additional mixtures are shown. Surprisingly, also for the aqueous mixtures (middle and bottom panel) good predictions are found, although considerable shifting of peaks can be expected. This clearly demonstrates the robustness of the proposed approach.

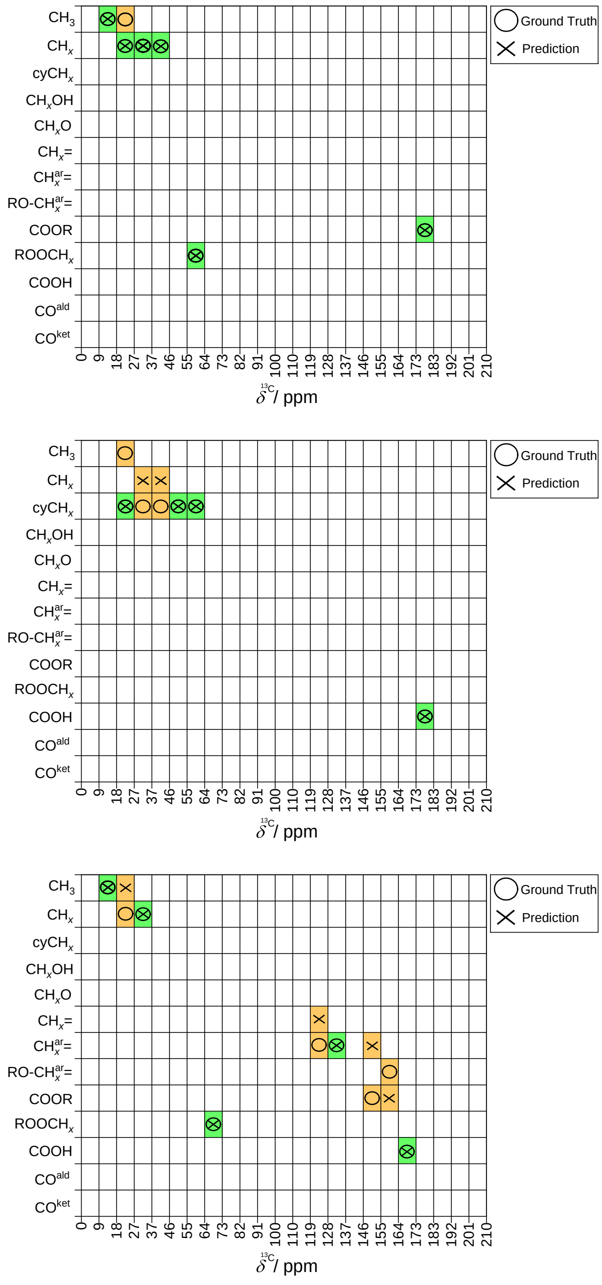


Figure A.8: Prediction of structural groups in pure components and assignment to sections in the ^{13}C NMR spectrum with the group-assignment method. Green areas indicate correct predictions, orange areas indicate mistakes. Top: isopentyl dodecanoate (1677), M = 270.4 g mol^{-1}. Middle: (+)-camphoric acid (6794), M = 200.2 g mol^{-1}. Bottom: p-(pentyloxycarbonyloxy)benzoic acid (7220), M = 252.3 g mol^{-1}. The number in parentheses refers to the SDBS No. of the component in the database.

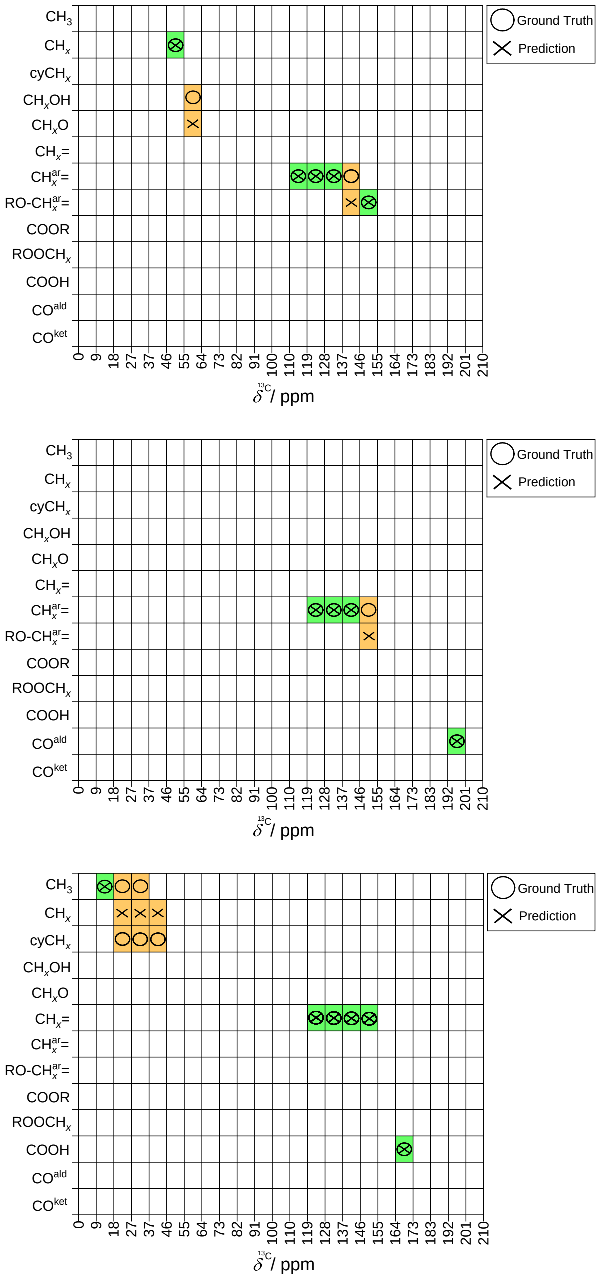


Figure A.9: Prediction of structural groups in pure components and assignment to sections in the ^{13}C NMR spectrum with the group-assignment method. Green areas indicate correct predictions, orange areas indicate mistakes. Top: o-(bis(p-hydroxyphenyl)methyl)benzyl alcohol (7550), M = 306.3 g mol^{-1}. Middle: 4-biphenylcarbaldehyde (10910), M = 182.2 g mol^{-1}. Bottom: all-trans-retinoic acid (21421), M = 300.4 g mol^{-1}. The number in parentheses refers to the SDBS No. of the component in the database.

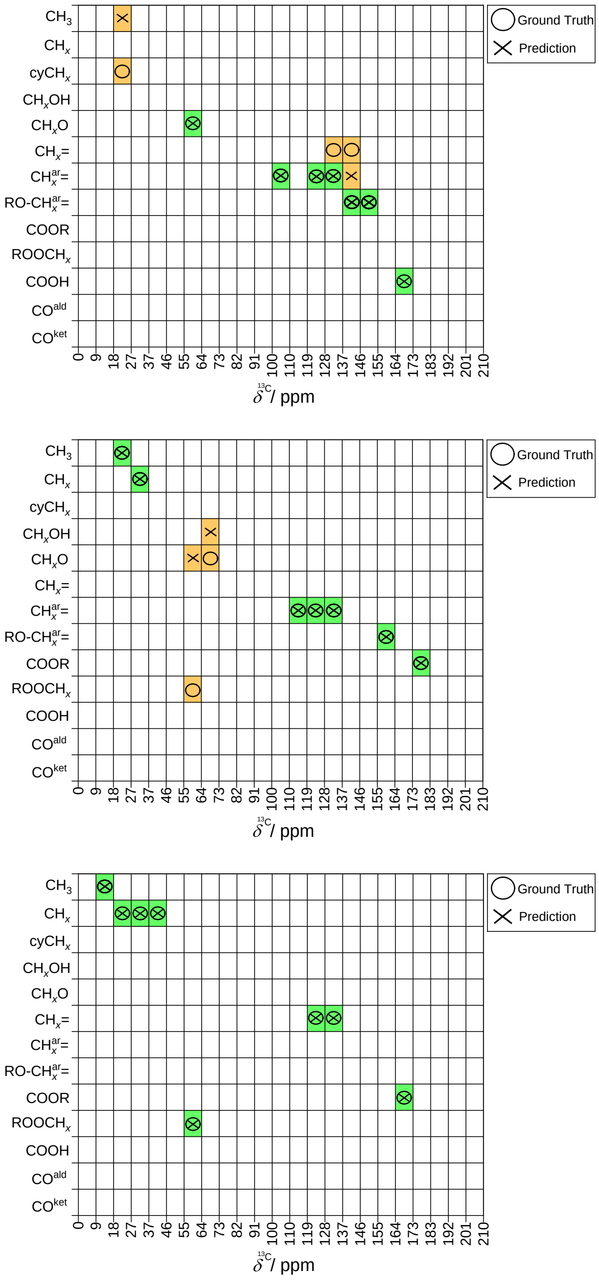


Figure A.10: Prediction of structural groups in pure components and assignment to sections in the ^{13}C NMR spectrum with the group-assignment method. Green areas indicate correct predictions, orange areas indicate mistakes. Top: 5,6,7-trimethoxy-3,4-dihydro-2-naphthoic acid (23750), M = 264.3 g mol^{-1}. Middle: 2-phenoxyethyl isobutyrate (52352), M = 208.2 g mol^{-1}. Bottom: ethyl trans-3-octenoate (52736), M = 170.2 g mol^{-1}. The number in parentheses refers to the SDBS No. of the component in the database.

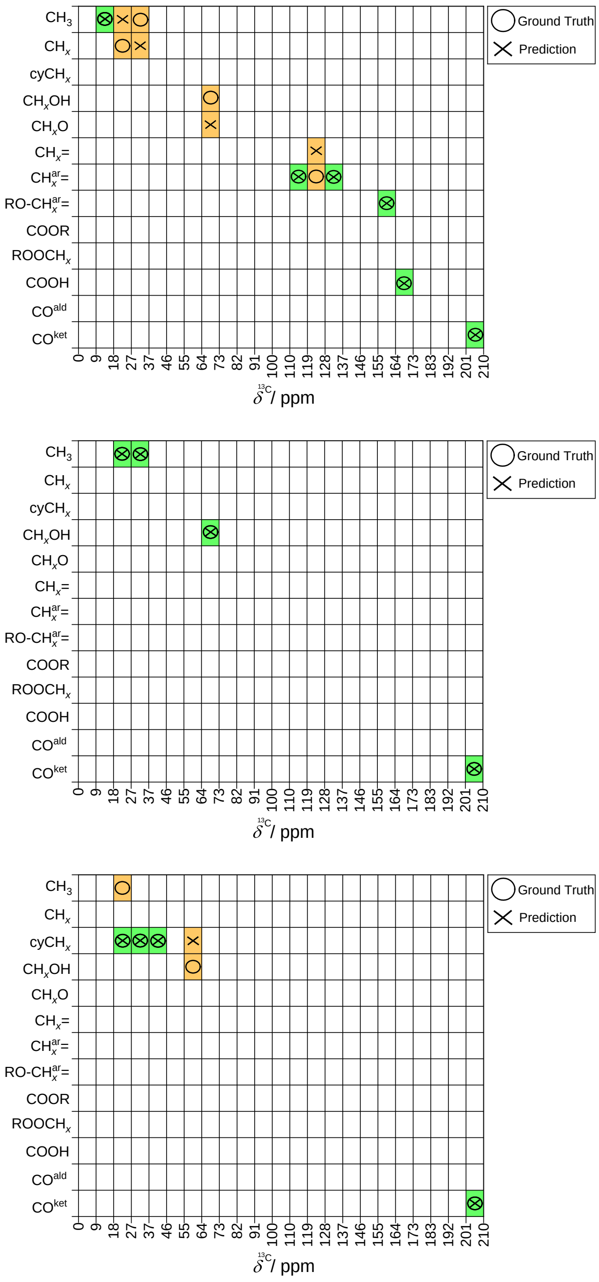


Figure A.11: Prediction of structural groups in mixtures and assignment to sections in the ^{13}C NMR spectrum with the group-assignment method. Green areas indicate correct predictions, orange areas indicate mistakes. Top: 4-hydroxybenzoic acid (0.026 mol mol^{-1}) + 1-propanol (0.616 mol mol^{-1}) + acetone. Middle: 2-propanol (0.025 mol mol^{-1}) + acetone (0.037 mol mol^{-1}) + water. Bottom: ethanol (0.042 mol mol^{-1}) + cyclohexanone (0.005 mol mol^{-1}) + water. The water peak was ignored for evaluation of the aqueous spectra.

A.9 Information on Studied Pure Components

A.9.1 Priority Rules for Structural Group Division

In the following, the priority rules that were applied during the manual division of the considered components into structural groups are described, which were necessary in a few cases. The carbon in CH_xO is labeled as such, even if it is bound to a hydroxyl group or carbon next to an ester. The carbon in the $RO{-}CH_x^{ar}{=}$ group is labeled as such, even if the carbon is bound to another group, e.g., a CH_xO or $ROOCH_x$ group.

A.9.2 Division of All Considered Pure Components

In Table A.5, all pure components that are considered in Chapter 2 are listed. Furthermore, the solvents for which the ^{13}C / ^{1}H NMR spectra of each component were recorded and the structural groups (according to the group division scheme, cf. Table 1) in each component are specified. It is noted that Table A.5 does not represent the stoichiometry of the components as only indicators of whether a structural group is part of the component or not are given; no quantitative information is included in Table A.5. If NMR spectra of a component were available for different solvents, the spectrum obtained with $CDCl_3$ was used, which was the case for most components. If no spectrum of the component in $CDCl_3$ was available, the first in the list of the SDBS database [49] was used. Additionally, if the component contains a carboxyl group (COOH), the presence of its characteristic peak in the ^{1}H NMR spectrum (usually >10 ppm) was checked; if it was not present in the spectrum, another spectrum obtained with another solvent in which the peak was present, or if this was not possible, the component was omitted. In a few cases, for some carbons of a component, no chemical shifts were reported, in that case, the component was not considered. The data were collected in January and February 2020.

Table A.5: List of the pure components considered in Chapter 2. Also information on the solvent with which the NMR spectrum was recorded according to the SDBS database [49] is provided as well as the structural groups that are (according to the group division scheme used here) part of the components. The SDBS No. is a unique identifier for each component.

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-furoic acid	1	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	1	0	0
trans-1,3-dimethylcyclohexane	63	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
5-hexen-2-one	70	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	1
trans-2-butenal	72	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	1	0
toluene	97	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	0	0
cyclohexyl acetate	102	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
2,4-dimethyl-1-hexene	203	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
trans-4,4-dimethyl-2-pentene	204	NEAT	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
(E)-3,4-dimethyl-2-pentene	205	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
(Z)-3,4-dimethyl-2-pentene	206	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-4-methyl-2-pentene	208	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
1,5-hexadiene	211	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	0	0	0	0	0
4-methyl-1-pentene	212	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
trans-1,4-hexadiene	213	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO{-}CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3,3-dimethyl-1-hexene	219	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
3,3-dimethyl-1-butene	222	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2-methyl-1-pentene	225	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
1,4-pentadiene	226	CCl_4	$CDCl_3$	0	1	0	0	0	1	0	0	0	0	0	0	0
trans-1,3-pentadiene	227	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
2-ethyl-1-hexene	274	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
1-hexene	275	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-4-octene	277	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
trans-4-octene	278	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-3-octene	279	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-2-octene	280	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
trans-2-octene	281	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-1,3-pentadiene	284	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
2-methyl-1,5-hexadiene	294	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
4-methyl-1,3-pentadiene	295	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
2,5-dimethyl-1,5-hexadiene	296	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
1,7-octadiene	297	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	0	0	0	0	0
2,3-dimethyl-1,3-butadiene	300	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
acetaldehyde	305	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	0	0	0	1	0
acetic acid	306	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	0	0	1	0	0
acetone	319	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	0	0	0	0	1
allyl alcohol	320	$CDCl_3$	$CDCl_3$	0	0	0	1	0	1	0	0	0	0	0	0	0
2,5-dimethylhexane	338	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,3-dimethylhexane	340	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
isovaleraldehyde	350	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
cyclopentene	351	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
2-cyclopenten-1-one	382	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	1
2-(2-ethoxyethoxy)ethanol	416	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
1,2-propanediol	472	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
maleic anhydride	475	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	1	0	0	0	0
succinic anhydride	478	DMSO-d6	$CDCl_3$	0	0	1	0	0	0	0	0	1	0	0	0	0
2,5-dihydrofuran	484	$CDCl_3$	$CDCl_3$	0	0	0	0	1	1	0	0	0	0	0	0	0
propylene carbonate	486	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
tetrahydrofuran	497	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
acetoin	498	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	1
2-methyl-2-propanol	506	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-butanol	507	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-methyl-1-propanol	508	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1,3-butanediol	509	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
ethylene glycol monoethyl ether	510	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
citraconic anhydride	516	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	0	0	0	0
methylenecyclobutane	522	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
3-methyl-2-butanone	528	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-pentanone	529	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-methyl-1-butanol	541	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3-methoxy-1-butanol	542	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
p-benzoquinone	549	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	0	0	0	0	1
phenol	554	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	0	0
methyl 2-furoate	556	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	1	1	0	0	0
cyclohexene	569	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
cyclohexanone	571	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
diethyl oxalate	575	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
methylcyclopentane	578	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
4-methyl-2-pentanone	580	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
cyclohexanol	581	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
hexanoic acid	582	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
1,1-cyclobutanedicarboxylic acid	612	DMSO-d6	DMSO-d6	0	0	1	0	0	0	0	0	0	0	1	0	0
2,2-dimethylbutane	652	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
3-methylpentane	653	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,3-dimethylbutane	654	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-methylpentane	655	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-ethyl-1-butanol	661	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
acetaldehyde diethyl acetal	663	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
benzaldehyde	672	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	0	1	0
benzoic acid	673	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	1	0	0
benzyl alcohol	685	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	0	0	0	0	0	0
m-cresol	686	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
trans-2-heptene	699	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
methylcyclohexane	700	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
4-heptanone	703	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3-heptanone	704	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
cis-4-methylcyclohexanol	706	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
trans-4-methylcyclohexanol	707	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
cis-2-methylcyclohexanol	708	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
trans-2-methylcyclohexanol	709	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
methyl hexanoate	710	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2,2-diethyl-1,3-propanediol	712	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
triethyl orthoformate	713	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
acetophenone	722	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	0	1
2'-hydroxyacetophenone	724	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	1
methyl benzoate	725	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	1	1	0	0	0
vanillin	726	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
methyl salicylate	727	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	1	1	0	0	0
p-anisyl alcohol	733	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	1	1	0	0	0	0	0
2,2,4,4-tetramethyl-1,3-cyclobutanedione	738	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
propylcyclopentane	784	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	0	0	0	0	0
1-octene	788	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-ethylhexanoic acid	789	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
ethyl hexanoate	790	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2-ethyl-1-hexanol	792	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-octanol	793	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,2,4-trimethyl-1,3-pentanediol	794	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-butanone	888	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
ethyl acetate	889	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
1,4-dioxane	890	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
2,4-dimethylpentane	892	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
cyclohexane	897	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	0
benzene	898	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	0	0	0
p-xylene	899	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	0	0
2,3-dimethylpentane	902	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,2-dimethylhexane	903	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
3-ethylhexane	904	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-hydroxyethyl acrylate	940	$CDCl_3$	$CDCl_3$	0	0	0	1	0	1	0	0	1	1	0	0	0
m-toluic acid	1000	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	$\mathrm{CH_3}$	CH_x	cyCH_x	$\mathrm{CH}_x\mathrm{OH}$	$\mathrm{CH}_x\mathrm{O}$	$\mathrm{CH}_x=$	$\mathrm{CH}_x^{\mathrm{ar}}=$	$\mathrm{RO{-}CH}_x^{\mathrm{ar}}=$	COOR	ROOCH_x	COOH	$\mathrm{CO^{ald}}$	$\mathrm{CO^{ket}}$
4'-hydroxyacetophenone	1008	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	0	0	0	0	1	1	0	0	0	0	1
terephthalaldehydic acid	1010	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	0	0	0	1	1	0
3'-hydroxyacetophenone	1016	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	0	0	0	0	1	1	0	0	0	0	1
2,3-dihydroxybenzoic acid	1017	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
o-xylene	1028	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	0	0	0	0	1	0	0	0	0	0	0
m-xylene	1032	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	0	0	0	0	1	0	0	0	0	0	0
propionic acid	1033	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	1	0	0	0	0	0	0	0	0	1	0	0
acetic anhydride	1034	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	0	0	0	0	0	0	1	0	0	0	0
maleic acid	1065	DMSO-d6	DMSO-d6	0	0	0	0	0	1	0	0	0	0	1	0	0
methyl oxalate	1068	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	0	0	0	0	0	0	1	1	0	0	0
L-(-)-malic acid	1069	DMSO-d6	$\mathrm{D_2O}$	0	1	0	1	0	0	0	0	0	0	1	0	0
glutaric anhydride	1091	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	0	0	0	1	0	0	0	0
2-oxoglutaric acid	1092	DMSO-d6	DMSO-d6	0	1	0	0	0	0	0	0	0	0	1	0	1
hydroquinone	1128	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	0	0	0
dimethyl fumarate	1132	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	0	0	0	1	0	0	1	1	0	0	0
2-ethyl-2-hydroxymethyl-1,3-propanediol	1144	$\mathrm{D_2O}$	$\mathrm{D_2O}$	1	1	0	1	0	0	0	0	0	0	0	0	0
anisole	1154	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	0	0	1	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
phthalide	1158	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	1	1	0	0	0
p-dimethoxybenzene	1160	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
methyl formate	1208	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	0	0	1	1	0	0	0
3-buten-2-ol	1211	$CDCl_3$	$CDCl_3$	1	0	0	1	0	1	0	0	0	0	0	0	0
1-propanol	1212	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
formaldehyde dimethyl acetal	1213	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
ethyl formate	1216	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
methyl acrylate	1220	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	1	1	0	0	0
methacrylic acid	1222	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	1	0	0
isobutyric acid	1224	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
methyl propionate	1226	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
butyric acid	1227	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
1,3-dioxane	1231	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
acetaldehyde dimethyl acetal	1232	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
cyclopentane	1235	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	0
2,2'-oxydiethanol	1236	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0
methyl isobutyrate	1237	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
acetone dimethyl acetal	1291	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
p-anisaldehyde	1293	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
ethyl alcohol	1300	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
hexyl alcohol	1303	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,5-dimethyl-2,4-hexadiene	1305	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
1,2-epoxycyclohexane	1308	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
2-furaldehyde	1309	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
furfuryl alcohol	1310	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	1	0	0	0	0	0
ethylene glycol monomethyl ether	1311	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0
furan	1313	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	0	0
ethyl methacrylate	1316	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
(1,2-epoxyethyl)benzene	1319	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	0	0	0	0	0	0
2,4,6-trimethyl-1,3,5-trioxane	1320	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
butyric anhydride	1323	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	0	0	0	0
gamma-butyrolactone	1325	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	1	0	0	0
1,2-dimethoxyethane	1328	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
ethylbenzene	1332	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	0	0	0	0	0	0
benzofuran	1353	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
m-hydroxyphenylacetic acid	1363	DMSO-d6	D_2O	0	1	0	0	0	0	1	1	0	0	1	0	0
o-cresol	1368	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
acrylaldehyde dimethyl acetal	1371	$CDCl_3$	$CDCl_3$	0	0	0	0	1	1	0	0	0	0	0	0	0
1-butanol	1374	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,3-xylenol	1375	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
p-hydroxybenzoic acid	1376	DMSO-d6	D_2O	0	0	0	0	0	0	1	1	0	0	1	0	0
m-hydroxybenzoic acid	1390	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
2,5-dihydroxybenzoic acid	1398	DMSO-d6	D_2O	0	0	0	0	0	0	1	1	0	0	1	0	0
4-oxovaleric acid	1418	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
ethyl acrylate	1424	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
methyl methacrylate	1425	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
2-ethylbutyraldehyde	1432	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
2,4-dimethylhexane	1433	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
phthalaldehyde	1434	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	0	1	0
2-methyl-1-cyclohexanone	1437	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
p-toluic acid	1448	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	0	0	0	1	0	0
adipic acid	1456	DMSO-d6	DMSO-d6	0	1	0	0	0	0	0	0	0	0	1	0	0
butyl ether	1458	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-hydroxy-2-methylpropionic acid	1483	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	1	0	0
methyl m-hydroxybenzoate	1492	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	1	1	0	0	0
resorcinol	1501	D_2O	D_2O	0	0	0	0	0	0	1	1	0	0	0	0	0
isophthalaldehyde	1508	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	0	1	0
5-methylresorcinol	1509	D_2O	D_2O	1	0	0	0	0	0	1	1	0	0	0	0	0
2-methyl-3-buten-2-ol	1539	$CDCl_3$	$CDCl_3$	1	0	0	1	0	1	0	0	0	0	0	0	0
dimethyl succinate	1545	$CDCl_3$	$CDCl_3$	0	1	0	0	0	0	0	0	1	1	0	0	0
anisic acid	1607	DMSO-d6	DMSO-d6	0	0	0	0	1	0	1	1	0	0	1	0	0
mandelic acid	1609	DMSO-d6	DMSO-d6	0	0	0	1	0	0	1	0	0	0	1	0	0
allyl acetate	1644	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
allyl butyrate	1648	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
acrylic acid	1649	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	0	0	1	0	0
alpha-acetyl-gamma-butyrolactone	1650	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	1
isopentyl formate	1676	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
m-methoxybenzaldehyde	1678	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
o-methoxybenzaldehyde	1680	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
pentyl acetate	1683	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2-pentanol	1708	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-methyl-2-butanol	1709	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
diethylene glycol dimethyl ether	1732	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
isobutyl propionate	1781	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
butyl propionate	1788	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
o-methoxytoluene	1798	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
m-methoxytoluene	1799	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
2,6-dihydroxytoluene	1810	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
3,5-xylenol	1850	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
2,6-xylenol	1851	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
o-ethylphenol	1857	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
trans-1,2-dimethylcyclohexane	1862	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
cis-1,2-dimethylcyclohexane	1863	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
1-methoxy-2-propanol	1884	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
3-methoxy-1,2-propanediol	1885	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
1,4-butanediol	1892	DMSO-d6	$CDCl_3$	0	1	0	1	0	0	0	0	0	0	0	0	0
2-ethyloxirane	1896	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
isobutyric anhydride	1900	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	0	0	0	0
butyl methacrylate	1907	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
butyl isobutyrate	1921	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isobutyl butyrate	1923	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
butyraldehyde	1925	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
isobutyraldehyde	1926	NEAT	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
isobutyl acetate	1928	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isobutyl acrylate	1930	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
butyl acrylate	1931	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
isopentyl acetate	1935	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
1-octanol	1938	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
guaiacol	1942	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
octanoic acid	2010	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
cyclooctane	2031	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	0
2,3-butanedione	2050	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	0	0	0	0	1
cyclooctene	2051	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
1,5-cyclooctadiene	2054	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
cis-1,2-cyclohexanedicarboxylic anhydride	2066	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	0	0	0	0
3-methylsalicylic acid	2069	$CDCl_3$	DMSO-d6	1	0	0	0	0	0	1	1	0	0	1	0	0
dimethyl malonate	2089	$CDCl_3$	$CDCl_3$	0	1	0	0	0	0	0	0	1	1	0	0	0
hexane	2118	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
m-dimethoxybenzene	2120	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
2-methyl-2-propen-1-ol	2124	$CDCl_3$	$CDCl_3$	1	0	0	1	0	1	0	0	0	0	0	0	0
2-propanol	2149	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
diethyl ether	2150	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
isovaleric acid	2157	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
ethylene diacetate	2162	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
2-methyl-1-butene	2176	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
dimethyl maleate	2180	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	1	1	0	0	0
1,2-ethanediol	2185	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	0	0	0	0	0	0	0
dimethyl methylenesuccinate	2190	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	1	1	0	0	0
3,4-xylenol	2194	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-methyl-2,4-pentanediol	2195	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3,6-dioxa-1-heptanol	2198	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0
dimethyl carbonate	2209	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	0	0	1	1	0	0	0
2,2-dimethyl-1,3-propanediol	2229	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1,6-hexanediol	2231	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	0	0	0	0	0	0	0
ethyl butyrate	2259	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
ethyl propionate	2285	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
3-hydroxy-4-methoxybenzaldehyde	2328	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
3-methoxysalicylaldehyde	2329	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
o-tolualdehyde	2338	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	1	0
2,2,4-trimethyl-1-pentanol	2342	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,2,4-trimethylpentane	2353	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-ethyl-2-hexenal	2355	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
(+-)-alpha-methylbenzyl alcohol	2365	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	1	0	0	0	0	0	0
(+-)-2-methylbutyric acid	2370	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
1,3-cyclooctadiene	2372	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
methyl pyruvate	2375	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	1
p-methoxyphenol	2379	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
ethylene carbonate	2392	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	0	0	1	1	0	0	0
ethylcyclohexane	2395	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	0	0	0	0	0
heptane	2396	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-methyltetrahydrofuran	2397	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
tetrahydropyran	2399	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
4-methoxy-2-butanone	2403	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
gamma-valerolactone	2408	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
trans-1,2-cyclohexanediol	2416	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
m-ethylphenol	2427	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
1,4-benzodioxane	2428	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
1,3-dihydroisobenzofuran	2437	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	0	0	0	0	0	0
p-ethylphenol	2443	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
isobutyl methacrylate	2452	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
crotonic anhydride	2463	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	0	0	0	0
2-ethoxy-2-methylpropane	2468	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
2-methoxy-2-methylpropane	2470	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
m-tolualdehyde	2473	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	1	0
pentane	2475	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,4,4-trimethyl-1-pentene	2476	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2-isopropoxyethanol	2477	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
1,2-diethoxyethane	2508	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
2-phenoxyethanol	2514	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	1	1	0	0	0	0	0
2-methoxyethyl acetate	2516	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	1	0	0	0
glycerol	2517	DMSO-d6	DMSO-d6	0	0	0	1	0	0	0	0	0	0	0	0	0
furfuryl acetate	2526	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	1	1	0	0	0
heptanal	2532	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
methyl p-hydroxybenzoate	2537	DMSO-d6	$CDCl_3$	0	0	0	0	0	0	1	1	1	1	0	0	0
p-methoxytoluene	2548	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
o-toluic acid	2628	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	1	0	0
methylsuccinic acid	2629	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
pyruvic acid	2642	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	0	0	1	0	1
1,5-pentanediol	2668	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	0	0	0	0	0	0	0
phenethyl alcohol	2670	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	1	0	0	0	0	0	0
octane	2672	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-pentanone	2673	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
DL-pantolactone	2705	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	1	1	0	0	0
hexyl acetate	2767	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
5-methyl-3-heptanone	2777	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
methyl acetate	2778	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
2-methylfuran	2803	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
5-methyl-2-hexanone	2810	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
m-methoxyphenol	2812	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
2-ethylbutyric acid	2823	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
methyl crotonate	2825	CCl_4	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
2,3-butanediol	2827	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
2-methylbutanal	2831	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
isopropyl formate	2834	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
diisopropyl ether	2837	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
phenetole	2838	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
1-methoxypropane	2841	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
methyl isovalerate	2885	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
methyl valerate	2890	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	RO$-CH_x^{ar}=$	COOR	ROOCH_x	COOH	CO^{ald}	CO^{ket}
propionaldehyde	2899	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
2-methyloxirane	2907	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
propyl butyrate	2908	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
3,6-dioxa-1,8-octanediol	2918	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0
propyl propionate	2919	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
propyl isobutyrate	2920	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isopropyl isobutyrate	2922	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isopropyl propionate	2923	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
propyl isovalerate	2940	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isopropyl butyrate	2941	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
veratrole	2942	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
1,3-propanediol	2951	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	0	0	0	0	0	0	0
pyrogallol	2979	D_2O	D_2O	0	0	0	0	0	0	1	1	0	0	0	0	0
3,5-dihydroxybenzoic acid	2988	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
succinic acid	3001	DMSO-d6	D_2O	0	1	0	0	0	0	0	0	0	0	1	0	0
pentyl formate	3030	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
trimethyl orthoacetate	3040	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
styrene	3044	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	1	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
dipropyl ether	3047	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
diethyl carbonate	3079	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
pivalic acid	3139	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
2,6-dihydroxybenzoic acid	3142	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
2-cyclohexen-1-ol	3231	$CDCl_3$	$CDCl_3$	0	0	1	1	0	1	0	0	0	0	0	0	0
2-ethoxyethyl acetate	3291	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	1	0	0	0
butyl acetate	3292	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
4-methyl-3-penten-2-one	3298	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
2-octanone	3299	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
methanol	3302	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	0	0	0	0	0	0	0
butyl lactate	3303	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
p-tolualdehyde	3376	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	0	0	0	0	1	0
valeric acid	3381	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
3-allyloxy-1,2-propanediol	3396	$CDCl_3$	$CDCl_3$	0	0	0	1	1	1	0	0	0	0	0	0	0
alpha-monopropionin	3399	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
trimethyl orthoformate	3401	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
4,4-dimethoxy-2-butanone	3407	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
ethyl valerate	3432	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
m-hydroxybenzaldehyde	3438	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
2-methylhydroquinone	3439	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	0	0	0
p-hydroxybenzaldehyde	3444	DMSO-d6	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
4-methylpyrocatechol	3450	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
methyl cyclohexanecarboxylate	3464	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	1	0	0	0
ethyl (E)-2-methyl-2-butenoate	3466	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
tetrahydrofurfuryl acetate	3467	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	1	1	0	0	0
methyl trans,trans-2,4-hexadienoate	3468	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
cyclooctanol	3470	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
ethyl 3-ethoxypropionate	3471	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	1	1	0	0	0
cyclopentaneethanol	3473	$CDCl_3$	$CDCl_3$	0	1	1	1	0	0	0	0	0	0	0	0	0
1-butoxy-2-propanol	3474	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
2-methyl-2-propyl-1,3-propanediol	3491	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
phenoxyacetic acid	3508	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	1	0	0
cyclohexanecarboxylic acid	3512	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
tert-butyl methacrylate	3514	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
1-phenyl-1,2-ethanediol	3526	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	0	0	0	0	0	0
acrylaldehyde diethyl acetal	3541	$CDCl_3$	$CDCl_3$	1	0	0	0	1	1	0	0	0	0	0	0	0
1,1-dimethoxycyclohexane	3542	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
ethyl 2-furoate	3567	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	1	1	0	0	0
alpha-methoxytoluene	3571	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	0	0	0	0	0	0
methyl heptanoate	3572	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
o-ethoxyphenol	3573	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
6-methyl-5-hepten-2-one	3575	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	1
terephthalaldehyde	3580	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	0	1	0
beta-butyrolactone	3667	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
cyclohexanemethanol	3728	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
3-methylcyclohexanone	3734	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
ethyl lactate	3794	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
3-methyl-2-pentanone	3795	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
1-heptene	3876	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
diisobutyl ether	3877	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
2,4-dimethyl-3-pentanol	3884	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
(butoxymethyl)oxirane	3885	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
cycloheptatriene	3886	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
ethyl 2-ethylbutyrate	3907	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
ethyl isovalerate	3909	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
cycloheptanone	3947	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
2,4-dimethyl-3-pentanone	3978	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-methoxybutyl acetate	3981	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	1	1	0	0	0
2,3-epoxypropyl methyl ether	4004	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
3-buten-2-one	4018	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
1,2-butanediol	4050	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(E)-2-methyl-2-butenoic acid	4079	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	1	0	0
erythritol	4081	D_2O	D_2O	0	0	0	1	0	0	0	0	0	0	0	0	0
allyl formate	4088	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	1	1	0	0	0
oxetane	4097	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
1,3-dioxolane	4103	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
methacrylaldehyde	4117	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	1	0
1,2:3,4-diepoxybutane	4142	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
trans-2-pentenoic acid	4233	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
octanal	4269	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
1-pentanol	4321	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
cis-1,3-dimethylcyclohexane	4322	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
3-methylheptane	4324	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
3-ethyl-3-methylpentane	4325	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-methylheptane	4326	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
acrylaldehyde	4448	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	0	0	0	1	0
2-methyl-1,3-dioxolane	4471	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
DL-2-hydroxy-4-methylvaleric acid	4555	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	1	0	0
1,8-octanediol	4559	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	0	0	0	0	0	0	0
vanillyl alcohol	4563	DMSO-d6	$CDCl_3$	0	0	0	1	1	0	1	1	0	0	0	0	0
methyl (E)-2-methyl-2-butenoate	4573	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
methyl 4-oxovalerate	4575	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
bipropionyl	4577	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-methyl-3-pentanol	4588	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
4-octanolide	4594	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	CH_x=	CH^{ar}_x=	RO−CH^{ar}_x=	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
4-methyl hydrogen methylenesuccinate	4600	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	1	1	1	0	0
2,2-dimethyl-1,3-dioxolane	4635	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
3-ethoxypropionic acid	4639	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	1	0	0
2-(allyloxy)ethanol	4640	$CDCl_3$	$CDCl_3$	0	0	0	1	1	1	0	0	0	0	0	0	0
cyclooctanone	4641	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
propyl acetate	4643	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2,2-dimethyl-1-propanol	4708	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3-hexanol	4711	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-isobutoxyethanol	4715	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
3,3-dimethyl-2-butanol	4716	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
methyl (+-)-2-methylbutyrate	4718	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
methyl pivalate	4720	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2-hexanol	4722	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,5-hexanedione	4729	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
2-ethyl-2-butenal	4759	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
3-hexanone	4760	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
cyclopentanecarboxylic acid	4780	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	1	0	0
2,4,4-trimethyl-2-pentene	4798	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
1-pentene	4812	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-2-pentene	4814	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2,3-dimethyl-1-butene	4815	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2,3-dimethyl-2-butene	4816	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
2-methyl-2-pentene	4817	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
trans,trans-3,5-heptadien-2-one	4876	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
cis-9-oxabicyclo(6.1.0)nonane	4884	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
1,3-benzodioxole	4895	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
L-(+)-alpha-hydroxyphenylacetic acid	4968	DMSO-d6	DMSO-d6	0	0	0	1	0	0	1	0	0	0	1	0	0
ethylcyclopentane	5134	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	0	0	0	0	0
3-ethylpentane	5141	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
3,4,5-toluenetriol	5157	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	0	0	0
cyclopropanecarboxylic acid	5317	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	1	0	0
methylmalonic acid	5318	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	RO$-CH_x^{ar}=$	COOR	ROOCH_x	COOH	CO^{ald}	CO^{ket}
3-buten-1-ol	5328	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
3-methyl-3-buten-2-one	5390	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
cyclobutanecarboxylic acid	5396	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	1	0	0
2,3-pentanedione	5398	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
4-pentenoic acid	5399	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	0	0	1	0	0
2-oxopentanoic acid	5401	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
trans-2-pentene	5405	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
pivalaldehyde	5412	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
methyl dimethoxyacetate	5414	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	1	1	0	0	0
(R)-(-)-2-pentanol	5429	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
5-(hydroxymethyl)furfural	5478	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	1	0	0	0	1	0
1,3-cyclohexadiene	5491	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
2-cyclohexen-1-one	5498	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	1
(3S-cis)-(-)-dilactide	5500	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
(+-)-dilactide	5501	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
1-methyl-1-cyclopentene	5507	$CDCl_3$	$CDCl_3$	1	0	1	0	0	1	0	0	0	0	0	0	0
1-methylallyl acetate	5514	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
gamma-caprolactone	5515	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3-(allyloxy)propionic acid	5516	$CDCl_3$	$CDCl_3$	0	1	0	0	1	1	0	0	0	0	1	0	0
4-acetylbutyric acid	5517	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
(+-)-mevalonic acid delta-lactone	5518	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	1	1	0	0	0
meso-2,3-dimethylsuccinic acid	5519	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
ethylidene diacetate	5520	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
trans-2-hexene	5531	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cis-3-hexene	5532	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
4-ethoxy-2-butanone	5540	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
3-methyl-1-pentanol	5603	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
methyl-p-benzoquinone	5658	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
2,5-dihydroxybenzaldehyde	5660	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
3,4-dihydroxybenzaldehyde	5661	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	0	1	0
3,4-dihydroxybenzoic acid	5663	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
p-hydroxybenzyl alcohol	5702	DMSO-d6	DMSO-d6	0	0	0	1	0	0	1	1	0	0	0	0	0
3-methylpyrocatechol	5703	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
3-methoxypyrocatechol	5707	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	$\mathrm{CH_3}$	CH_x	cyCH_x	$\mathrm{CH}_x\mathrm{OH}$	$\mathrm{CH}_x\mathrm{O}$	$\mathrm{CH}_x=$	$\mathrm{CH}_x^{\mathrm{ar}}=$	$\mathrm{RO-CH}_x^{\mathrm{ar}}=$	COOR	ROOCH_x	COOH	$\mathrm{CO^{ald}}$	$\mathrm{CO^{ket}}$
methoxyhydroquinone	5708	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	0	0	1	0	1	1	0	0	0	0	0
norbornylene	5719	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	1	0	0	0	0	0	0	0
methyl 2-oxocyclopentanecarboxylate	5732	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	0	0	0	1	1	0	0	1
3,3-dimethylglutaric anhydride	5734	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	1	0	0	0	0	0	1	0	0	0	0
4-ethyl hydrogen itaconate	5735	DMSO-d6	DMSO-d6	1	1	0	0	0	1	0	0	1	1	1	0	0
cycloheptene	5744	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	1	0	0	0	0	0	0	0
1-ethyl-1-cyclopentene	5745	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	1	1	0	0	1	0	0	0	0	0	0	0
1-methyl-1-cyclohexene	5749	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	1	0	0	1	0	0	0	0	0	0	0
cyclohexanecarbaldehyde	5755	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	0	0	0	0	0	0	1	0
4-methylcyclohexanone	5761	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	0	1	0	0	0	0	0	0	0	0	0	1
cyclopentylacetic acid	5762	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	1	1	0	0	0	0	0	0	0	1	0	0
2,3-heptanedione	5765	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-heptenoic acid	5767	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	1	0	0	0	1	0	0	0	0	1	0	0
cycloheptane	5790	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	0	0	1	0	0	0	0	0	0	0	0	0	0
phloroglucinol	5851	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	0	0	0
(S)-(+)-2-octanol	5855	$\mathrm{CDCl_3}$	$\mathrm{CDCl_3}$	1	1	0	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2,4-dimethyl-1-pentene	6018	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2,4-dimethyl-2-pentene	6019	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
4,4-dimethyl-1-pentene	6020	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
3-ethyl-2-pentene	6021	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2-methyl-1-hexene	6024	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2-methyl-2-hexene	6025	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
3-methyl-1-hexene	6026	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
4-methyl-1-hexene	6029	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
5-methyl-1-hexene	6031	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2,3,3-trimethyl-1-butene	6033	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
cycloheptanol	6038	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
1,2-epoxyheptane	6039	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
5-methyl-3-hexanone	6040	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
1-methylcyclohexanol	6041	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
2-methyl-3-hexanone	6042	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
3-methyl-2-hexanone	6043	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
ethyl (+-)-2-methylbutyrate	6046	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2-methylhexane	6055	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3-methylhexane	6056	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,2,3-trimethylbutane	6057	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-heptanol	6060	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3-heptanol	6061	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
4-heptanol	6062	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-methyl-2-hexanol	6063	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,2-diethoxypropane	6064	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
1,3-diethoxy-2-propanol	6065	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
cyclooctatetraene	6117	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	0	0	0	0	0
2,5-dimethyl-p-benzoquinone	6136	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
o-anisic acid	6137	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	1	0	0
p-methylbenzyl alcohol	6196	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	1	0	0	0	0	0	0
p-ethoxyphenol	6198	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
2,6-dimethoxyphenol	6199	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
3,5-dimethoxyphenol	6200	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
ethyl 2-oxo-1-cyclopentanecarboxylate	6217	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
ethyl tetrahydrofuran-2-acetate	6229	$CDCl_3$	$CDCl_3$	1	1	1	0	1	0	0	0	1	1	0	0	0
trans-1,4-dimethylcyclohexane	6236	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
4,4-dimethyl-1-hexene	6238	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2-methyl-1-heptene	6239	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
2,5-dimethyl-3-hexanone	6245	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
2-methyl-3-heptanone	6246	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
2-methyl-4-heptanone	6247	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
6-methyl-2-heptanone	6248	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
4-isobutoxy-2-butanone	6249	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
2-propylvaleric acid	6250	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
3-ethyl-2-methylpentane	6256	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
4-methylheptane	6257	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2,2,3,3-tetramethylbutane	6258	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
6-methyl-3-heptanol	6261	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(R)-(-)-2-octanol	6262	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(Z)-2-methyl-2-butenoic acid	7048	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
trans-2-heptenoic acid	7623	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0
cyclopentanone	7939	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
isoprene	7940	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
methyl acetoacetate	7942	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
ethyl propyl ether	7950	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
1,2-epoxy-3-isopropoxypropane	8240	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
cis-1,4-cyclohexanediol	8456	DMSO-d6	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
trans-1,4-cyclohexanediol	8457	DMSO-d6	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
cis-1,4-dimethylcyclohexane	8461	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
methyl trans-3-acetylacrylate	8483	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	1
4,4-dimethyl-2-cyclohexen-1-one	8507	$CDCl_3$	$CDCl_3$	1	0	1	0	0	1	0	0	0	0	0	0	1
dicyclopropyl ketone	8552	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
cis-2,4-dimethyl-1,3-dioxane	8878	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
trans-3-heptene	8885	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
acetylcyclopropane	8901	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
2-hydroxy-p-toluic acid	8995	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2,3-epoxy-1-propanol	9283	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
3-methylfuran	9353	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
5,6-dihydro-2-pyrone	9850	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	1	1	0	0	0
propionaldehyde dimethyl acetal	10046	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
1-methoxy-2-butanol	10084	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
1,1,2-trimethoxyethane	10086	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
pinacol	10089	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
2-ethyl-4-methyl-1-pentanol	10101	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-(hexyloxy)ethanol	10102	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
1,2-epoxyoctane	10103	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
cis-3-hexenyl acetate	10143	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
butyl crotonate	10144	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
formaldehyde diethyl acetal	10145	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
(+-)-2-methyl-1-butanol	10147	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
tert-butyl acetoacetate	10148	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
trans-2-hexenyl acetate	10151	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
sec-butyl acetoacetate	10153	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH^x	$cyCH^x$	CH^xOH	CH^xO	$CH^x=$	$CH^x_{ar}=$	$RO-CH^x_{ar}=$	COOR	$ROOCH^x$	COOH	CO_{ald}	CO_{ket}
ethyl trans,trans-2,4-hexadienoate	10173	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
4-ethylresorcinol	10199	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
4-methyl-1,3-dioxane	10201	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
trans-3-hexenoic acid	10217	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0
3-methylvaleric acid	10223	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
2-methylpentanal	10224	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
3,3-dimethylbutyric acid	10226	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
trans-2-hexenal	10228	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
isopropyl acetate	10237	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
3-methyl-2-butanol	10240	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,5-dimethyl-2,5-hexanediol	10259	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-acetylfuran	10263	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	1
trans-2-hexenoic acid	10264	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0
4-hydroxy-4-methyl-2-pentanone	10287	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1
ethyl DL-3-hydroxybutyrate	10288	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
oxepane	10291	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
D-(-)-panto-1,4-lactone	10296	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	1	1	0	0	0
trans-2-hexen-1-ol	10321	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
cis-3-hexen-1-ol	10323	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
ethyl 2-hydroxy-2-methylpropionate	10324	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
tetrahydropyran-2-methanol	10334	$CDCl_3$	$CDCl_3$	0	0	1	1	1	0	0	0	0	0	0	0	0
diallyl ether	10339	$CDCl_3$	$CDCl_3$	0	0	0	0	1	1	0	0	0	0	0	0	0
2-ethyl-1-butene	10346	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
3-methoxy-3-methyl-1-butanol	10347	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
3-octanone	10370	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
2-cyclohexylethanol	10373	$CDCl_3$	$CDCl_3$	0	1	1	1	0	0	0	0	0	0	0	0	0
methyl 2-hydroxy-2-methylpropionate	10374	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
3-octanol	10378	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1-octene-3-ol	10382	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
L-arabinitol	10413	DMSO-d6	DMSO-d6	0	0	0	1	0	0	0	0	0	0	0	0	0
2-methyl-2-butene	10424	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO{-}CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2,3,4-trimethylpentane	10427	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
2-ethylbutyl acetate	10428	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
4-butoxy-2-butanone	10431	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
3-hydroxy-3-methyl-2-butanone	10432	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	1
4-octanol	10440	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,5-dimethylfuran	10444	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	0
5-methyl-2-furaldehyde	10445	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	1	0
5-methylfurfuryl alcohol	10448	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	1	1	0	0	0	0	0
(S)-(+)-2-hydroxy-2-methylsuccinic acid	10463	DMSO-d6	DMSO-d6	1	1	0	1	0	0	0	0	0	0	1	0	0
L-2-hydroxy-4-methylvaleric acid	10517	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	1	0	0
formic acid	10523	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	0	0	0	0	1	0	0
cyclopentanol	10527	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
1,4-cyclohexanedione	10546	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
2-methoxy-4-methylphenol	10598	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
delta-valerolactone	10602	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
cyclobutanol	10603	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
isopentane	10633	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
valeraldehyde	10637	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	1	0
1,6-heptadiene	10723	$CDCl_3$	$CDCl_3$	0	1	0	0	0	1	0	0	0	0	0	0	0
2-methylpentyl formate	10865	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
dimethyl citraconate	10867	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
2,3-dihydroxybenzaldehyde	11037	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
2,2-dimethyloxirane	11268	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
beta-D-arabinopyranose	11512	DMSO-d6	D_2O	0	0	0	1	1	0	0	0	0	0	0	0	0
2,3-dimethylmaleic anhydride	11600	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	0	0	0	0
trans-furfurylideneacetone	11622	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	1	1	0	0	0	0	1
cis-4-hexen-1-ol	11748	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
3,5-dimethyl-2-furyl methyl ketone	11799	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	1
allyl valerate	12085	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
allyl isovalerate	12086	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
1-methylbutyl acetate	12271	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
tert-pentyl acetate	12272	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
isopentyl propionate	12283	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
m-anisic acid	12299	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	1	0	0
isophthalaldehydic acid	12383	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	0	0	0	1	1	0
1,1-dimethylcyclohexane	12465	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
2,2-dimethylpentane	12466	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
3,3-dimethylpentane	12467	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	0
4,4-dimethyl-2-pentanone	12468	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
ethyl 2-ethyl-3-oxobutyrate	12498	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
ethyl glycolate	12500	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
benzoylformic acid	12538	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	0	0	0	1	0	1
lactic acid	12682	DMSO-d6	DMSO-d6	1	0	0	1	0	0	0	0	0	0	1	0	0
(-)-methyl (R)-3-hydroxybutyrate	12745	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
(+)-methyl (S)-3-hydroxybutyrate	12746	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
(R)-(-)-2-methyl-2,4-pentanediol	12753	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
7-oxabicyclo(2.2.1)heptane	12789	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
(2R,4R)-(-)-2,4-pentanediol	12843	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
propyl valerate	12873	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2',5'-dihydroxyacetophenone	13050	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	0	0	1
2',6'-dihydroxyacetophenone	13051	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	1
2,2-dimethylsuccinic acid	13133	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
2,3-dimethylsuccinic acid	13134	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
ethoxyacetic acid	13172	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	1	0	0
ethyl 4-oxovalerate	13197	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
2-ethyl-2-methyl-1,3-propanediol	13198	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
ethyl pyruvate	13212	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	1
1-penten-3-one	13219	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	1
3-furoic acid	13253	$CDCl_3$	DMSO-d6	0	0	0	0	0	0	1	1	0	0	1	0	0
cis-2,trans-4-hexadiene	13275	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
1-hexen-3-ol	13279	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
hydroxyacetone	13286	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	1
3-hydroxybenzyl alcohol	13289	DMSO-d6	DMSO-d6	0	0	0	1	0	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	RO$-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-hydroxy-5-methoxybenzaldehyde	13300	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	1	0
4-hydroxy-3-methyl-2-butanone	13306	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1
2-hydroxy-2-methylbutyric acid	13307	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	1	0	0
methoxyacetone	13380	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	1
2-furyl methyl ether	13396	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
o-methylbenzyl alcohol	13429	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	1	0	0	0	0	0	0
3-methylbenzyl alcohol	13430	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	1	0	0	0	0	0	0
3-methyl-2-cyclohexen-1-ol	13439	$CDCl_3$	$CDCl_3$	1	0	1	1	0	1	0	0	0	0	0	0	0
3-methyl-2-cyclohexen-1-one	13440	$CDCl_3$	$CDCl_3$	1	0	1	0	0	1	0	0	0	0	0	0	1
(+-)-3-methylcyclopentanone	13444	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
methylenecyclohexane	13449	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
3-methylene-2-norbornanone	13451	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	1
5-methylene-2-norbornene	13452	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
2-methyl-3-pentanol	13488	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
4-methyl-1-pentanol	13490	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-norbornanone	13580	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
1-penten-3-ol	13593	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
4-penten-1-ol	13595	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
4-penten-2-ol	13596	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
3,4-methylenedioxyphenol	13685	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
vinylcyclohexane	13821	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
cis-bicyclo(3.3.0)octane-3,7-dione	15108	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	1
(R)-(-)-1,3-butanediol	15151	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(S)-(+)-1,3-butanediol	15152	DMSO-d6	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
cycloheptanecarboxylic acid	15250	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	1	0	0
1,3-cyclopentanedione	15258	DMSO-d6	DMSO-d6	0	0	1	0	0	0	0	0	0	0	0	0	1
cyclopropylmethanol	15263	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
ethyl cyclopropanecarboxylate	15380	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
trans-3-hexen-1-ol	15430	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
4-hydroxy-2-butanone	15443	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
methoxyacetaldehyde diethyl acetal	15506	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
o-methoxybenzyl alcohol	15507	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	1	1	0	0	0	0	0
m-methoxybenzyl alcohol	15508	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	1	1	0	0	0	0	0
2-methoxy-1,3-dioxolane	15510	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
alpha-methylene-gamma-butyrolactone	15532	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	1	1	0	0	0
3-methylglutaric anhydride	15535	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	0	0	0	0
methyl hydrogen succinate	15562	$CDCl_3$	$CDCl_3$	0	1	0	0	0	0	0	0	1	1	1	0	0
1,4-pentanediol	15603	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
pyruvaldehyde dimethyl acetal	15665	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	1
tetrahydropyran-4-ol	15725	$CDCl_3$	$CDCl_3$	0	0	1	1	1	0	0	0	0	0	0	0	0
1,2-dimethylcyclopentene	16141	$CDCl_3$	$CDCl_3$	1	0	1	0	0	1	0	0	0	0	0	0	0
ethyl 5-oxohexanoate	16142	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
2-methylcyclopentanone	16143	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
2-methyl-3-buten-1-ol	16151	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
trans-2,7-octadien-1-ol	16223	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
trans-3-hepten-2-one	16338	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	1
cis-1,3-cyclohexanediol	16407	DMSO-d6	DMSO-d6	0	0	1	1	0	0	0	0	0	0	0	0	0
trans-1,3-cyclohexanediol	16408	DMSO-d6	DMSO-d6	0	0	1	1	0	0	0	0	0	0	0	0	0
bicyclo(4.2.0)octa-1,3,5-triene	16423	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	1	0	0	0	0	0	0
4-ethyl-1,3-dioxolan-2-one	16486	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
4-ethylpyrocatechol	16708	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
(S)-tetrahydro-3-furanol	16712	$CDCl_3$	$CDCl_3$	0	0	1	1	1	0	0	0	0	0	0	0	0
1,1,2,2-tetramethylcyclopropane	16713	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	0
1-methoxyhexane	16726	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
2,2,4-trimethylcyclopentanone	16735	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
2,4,4-trimethylcyclopentanone	16736	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
(R)-(+)-tetrahydro-2-furancarboxylic acid	16737	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	1	0	0
(S)-(-)-tetrahydro-2-furancarboxylic acid	16738	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	1	0	0
3-butoxypropene	16798	$CDCl_3$	$CDCl_3$	1	1	0	0	1	1	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-ethyl-2-methyl-1,3-dioxolane	16846	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
1,3-dimethylbutyl acetate	17021	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
4,4-dimethyl-1,3-dioxane	17023	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
2,2-dimethyl-3-pentanol	17024	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,3-dimethyl-2-pentanol	17026	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,4-dimethyl-2-pentanol	17027	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3,3-dimethyl-2-pentanol	17028	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-ethyl-4,6-dimethyl-1,3,5-trioxane	17039	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
ethyl 2-methylvalerate	17042	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isobutyraldehyde diethyl acetal	17059	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
2-methoxyethyl butyrate	17070	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	1	1	0	0	0
methyl 2-methylvalerate	17075	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2-methylpentyl acetate	17076	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
ethylene diformate	17105	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	0	0	1	1	0	0	0
4-isopropoxy-2-butanone	17109	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO{-}CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3-methyl-1-heptanol	17114	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1,7-octadien-3-ol	17115	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
methyl glycolate	17245	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	0	0	1	1	0	0	0
2-methylhexanoic acid	17249	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
2-methyl-2-pentanol	17250	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-hydroxyethyl acetate	17252	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
ethyl L-(-)-lactate	17269	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
2,2-dimethyl-3-hexanol	18028	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
4-propoxy-2-butanone	18042	DMSO-d6	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	1
1,3,5-cyclohexanetriol	18062	DMSO-d6	D_2O	0	0	1	1	0	0	0	0	0	0	0	0	0
2-methoxy-5-methylphenol	18140	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
3-hydroxy-3-methylvaleric acid	18142	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	1	0	0
propioin	18160	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1
methyl methoxyacetate	18286	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	1	1	0	0	0
5-methoxy-m-cresol	18454	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	1	1	0	0	0	0	0
3-hydroxy-2,2-dimethylpropionic acid	18534	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
methyl 3,3-dimethoxypropionate	18559	$CDCl_3$	$CDCl_3$	0	1	0	0	1	0	0	0	1	1	0	0	0
2,2-dimethyl-5-oxotetrahydro-3-furoic acid	18639	DMSO-d6	DMSO-d6	1	0	1	0	0	0	0	0	1	1	1	0	0
1,2,3-cyclohexanetriol	18649	D_2O	D_2O	0	0	1	1	0	0	0	0	0	0	0	0	0
2-methylbutyl acetate	18676	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2,4-dimethyl-2,4-pentanediol	18802	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1,2-hexanediol	18855	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
methyl (R)-3-hydroxy-2-methylpropionate	18923	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
4-ethylcyclohexanone	18938	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	0	0	0	0	1
methyl (+-)-2-hydroxy-3-butenoate	18983	$CDCl_3$	$CDCl_3$	0	0	0	1	0	1	0	0	1	1	0	0	0
delta-caprolactone	18984	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
methyl 4-methoxy-3-oxobutyrate	18991	$CDCl_3$	$CDCl_3$	0	1	0	0	1	0	0	0	1	1	0	0	1
(S)-(+)-1,2-propanediol	18996	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
4-hydroxyisophthalaldehyde	19016	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	0	0	0	1	0
butoxyacetic acid	19032	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2,3-dimethyl-1-butanol	19069	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
ethyl 3,3-dimethylbutyrate	19110	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
allyl ethyl carbonate	19117	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
methyl 3,3-dimethyl-4-pentenoate	19123	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
2-methyl-4,5-dihydro-3(2H)-furanone	19132	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	1
3-methyl-4-oxovaleric acid	19169	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
2-methyl-4-oxovaleric acid	19170	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
1,4-dioxaspiro(4.5)decan-8-one	19189	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	1
2-hydroxy-gamma-butyrolactone	19196	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	1	1	0	0	0
isobutyl trans-crotonate	19243	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
allyl methyl carbonate	19245	$CDCl_3$	$CDCl_3$	0	0	0	0	0	1	0	0	1	1	0	0	0
2-methyl-1,3-propanediol	19339	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-methoxy-1-methylethyl acetate	19342	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	1	0	0	0
1,2-epoxyhexane	19364	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
cis-4-hepten-1-ol	19442	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
2,2-dimethylvaleric acid	19470	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
2-ethylfuran	19471	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
methyl 3-oxovalerate	19476	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
levoglucosenone	19477	$CDCl_3$	$CDCl_3$	0	0	0	0	1	1	0	0	0	0	0	0	1
2-methyl-4-pentenoic acid	19480	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0
3-octen-2-one	19484	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	1
methyl 2,2-dimethoxypropionate	19641	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	1	0	0	0
methyl 3-methoxy-2-methylpropionate	19673	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	1	1	0	0	0
3-ethyl-2-methyl-3-pentanol	19706	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
3-hydroxy-o-toluic acid	19709	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	1	0	0
3-methylpentyl acetate	19722	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
2,3,4-trimethyl-3-pentanol	19733	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
tert-butyl acrylate	19816	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
4-methyl hydrogen (R)-methylsuccinate	19841	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
(R)-(+)-methylsuccinic acid	19852	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
(S)-(-)-methylsuccinic acid	19853	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
(R)-(-)-2-methylglutaric acid	19854	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
(S)-(+)-2-methylglutaric acid	19855	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
3-methyl-1-hexanol	19860	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(R)-(+)-2-methyl-1,4-butanediol	19876	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(S)-(-)-2-methyl-1,4-butanediol	19877	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2-methoxyethyl acrylate	19887	$CDCl_3$	$CDCl_3$	0	0	0	0	1	1	0	0	1	1	0	0	0
dimethyl 1,1-cyclopropanedicarboxylate	19891	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	1	0	0	0
3-methyl-2(5H)-furanone	19905	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
1-cyclopropylethanol	19910	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
(E)-2-methyl-2-butenal	19912	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	1	0
methyl 2-hydroxy-2-methoxyacetate	19913	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	1	1	0	0	0
3-methyl-2-butenal	19914	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	1	0
ethyl 3-methoxypropionate	19919	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
methyl 3-methoxypropionate	19926	$CDCl_3$	$CDCl_3$	0	1	0	0	1	0	0	0	1	1	0	0	0
tetrahydro-2-furancarboxylic acid	19931	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	1	0	0
3-methyl-1,3-butanediol	19949	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
trans,trans-2,4-hexadienal	19961	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	1	0
3-propoxypropene	19963	$CDCl_3$	$CDCl_3$	1	1	0	0	1	1	0	0	0	0	0	0	0
2-tert-butoxyethanol	19973	$CDCl_3$	$CDCl_3$	1	0	0	1	1	0	0	0	0	0	0	0	0
5-hexen-1-ol	19976	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
propyl lactate	19977	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
quadricyclane	21034	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	0	0	0	0	0
ethyl 3-furoate	21157	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	1	1	0	0	0
exo-norboneol	21170	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
allylcyclopentane	21184	$CDCl_3$	$CDCl_3$	0	1	1	0	0	1	0	0	0	0	0	0	0
2,4,4-trimethyl-1-pentanol	21207	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
endo-norborneol	21209	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
o-xylene-alpha,alpha'-diol	21210	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	0	0	0	0	0	0
3-cyclopentyl-1-propanol	21222	$CDCl_3$	$CDCl_3$	0	1	1	1	0	0	0	0	0	0	0	0	0
2-hydroxyphenethyl alcohol	21223	$CDCl_3$	$CDCl_3$	0	1	0	1	0	0	1	1	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH_x^{ar}{=}$	$RO{-}CH_x^{ar}{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
1,2-octanediol	21366	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
2,2-pentamethylene-1,3-dioxolane	21564	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
4-hydroxy-2,5-dimethyl-3-hexanone	21593	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1
ethyl 3-oxovalerate	21610	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
tert-butyl 2,3-epoxypropyl ether	21686	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
methyl 4,4-dimethyl-3-oxovalerate	21794	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
trans-2-octenal	21812	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
(E)-2-methyl-2-pentenoic acid	21817	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0
bicyclo(2.2.2)oct-2-ene	21824	$CDCl_3$	$CDCl_3$	0	0	1	0	0	1	0	0	0	0	0	0	0
5,6-epoxy-1-hexene	21892	$CDCl_3$	$CDCl_3$	0	1	0	0	1	1	0	0	0	0	0	0	0
(+-)-3-methyltetrahydropyran	21907	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
ethyl isobutyrylacetate	22093	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
2,4,6-trihydroxybenzaldehyde	22255	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	0	1	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	cyCH_x	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3,4,5,6-tetrahydrophthalic anhydride	22262	DMSO-d6	DMSO-d6	0	0	1	0	0	1	0	0	1	0	0	0	0
2-ethoxyethyl methacrylate	22331	$CDCl_3$	$CDCl_3$	1	0	0	0	1	1	0	0	1	1	0	0	0
2,3-dimethyl-2-butanol	22360	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1-methylcyclopropanecarboxylic acid	22531	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	1	0	0
3,4-furandimethanol	22533	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	1	0	0	0	0	0
3,3-dimethyl-1,2-butanediol	22740	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
1,2-pentanediol	22741	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
trans-2-pentenal	22747	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
ethyl 2-methylcyclopropanecarboxylate	22772	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
3-methyl-1,5-hexadiene	22773	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	0	0
methyl cyclopropanecarboxylate	22839	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	0	0	1	1	0	0	0
2-methyl-3-hexanol	22840	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
ethyl 3-oxohexanoate	22841	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
2,6-dimethyl-p-benzoquinone	22842	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
tetrahydrofuran-3-ol	22881	$CDCl_3$	$CDCl_3$	0	0	1	1	1	0	0	0	0	0	0	0	0
3-hydroxy-4-methylbenzoic acid	22895	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	1	0	0
trans-2-methylcyclopentanol	22995	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
ethyl hydrogen fumarate	23123	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	1	0	0
tetramethyl orthocarbonate	23132	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
(+-)-5-methyl-2-hexanol	23157	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
cis-2-hexen-1-ol	23181	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
2,3,4-trihydroxybenzaldehyde	23184	DMSO-d6	DMSO-d6	0	0	0	0	0	0	1	1	0	0	0	1	0
(1-methylcyclopropyl)methanol	23188	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	0	0	0	0	0
7-octene-1,2-diol	23205	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
5-methoxyresorcinol	23237	DMSO-d6	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
o-(2-hydroxyethoxy)phenol	23267	DMSO-d6	DMSO-d6	0	0	0	1	1	0	1	1	0	0	0	0	0
(+-)-beta,beta-dimethyl-gamma-hydroxymethyl-gamma-butyrolactone	23325	$CDCl_3$	$CDCl_3$	1	0	1	1	0	0	0	0	1	1	0	0	0
6-methyl-2-heptanol	23333	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
1,6-dioxaspiro(4.4)nonane-2,7-dione	23352	DMSO-d6	DMSO-d6	0	0	1	0	0	0	0	0	1	1	0	0	0
1-acetylcyclohexene	23372	$CDCl_3$	$CDCl_3$	1	0	1	0	0	1	0	0	0	0	0	0	1
1-tert-butoxy-2-ethoxyethane	23406	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
1-tert-butoxy-2-methoxyethane	23407	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	0	0	0	0	0
dimethyl ethylidenemalonate	23443	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
4-hydroxybutyl acrylate	23446	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	1	1	0	0	0
cycloheptylmethanol	23489	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
2,4-dimethyl-1,3-pentadiene	23508	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	0
2-methyl-3-pentanone	23514	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	0	0	1
2(3H)-benzofuranone	23522	$CDCl_3$	$CDCl_3$	0	0	1	0	0	0	1	1	1	0	0	0	0
ethyl 3,3-dimethylacrylate	23580	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
3-cyclohexene-1,1-dimethanol	23631	$CDCl_3$	$CDCl_3$	0	0	1	1	0	1	0	0	0	0	0	0	0
(+-)-2-hydroxy-3-methylbutyric acid	23688	DMSO-d6	DMSO-d6	1	1	0	1	0	0	0	0	0	0	1	0	0
3’,5’-dihydroxyacetophenone	23697	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	0	0	1
(+-)-trans-1,2-cycloheptanediol	23698	DMSO-d6	DMSO-d6	0	0	1	1	0	0	0	0	0	0	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
ethyl pivalate	23704	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
3-furanmethanol	23734	$CDCl_3$	$CDCl_3$	0	0	0	1	0	0	1	1	0	0	0	0	0
(+-)-2-ethoxytetrahydrofuran	23769	$CDCl_3$	$CDCl_3$	1	0	1	0	1	0	0	0	0	0	0	0	0
2-methoxycyclohexanone	23819	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	1
ethyl methoxyacetate	23835	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	1	0	0	0
isopropylmalonic acid	28487	DMSO-d6	DMSO-d6	1	1	0	0	0	0	0	0	0	0	1	0	0
2-methylglutaric acid	32243	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
5-methylsalicylic acid	33556	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	1	0	0
3-penten-2-one	35146	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	0	0	0	0	1
methoxyacetic acid	36006	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	1	0	0
2-propoxyethanol	38702	$CDCl_3$	$CDCl_3$	1	1	0	1	1	0	0	0	0	0	0	0	0
1,2-heptanediol	41113	DMSO-d6	DMSO-d6	1	1	0	1	0	0	0	0	0	0	0	0	0
6-hepten-1-ol	41114	DMSO-d6	DMSO-d6	0	1	0	1	0	1	0	0	0	0	0	0	0
1,2,7-heptanetriol	41115	DMSO-d6	DMSO-d6	0	1	0	1	0	0	0	0	0	0	0	0	0
7-octen-1-ol	41125	$CDCl_3$	$CDCl_3$	0	1	0	1	0	1	0	0	0	0	0	0	0
2,2-bis(hydroxymethyl)butyric acid	41147	DMSO-d6	DMSO-d6	1	1	0	1	0	0	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2,6-dimethylhydroquinone	41173	DMSO-d6	DMSO-d6	1	0	0	0	0	0	1	1	0	0	0	0	0
9-oxabicyclo(6.1.0)nonane	41174	$CDCl_3$	$CDCl_3$	0	0	1	0	1	0	0	0	0	0	0	0	0
trans-4-hydroxycyclohexanecarboxylic acid	41224	DMSO-d6	DMSO-d6	0	0	1	1	0	0	0	0	0	0	1	0	0
L-(-)-lactic acid methyl ester	41227	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
alpha,alpha-dimethyl-gamma-butyrolactrone	41231	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	1	1	0	0	0
allylsuccinic anhydride	50201	$CDCl_3$	$CDCl_3$	0	1	1	0	0	1	0	0	1	0	0	0	0
(S,S)-(+)-2,3-butanediol	50480	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	0	0	0	0	0
2,2-dimethyl-hexanoic acid	50991	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	0
2,4,4-trimethyl-2-pentanol	51015	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
(R)-(+)-4-(methoxymethyl)-1,3-dioxolan-2-one	51056	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	1	1	0	0	0
(S)-(-)-4-(methoxymethyl)-1,3-dioxolan-2-one	51057	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	1	1	0	0	0
2-butylfuran	51163	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
(R)-propylene carbonate	51184	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0
(S)-propylene carbonate	51185	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x{=}$	$CH^{ar}_x{=}$	$RO{-}CH^{ar}_x{=}$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
2-propylfuran	51186	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	0
ethyl 2-hydroxyvalerate	51281	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0
trimethyl orthoisobutyrate	51289	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
butylsuccinic anhydride	51315	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	1	0	0	0	0
tetrahydrofurfuryl acrylate	51369	$CDCl_3$	$CDCl_3$	0	0	1	0	1	1	0	0	1	1	0	0	0
meso-1,2:3,4-diepoxybutane	51474	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	0	0	0	0	0	0	0
1,4-anhydroerythritol	51597	$CDCl_3$	$CDCl_3$	0	0	0	1	1	0	0	0	0	0	0	0	0
cyclobutanemethanol	51636	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
ethyl 4-pentenoate	51639	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
trimethyl orthopropionate	51641	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
ethyl 2-methyl-3-pentenoate	51675	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
2,2,3,3-tetramethylcyclopropanecarboxylic acid	51811	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	1	0	0
3,3-dimethylcyclohexanone	51915	$CDCl_3$	$CDCl_3$	1	0	1	0	0	0	0	0	0	0	0	0	1
4,10-dioxatricyclo(5.2.1.0(2,6))decan-8-en-3-one	52003	$CDCl_3$	$CDCl_3$	0	0	1	0	1	1	0	0	1	1	0	0	0
cis-5-octenoic acid	52258	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	1	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
1,1-dimethyl-(2-propenyl) acetate	52274	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	0	0	1	1	0	0	0
2-butenal diethyl acetal	52313	$CDCl_3$	$CDCl_3$	1	0	0	0	1	1	0	0	0	0	0	0	0
trans-4-(hydroxymethyl)cyclohexanol	52394	DMSO-d6	DMSO-d6	0	0	1	1	0	0	0	0	0	0	0	0	0
propyl 2-methylbutyrate	52451	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
isopropyl 2-methylbutyrate	52527	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
5-oxooctanoic acid	52568	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	0	0	1	0	1
isopropyl lactate	52572	$CDCl_3$	$CDCl_3$	1	0	0	1	0	0	0	0	1	1	0	0	0
1-(5-methyl-2-furyl)-1,2-propanedione	52577	$CDCl_3$	$CDCl_3$	1	0	0	0	0	0	1	1	0	0	0	0	1
trans-2-butene-1,4-diol	52670	DMSO-d6	DMSO-d6	0	0	0	1	0	1	0	0	0	0	0	0	0
butanal diethyl acetal	52725	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
hexanal dimethyl acetal	52728	$CDCl_3$	$CDCl_3$	1	1	0	0	1	0	0	0	0	0	0	0	0
ethyl 4-methylpentanoate	52735	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
trans-3-hexenyl acetate	52738	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	0
2-methylbutyl propionate	52742	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	0
methyl 3-hydroxybutyrate	52752	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH_x^{ar}=$	$RO-CH_x^{ar}=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
propyl levulinate	52753	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
furfuryl formate	52770	$CDCl_3$	$CDCl_3$	0	0	0	0	0	0	1	1	1	1	0	0	0
4-(methoxymethyl)phenol	52788	$CDCl_3$	$CDCl_3$	0	0	0	0	1	0	1	1	0	0	0	0	0
isopropyl levulinate	52796	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	0	0	1	1	0	0	1
allyl levulinate	52797	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	1	1	0	0	1
cis-3-heptenol	52800	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
1,3-octanediol	52802	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	0
delta-heptalactone	52805	$CDCl_3$	$CDCl_3$	1	1	1	0	0	0	0	0	1	1	0	0	0
methyl trans-4-hydroxycyclohexanecarboxylate	53045	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	1	1	0	0	0
1-hydroxy-4-methyl-2-pentanone	53127	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	0	0	0	0	1
1-ethoxyethyl acetate	53220	$CDCl_3$	$CDCl_3$	1	0	0	0	1	0	0	0	1	0	0	0	0
cyclopentanemethanol	53255	$CDCl_3$	$CDCl_3$	0	0	1	1	0	0	0	0	0	0	0	0	0
trans-2-octenol	53313	$CDCl_3$	$CDCl_3$	1	1	0	1	0	1	0	0	0	0	0	0	0
2-furyl-2-propanone	53328	$CDCl_3$	$CDCl_3$	1	1	0	0	0	0	1	1	0	0	0	0	1
2-methyl-3-(2-furyl)-2-propenal	53329	$CDCl_3$	$CDCl_3$	1	0	0	0	0	1	1	1	0	0	0	1	0

Name	SDBS No.	Solvent ^{1}H	Solvent ^{13}C	CH_3	CH_x	$cyCH_x$	CH_xOH	CH_xO	$CH_x=$	$CH^{ar}_x=$	$RO-CH^{ar}_x=$	COOR	$ROOCH_x$	COOH	CO^{ald}	CO^{ket}
3-hydroxy-alpha-methylbenzyl alcohol	53460	DMSO-d6	DMSO-d6	1	0	0	1	0	0	1	1	0	0	0	0	0
2-methyl-2-pentenal	53524	$CDCl_3$	$CDCl_3$	1	1	0	0	0	1	0	0	0	0	0	1	0
methyl 3-hydroxyhexanoate	53796	$CDCl_3$	$CDCl_3$	1	1	0	1	0	0	0	0	1	1	0	0	0

B Supporting Information for Chapter 3

B.1 Data

B.1.1 Processing of Pure-component Spectra

The raw data were downloaded from the NMRShiftDB data bank [66, 136] and the Biological Magnetic Resonance Data Bank (BMRB) [67, 137], whereby the format of the retrieved data was different as described in the following.

B.1.1.1 NMRShiftDB

The NMRShiftDB data bank [66] reports the raw data in a so-called structural data file (.sdf) [138], from which individual files in the NMReData format [139] were retrieved using RDKit [70]. Specifically, each NMReData file contains a component-specific ID ("MoleculeID"), a molfile (.mol) (including information about the atoms, bonds, connectivity, and coordinates of the respective component), and the assignment of the chemical shifts of the peaks in the NMR spectrum to individual atoms of the component. The assigned chemical shifts of all carbon and proton peaks were extracted from the NMReData file of each component. If multiple chemical shifts for the same atom were given in *one* NMReData file, the arithmetic mean of the shifts is used. If, furthermore, multiple NMReData files for the same component (same MoleculeID) were available, again, the arithmetic mean was used for averaging the chemical shifts.

B.1.1.2 BMRB

The BMRB database [67] reports the data in experimental sets ("bmsexxxxxx"), where "xxxxxx" is a 6-digit numerical identifier of the experimental set. Experimental sets were downloaded from the file transfer protocol (FTP) server [137]. Duplicate experimental sets for the same component were discarded, and only the first experimental set (corresponding to a lower 6-digit identifier) was kept. Duplicates were identified based on the simplified molecular-input line-entry system (SMILES) of the components. The experimental sets include, among other information that was not used here, a file in the NMRStar format [140], which was processed with the NMRPystar package [141], and a molfile for the component.

In most cases, the NMRStar file included an “atom label assignment tool using INCHI string” (ALATIS) [142] as identifier for the component, which is based on the international chemical identifier (INCHI) format [143] and, in addition, NMR spectroscopic information. Each ALATIS was then converted into a molfile by RDKit [70]. In very few cases, the ALATIS was missing. In these cases, the additionally downloaded molfile was used if the keyword “alatis” appeared in the header to ensure a consistent numbering of the atoms [142]; otherwise, the experimental set was discarded.

Subsequently, the chemical shifts of all carbon and proton atoms in one NMRStar file were extracted and arithmetically averaged if multiple shifts for the same atom were reported.

B.1.2 Further Processing Steps

Only data of components composed exclusively of the elements C, H, and O were kept since the structural groups considered in Chapter 3 are only made up of these elements, cf. Table 4. Furthermore, only components for which spectra of both ^{1}H and ^{13}C were available were adopted for the data set. In an additional step, the chemical shifts of all peaks of labile protons, i.e., protons that are directly bonded to oxygen, were discarded.

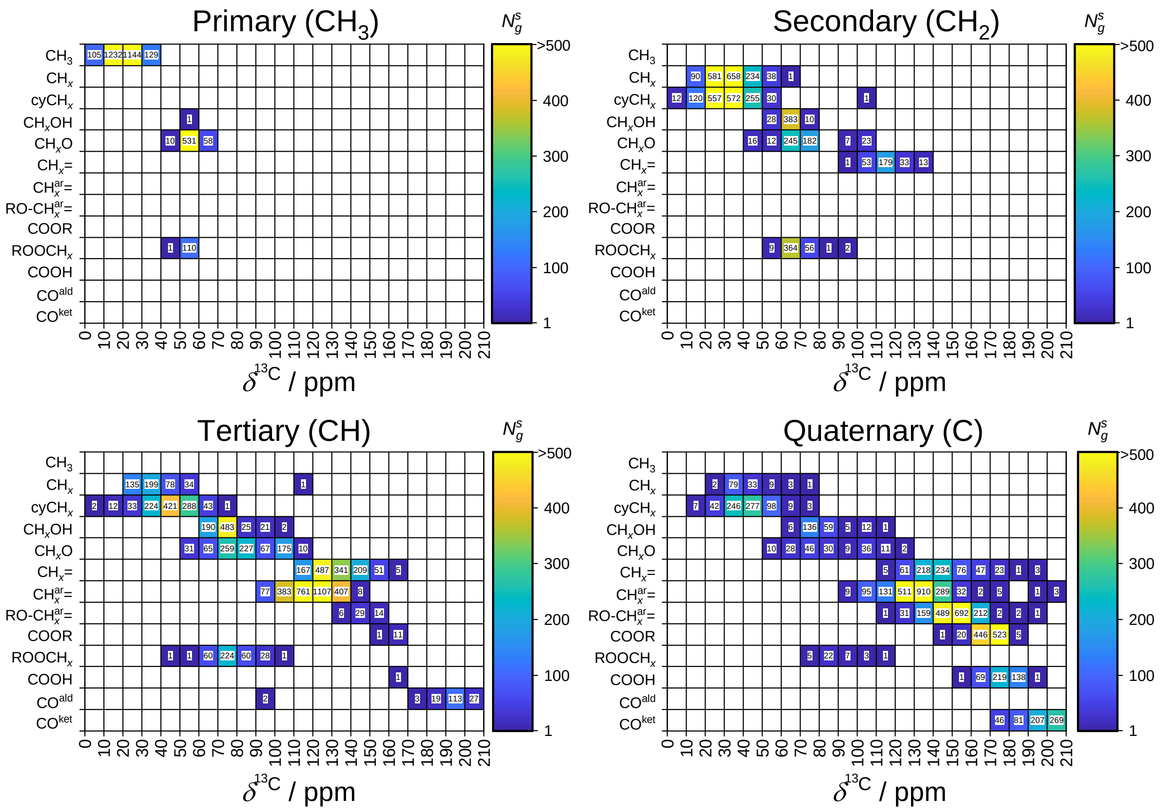
All carbon atoms were then assigned to one structural group by comparing the SMARTS string against the molfiles of the components by RDKit [70]. Only components in which each carbon atom could be assigned to one of the considered structural groups were kept. Carbon atoms to which a structural group was assigned but for which no chemical shift is reported were ignored in the generation of the input and output data. Additionally, some spectra were identified as erroneous and removed during manual consideration. A typical example for this case is that an unusual chemical shift for a structural group is observed, e.g., caused by miss-assignments of chemical shifts to the wrong atom.

The substitution degree of each carbon atom in a component was determined by RDKit [70] using the molfile of the respective component, i.e., to classify it as primary (3 bonded protons), secondary (2 bonded protons), tertiary (1 bonded proton), or quaternary (no bonded proton).

The final data set containing data from both data banks was finally merged with a priority on the NMRShiftDB; duplicate components from the BMRB were identified by comparing the SMILES of the components. The input and output vectors were finally generated as described in Chapter 3.

B.2 Data Overview by Consideration of Substitution Degrees of Carbon Atom

Figure B.1 indicates the frequency of the 13 distinguished structural groups in the considered components and also in which sections of the ^{13}C NMR spectrum the respective peaks appear. In contrast to Figure 9, the data is subdivided here based on the substitution degree of the respective carbon atom.



Primary (CH_3)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180	180–190	190–200	200–210
CH_3	105	1232	1144	129																	
CH_x																					
$cyCH_x$																					
CH_xOH						1															
CH_xO					10	531	58														
$CH_x=$																					
$CH_x^{ar}=$																					
$RO-CH_x^{ar}=$																					
COOR																					
$ROOCH_x$					1	110															
COOH																					
CO^{ald}																					
CO^{ket}																					

Secondary (CH_2)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180	180–190	190–200	200–210
CH_3																					
CH_x		90	581	658	234	38	1														
$cyCH_x$	12	120	557	572	255	30					1										
CH_xOH						28	383	10													
CH_xO					16	12	245	182		7	23										
$CH_x=$										1	53	179	33	13							
$CH_x^{ar}=$																					
$RO-CH_x^{ar}=$																					
COOR																					
$ROOCH_x$						9	364	56	1	2											
COOH																					
CO^{ald}																					
CO^{ket}																					

Tertiary (CH)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180	180–190	190–200	200–210
CH_3																					
CH_x			135	199	78	34						1									
$cyCH_x$	2	12	33	224	421	288	43	1													
CH_xOH							190	483	25	21	2										
CH_xO						31	65	259	227	67	175	10									
$CH_x=$												167	487	341	209	51	5				
$CH_x^{ar}=$										77	383	761	1107	407	8						
$RO-CH_x^{ar}=$														6	29	14					
COOR																1	11				
$ROOCH_x$					1	1	60	224	60	28	1										
COOH																	1				
CO^{ald}										2								3	19	113	27
CO^{ket}																					

Quaternary (C)	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180	180–190	190–200	200–210
CH_3																					
CH_x			2	79	33	9	3	1													
$cyCH_x$		7	42	246	277	98	9	3													
CH_xOH							6	136	59	5	12	1									
CH_xO						10	28	46	30	9	36	11	2								
$CH_x=$												5	61	218	234	76	47	23	1	3	
$CH_x^{ar}=$										9	95	131	511	910	289	32	2	5		1	3
$RO-CH_x^{ar}=$												1	31	159	489	692	212	2	2	1	
COOR															1	20	446	523	5		
$ROOCH_x$								5	22	7	3	1									
COOH																1	69	219	138	1	
CO^{ald}																					
CO^{ket}																		46	81	207	269

Figure B.1: Positions of the peaks of the 2839 components from the data set in the ^{13}C NMR spectrum for different substitution degrees (P, S, T, Q) of the carbon atoms. The color code and the numbers inside the cells denote N_g^s, which is the number of components that contain the structural group g (row) that induces a peak in the section s of the spectrum (column). White cells refer to $N_g^s = 0$.

B.3 Modeling and Computational Details

A non-linear soft-margin support vector classification (SVC) approach was used in the present thesis, which a priori contains one hyperparameter C that controls the penalty term in the SVC optimization problem for misclassified training data points [47]. The non-linearity was introduced here by using the radial basis function (RBF) kernel, cf. Eq. (3), which results in an additional hyperparameter γ.

In general, binary SVCs for combinations of structural groups and sections in the ^{13}C NMR spectrum were trained in the present thesis, which predict the presence or absence of the respective structural group in the respective section of the NMR spectrum. The training of the SVCs itself was done using scikit-learn [55] using a tolerance of 10^{-4} and using the option of balanced class weighting to account for label imbalances. Only for those structural groups that can, in principle, based on the training set, show peaks in the respective section of the ^{13}C NMR spectrum, a binary SVC was trained. Prior to training the SVCs, a zero-variance filter [55] was applied to the respective training data set to remove entries in the input vector (corresponding to sections s in the binned NMR spectra) that did not show any variance in the whole training data set. This was typically the case for sections s in the ^{13}C NMR spectrum in which, in general, no peaks associated with carbons with a specific substitution degree are present, e.g., the peaks of primary carbon atoms were exclusively observed at chemical shifts ranging from 0-70 ppm in the ^{13}C NMR spectrum.

Nested cross-validation (CV) [54] was carried out to determine the generalization error of the method by ensuring that the test set, which is built in the outer loop of the CV, is neither used for training the method nor for optimizing the hyperparameters. An outer loop with 10 folds was chosen here, whereas the hyperparameters were optimized in an inner loop with 5 folds. The different folds were built using the learning library scikit-multilearn [57] that uses the multi-label data stratification technique based on Ref. [58] to reduce imbalances between classes in the different folds.

In the inner loop of the nested CV, the hyperparameters C and γ were optimized using a Bayesian optimization approach with the scikit-optimize toolbox [72] and built-in function “BayesSearchCV” with 100 iterations and a hyperparameter range of 10^{-4} to 10^{4} for C and 10^{-5} to 10^{1} for γ; the target was thereby to maximize the F_1^{macro} score averaged (by the arithmetic mean) over the validation set of the five inner folds. The F_1^{macro} is the arithmetic mean over all $F_{1,g}$ scores of all groups g in the respective section with at least 10 positive examples in the outer loop, cf. Figure 10. In the default scikit-learn setting, the $F_{1,g}$ score is set to zero if it is ill-defined, i.e., when there are no predicted labels for class g and no true positives for class g, for details see the documentation of scikit-learn [144] and Section A.4.1. The behavior is changed by setting the “zero-divison” to 1, i.e., if the SVC correctly predicts the absence of class g (true negative), the $F_{1,g}$ score is set to 1. To further improve the robustness of the algorithms, the data set in the inner loop of the CV was augmented as follows: first, binary combinations of all data points in the inner loop were generated by the addition of the input and output vectors of the respective pure components, resulting in $N_{\text{inner}}(N_{\text{inner}} - 1)/2$ synthetic binary mixtures, where N_{inner} is the number of pure-component data points in the inner loop. From this combination list, $N_{\text{inner}}/2$ samples were randomly chosen and appended to the data set in the inner loop. After the optimal hyperparameters in the inner loop were found, the method was trained with all data from the inner loop and applied to the test data that were held back in the outer loop.

The same settings as described above were applied for training the final method for the application to the mixture spectra. For the training and optimization of the final method, the outer loop was not needed, and the data set was split into 90% training set (for training the SVCs) and 10% validation set (for optimization of hyperparameters). The process was repeated ten times so that each data point was exactly once in the validation set. Note that also in the training of the final method, only $F_{1,g}$ scores of all groups g in the respective section with at least 10 positive examples were taken into account to build the F_1^{macro}, cf. Figure 10.

After the training, the decision function of a sample $\boldsymbol{x}^*$ can be calculated for each group g in each section s as:

$$d_g^s(\boldsymbol{x}^*) = \sum_{i\in\text{SV}} (y_i \cdot \alpha_i)_g^s K_g^s(\boldsymbol{x}_i, \boldsymbol{x}^*) + b_g^s \tag{B.1}$$

where the sum is over all support vectors (SV), i.e., those training data points for which the dual coefficients ($\alpha_{i,g}^s$) are unequal to zero, and b_g^s is the offset [47, 144]. K_g^s is the RBF kernel for structural group g in section s of the binned ^{13}C NMR spectrum, cf. Eq. (3). Note that the hyperparameter γ in the RBF kernel is the same for all groups in one section. The interested reader can find an extensive description of the theory of SVC in Ref. [47].

B.3.1 Application of Method to Mixture Data

For the application of the SVC to experimental spectra of mixtures, the following heuristic was applied to ensure that to each peak p in the decoupled ^{13}C NMR spectrum of a mixture, exactly one structural group was assigned: for each peak p in the spectrum, the structural group with the highest value of the decision function in the respective section (in which peak p appears) was assigned. Thereby, also the information on the substitution degree was taken into account. For instance, if for a specific peak $p\star$ in section *s* the CH_3 group has the highest value of the decision function but the ^{13}C DEPT NMR experiments reveal that the associated structural group does *not* contain a primary carbon atom, the CH_3 group was not assigned (but the structural group with the highest value of the decision function among all groups that complied with the information on the substitution degree). This unique assignment of structural groups to peaks in the ^{13}C NMR spectrum also facilitates the retrieval of quantitative information about structural groups by integration of the peaks.

B.4 Experimental Methods

In Table B.1, information on the chemicals used for the preparation of the mixtures I-IV that were studied in Chapter 3, cf. Table 5, is summarized.

Table B.1: Suppliers and purities of chemicals used in Chapter 3. Purities are indicated as specified by the suppliers.

Chemical	Formula	Supplier	Purity
Acetone	C_3H_6O	Sigma Aldrich	≥99.90%
Butanal	C_4H_8O	Sigma Aldrich	≥99.50%
2-Butanone	C_4H_8O	Sigma Aldrich	≥99.70%
Cyclohexanone	$C_6H_{10}O$	Sigma Aldrich	≥99.80%
Deuterium oxide	D_2O	Cambridge Isotope Laboratories	≥99.50%
1,4-Dioxane	$C_4H_8O_2$	Sigma Aldrich	≥99.80%
Ethyl acetate	$C_4H_8O_2$	Sigma Aldrich	≥99.50%
Malic acid	$C_4H_6O_5$	Sigma Aldrich	≥99.20%
1-Propanol	C_3H_8O	Honeywell	≥99.50%
1-Octanol	$C_8H_{18}O$	Merck	≥99.00%
Oleic acid	$C_{18}H_{34}O_2$	Alfa Aesar	≥94.00%
Tert-butylhydroquinone	$C_{10}H_{14}O_2$	Merck	≥97.00%
Tetramethylsilane (TMS)	$C_4H_{12}Si$	Sigma Aldrich	≥99.90%

B.4.1 Preparation of Samples and Acquisition of NMR Spectra

Samples of mixtures I-IV were prepared gravimetrically in 20 ml glass vessels using a balance of Mettler Toledo XS603S Delta Range with an accuracy of ±0.001 g. The mass of each sample was about 10 g. A small amount of 1,4-dioxane was added to the samples of mixtures I and II and used as a reference in ^{13}C and ^{1}H NMR spectroscopy. To the samples of mixtures III and IV, a small amount of tetramethylsilane (TMS) was added and used as a reference in ^{13}C and ^{1}H NMR spectroscopy, respectively. Approx. 0.7 ml of each sample was transferred to a 5 mm NMR vial.

NMR spectra were recorded with an 80 MHz (proton frequency) Spinsolve Carbon NMR spectrometer from Magritek (Aachen, Germany). All measurements were performed at 28.5°C, resembling the internally set operating temperature of the NMR spectrometer. ^{1}H NMR spectra were recorded using a flip angle of 90°, a relaxation delay of 60 s, 16 scans, an acquisition time of 6.4 s, and a bandwidth of about 20 ppm. ^{1}H decoupled ^{13}C NMR spectra for the aqueous mixtures I and II were recorded with a flip angle of 45°, a relaxation delay of 15 s, 2048 scans, an acquisition time of 3.2 s and a bandwidth of about 320 ppm using nuclear Overhauser effect (NOE) enhancement. ^{1}H decoupled ^{13}C NMR spectra for the organic mixtures III and IV were recorded with a flip angle of 90°, a relaxation delay of 120 s, an acquisition time of 3.2 s, 256 scans, and a bandwidth of about 320 ppm. Additionally, on the ^{13}C and ^{1}H NMR spectra of all mixtures, exponential line broadening from MNova with 1.0 Hz was applied. Furthermore, ^{13}C DEPT NMR spectra were recorded with pulse angles of 45°, 90°, and 135°, a relaxation delay of 30 s, an acquisition time of 3.2 s, 256 scans and a bandwidth of about 320 ppm. All DEPT spectra were processed using MNova, whereby an exponential line broadening of 0.2 Hz was applied.

MNova was used for automatic baseline and phase correction. Then, the automatic peak picking available in MNova was applied to identify the peaks in the spectra, whereby artifacts were manually removed. Additionally, DEPT 45 spectra were used to determine peaks barely visible in the standard ^{13}C spectrum.

B.4.2 Determination of Substitution Degree

The substitution degree of each carbon atom was determined using the DEPT 90, DEPT 135, and standard ^{13}C NMR spectra. Primary, secondary, and tertiary carbon atoms were assigned using the standard phase angle behavior of such peaks in DEPT spectra [25]. Primary carbon atoms show almost no peak in DEPT 90 spectra and a positive peak in DEPT 135 spectra. In contrast, secondary carbon atoms show a negative peak in DEPT 135 spectra. Only tertiary carbon atoms show a peak in DEPT 90 spectra. Since quaternary carbon atoms show no peak in either of the DEPT spectra, they could easily be determined by comparison to the standard ^{13}C NMR spectra.

B.4.3 Determination of Labile Protons

To demonstrate the detection of labile protons, ^{1}H-^{13}C heteronuclear single quantum coherence (HSQC) NMR experiments were carried out on a 60 MHz (proton frequency) Spinsolve Carbon benchtop NMR spectrometer from Magritek for the organic mixtures. HSQC multiplicity-edited spectra of both organic mixtures were recorded with 16 scans, 512 steps, and a repetition time of 3 seconds. To reduce the measurement time, non-uniform sampling (NUS) with a density of 25% was applied. The HSQC spectra were phase-corrected and aligned to the TMS reference peak in the ^{1}H-^{13}C dimensions. ^{1}H-^{13}C HSQC experiments show carbon-proton single bond correlations, so labile protons that are directly bonded to oxygen should, in principle, show no correlation with any carbon in the sample [25]. In some cases, there might be a residual peak that, however, shows a much weaker correlation than the peak resembling the direct connection between the proton and a carbon. By comparison to the respective 1D NMR spectra, peaks in the ^{1}H dimension showing no or little correlation to the ^{13}C dimension in the HSQC spectrum were labeled as labile. Figure B.2 shows the HSQC spectrum of mixture IV as an example. Here, for each proton peak that is directly bonded to a carbon, a correlating carbon peak in the ^{13}C NMR spectrum can be observed. Only the proton peak at 11.8 ppm shows no correlation and is therefore labeled as a labile proton.

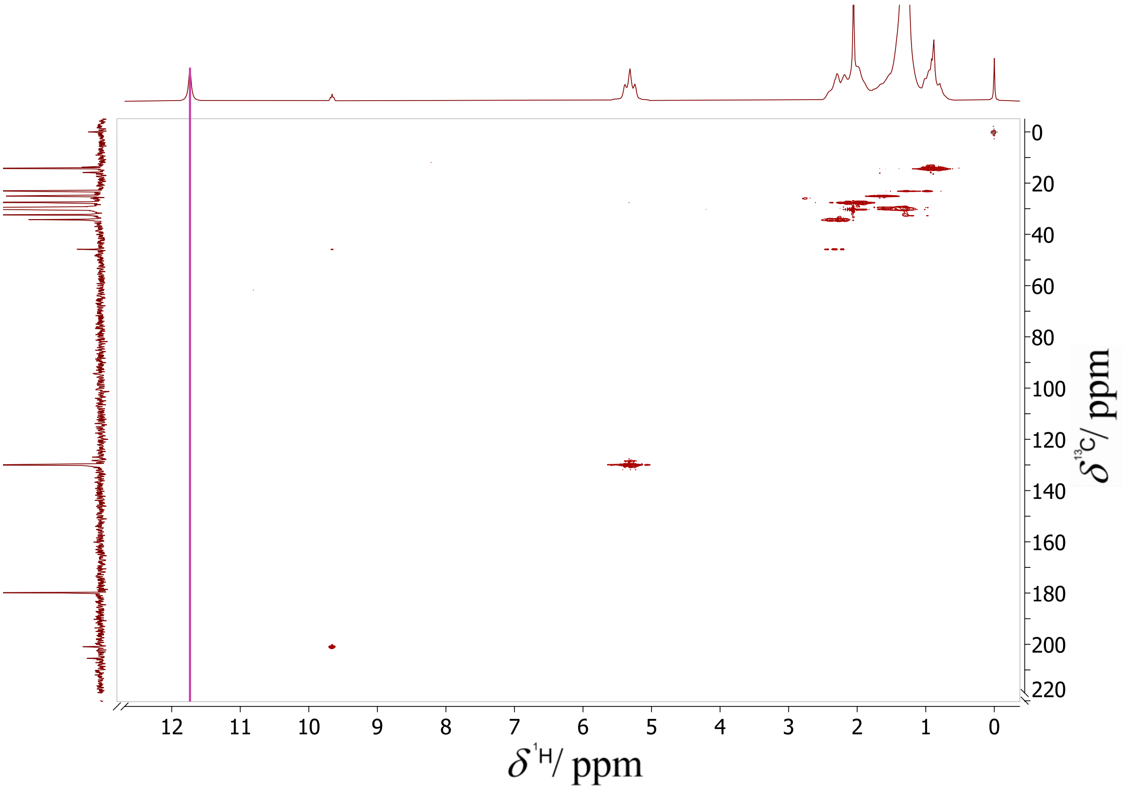


Figure B.2: ^{1}H-^{13}C HSQC spectrum of mixture IV with ^{1}H NMR as horizontal and the ^{13}C NMR as vertical dimension. The labile proton shown at 11.8 ppm in the ^{1}H NMR spectrum is marked with a magenta line. The projected spectra at the left and the top are 1D NMR spectra of the mixtures.

B.4.4 Quantitative Evaluation and Comparison to Ground Truth

Since NOE-enhancement was applied during the recording of the spectra of the aqueous mixtures I and II, no quantitative evaluation was possible. The application of NOE-enhancement was inevitable due to the low concentration of the mixtures and the otherwise extremely long measurement times necessary for ^{13}C detection. For the quantitative evaluation of the studied organic mixtures III and IV, cf. Table 5, the peak fitting method implemented in MNova (global spectral deconvolution) was applied to determine the area of the peaks. By summing up the areas of all peaks belonging to group g, a total area A_g was obtained. As each of the considered structural groups contains exactly one carbon nucleus, the mole fraction x_g^{pred} of each group g in the mixture as predicted based on the results of the developed method was calculated by using Eq. (A.1).

The predicted group mole fractions x_g^{pred} were compared to the true group mole fractions x_g, which were calculated from the known mole fractions x_i of the components i in each mixture and the known stoichiometry of each component i by using Eq. (A.2).

Table B.2: Stoichiometric coefficients ν_g^i of structural groups g in components i.

Component i	Structural group g	Stoichiometric coefficient ν_g^i
Acetone	CH_3	2
	CO^{ket}	1
Butanal	CH_3	1
	CH_x	2
	CO^{ald}	1
2-Butanone	CH_3	2
	CH_x	1
	CO^{ket}	1
Cyclohexanone	$cyCH_x$	5
	CO^{ket}	1
Ethyl acetate	CH_3	2
	COOR	1
	$ROOCH_x$	1
Malic acid	CH_x	1
	CH_xOH	1
	COOH	2
1-Propanol	CH_3	1
	CH_x	1
	CH_xOH	1
1-Octanol	CH_3	1
	CH_x	6
	CH_xOH	1
Oleic acid	CH_3	1
	CH_x	14
	$CH_x=$	2
	COOH	1
Tert-butylhydroquinone	CH_3	3
	CH_x	1
	$CH_x^{ar}=$	4
	$RO-CH_x^{ar}=$	2

B.5 NMR Spectra of Mixtures

In the following, the recorded ^{13}C and ^{1}H NMR spectra of mixtures I-IV (cf. Table 5) are shown. In general, the water peak was ignored for the evaluation. For the method described in Chapter 3, chemical shifts of labile protons were furthermore ignored and only used for the variant described in Section B.6. Small peaks that could not be assigned to the main components of the mixtures (presumably from contaminations of the utilized chemicals) were not considered. Additionally, the recorded DEPT spectra with phase angles of 45°, 90°, and 135° of mixtures I-IV are shown.

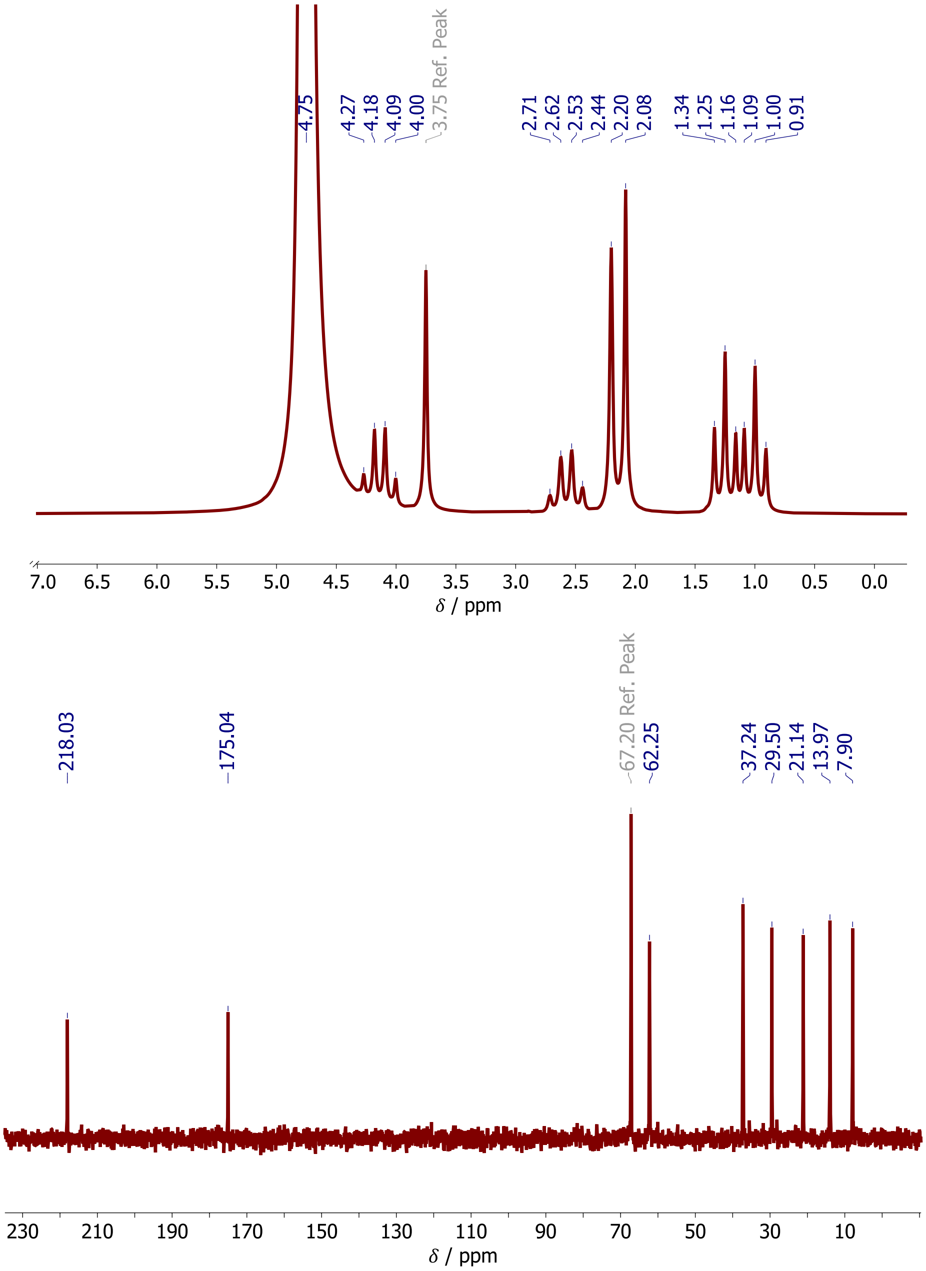


Figure B.3: ^{1}H and ^{13}C NMR spectra of mixture I (2-butanone, ethyl acetate, water), cf. Table 5. The peak at 67.2 ppm in the ^{13}C and at 3.75 ppm in the ^{1}H NMR spectrum belongs to 1,4-dioxane, which was used as reference [145]. Top: ^{1}H NMR spectrum. Bottom: ^{13}C NMR spectrum.

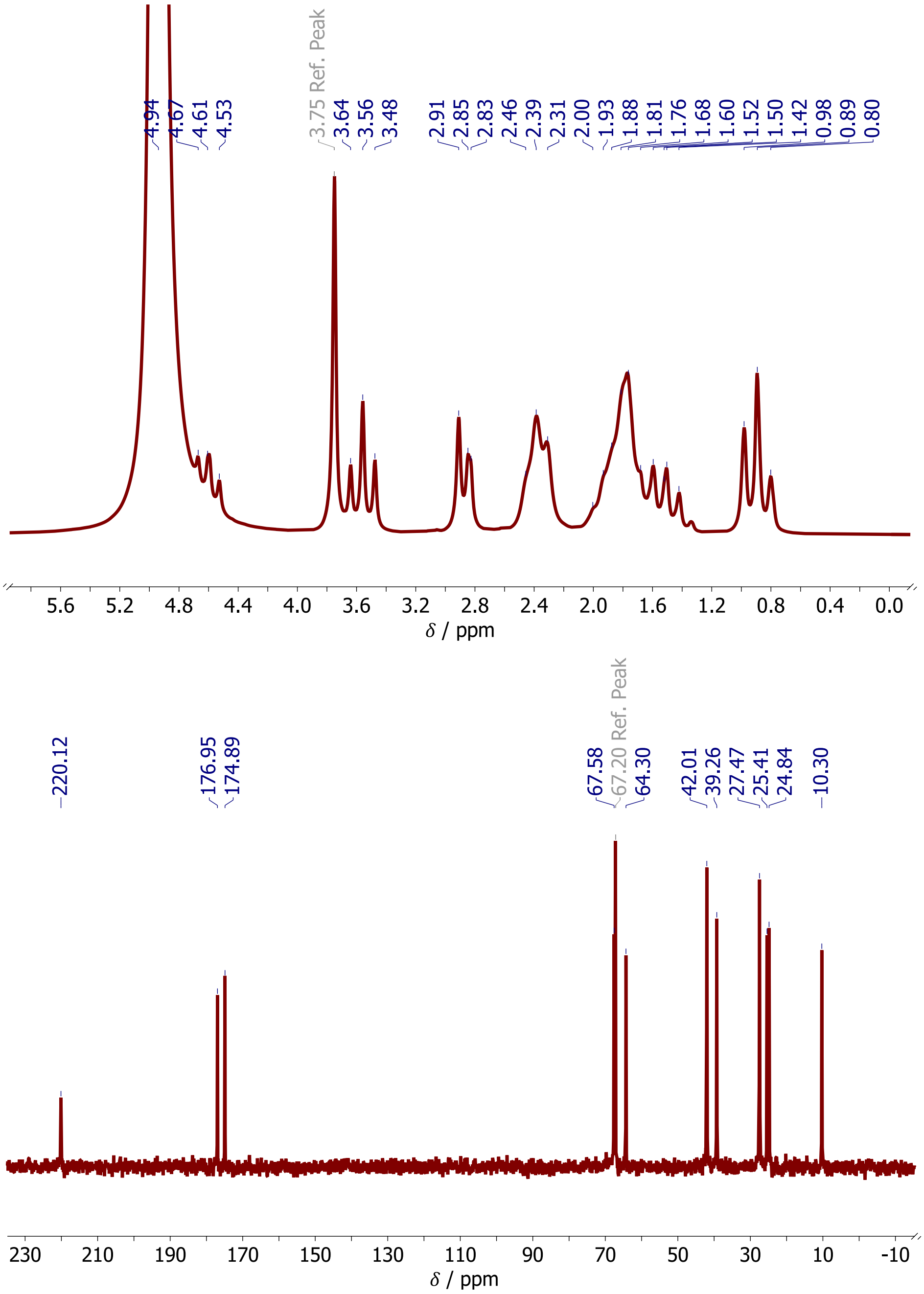


Figure B.4: ^{1}H and ^{13}C NMR spectra of mixture II (cyclohexanone, malic acid, 1-propanol, water), cf. Table 5. The peak at 67.2 ppm in the ^{13}C and at 3.75 ppm in the ^{1}H NMR spectrum belongs to 1,4-dioxane, which was used as reference [145]. Top: ^{1}H NMR spectrum. Bottom: ^{13}C NMR spectrum.

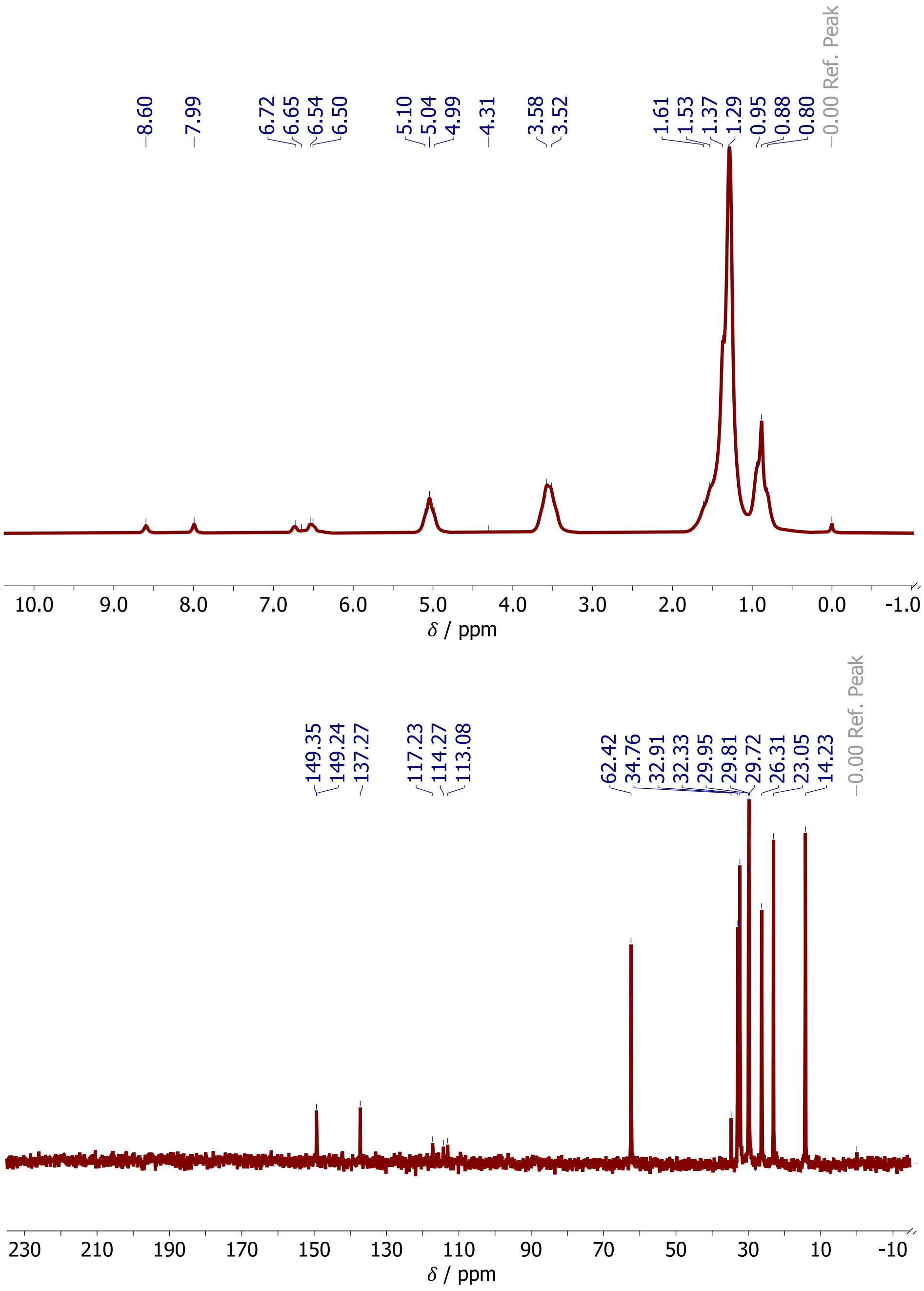


Figure B.5: ^{1}H and ^{13}C NMR spectra of mixture III (1-octanol, tert-butylhydroquinone), cf. Table 5. The peak at 0 ppm belongs to TMS, which was used as reference. Top: ^{1}H NMR spectrum. Bottom: ^{13}C NMR spectrum. The peak at 29.72 ppm in the ^{13}C NMR spectrum was ignored for the quantitative evaluation, since no meaningful peak area could be determined. The peaks at 4.99-5.10 ppm and 7.99/8.60 ppm in the ^{1}H NMR spectrum are classified as stemming from labile protons.

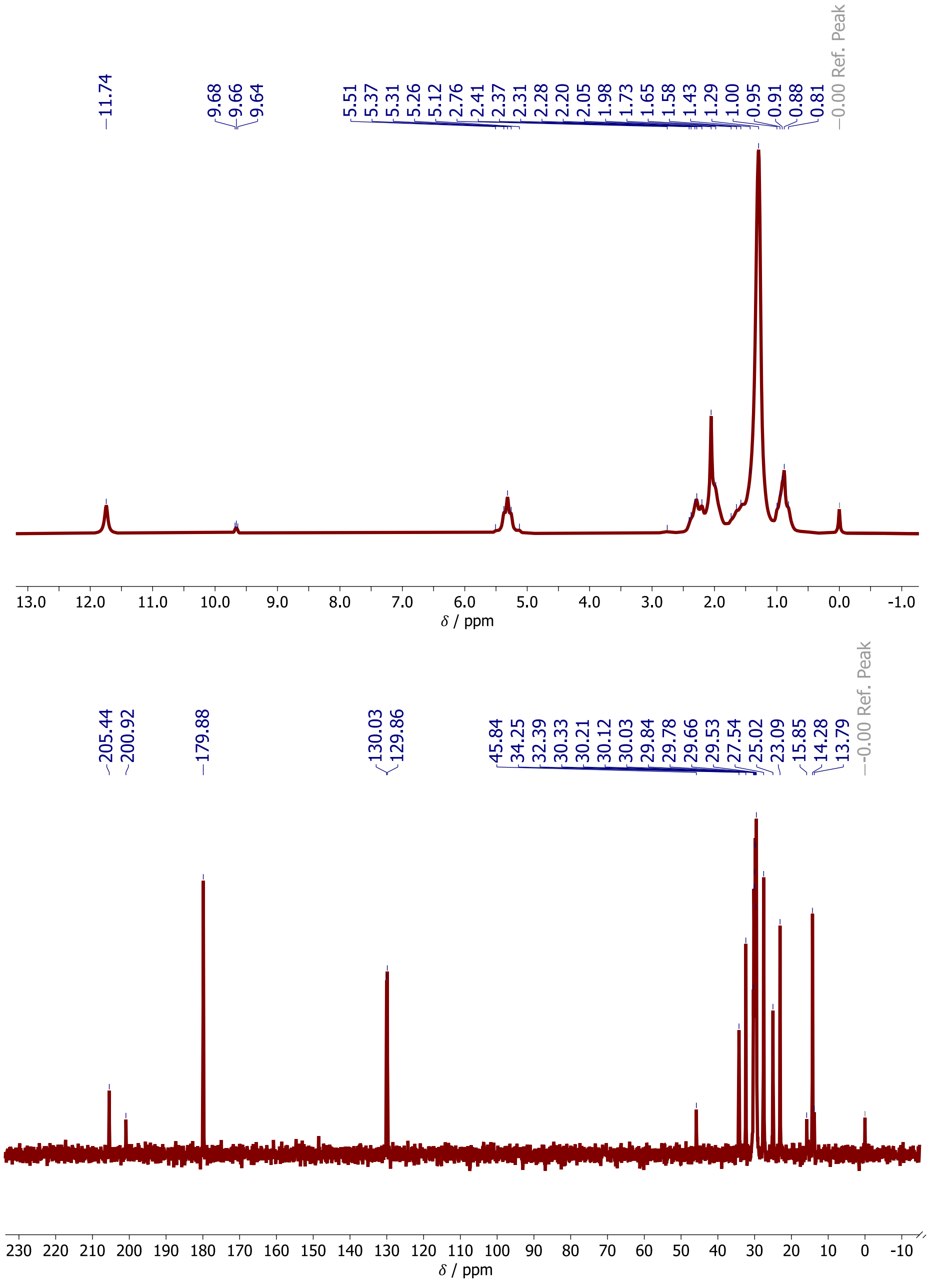


Figure B.6: ^{1}H and ^{13}C NMR spectra of mixture IV (acetone, butanal, oleic acid), cf. Table 5. The peak at 0 ppm belongs to TMS, which was used as reference. Top: ^{1}H NMR spectrum. Bottom: ^{13}C NMR spectrum. The peak at 11.70 ppm in the ^{1}H NMR spectrum is classified as stemming from labile protons.

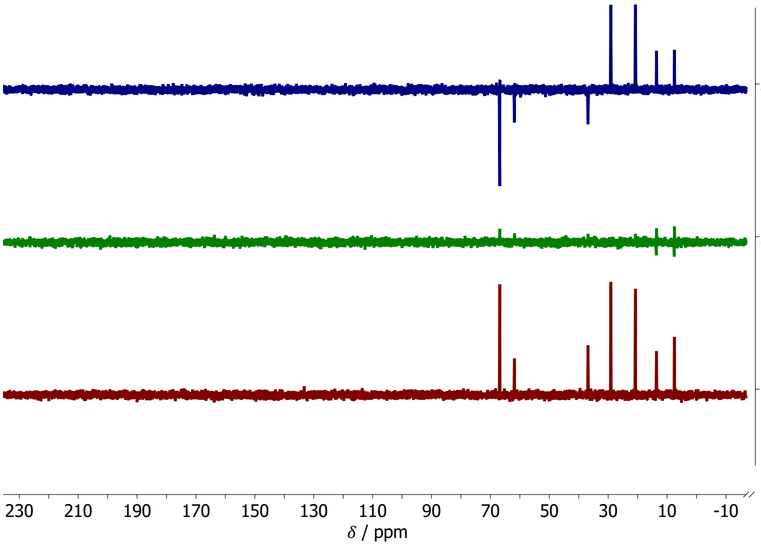


Figure B.7: ^{13}C DEPT NMR spectra of mixture I (2-butanone, ethyl acetate, water), cf. Table 5. The peak at 67.2 ppm in the ^{13}C belongs to 1,4-dioxane, which was used as reference. Top: DEPT spectrum with 135° phase angle. Middle: DEPT spectrum with 90° phase angle. Bottom: DEPT spectrum with 45° phase angle.

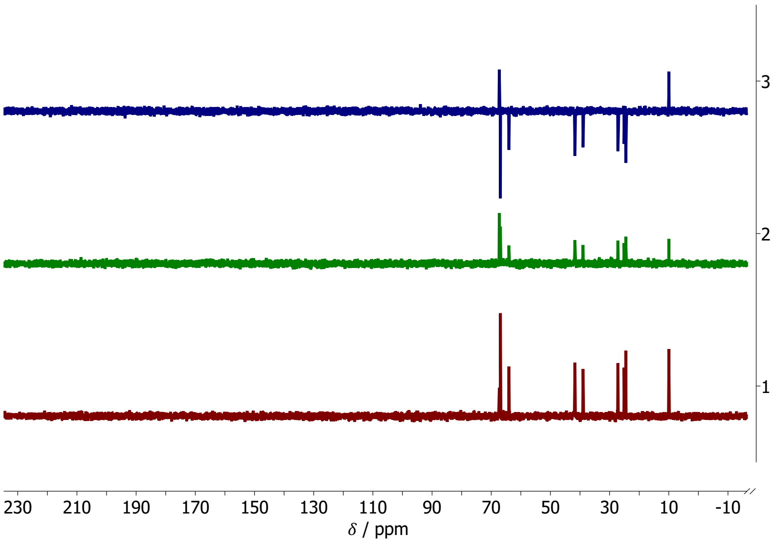


Figure B.8: ^{13}C DEPT NMR spectra of mixture II (cyclohexanone, malic acid, 1-propanol, water), cf. Table 5. The peak at 67.2 ppm in the ^{13}C belongs to 1,4-dioxane, which was used as reference. Top: DEPT spectrum with 135° phase angle. Middle: DEPT spectrum with 90° phase angle. Bottom: DEPT spectrum with 45° phase angle.

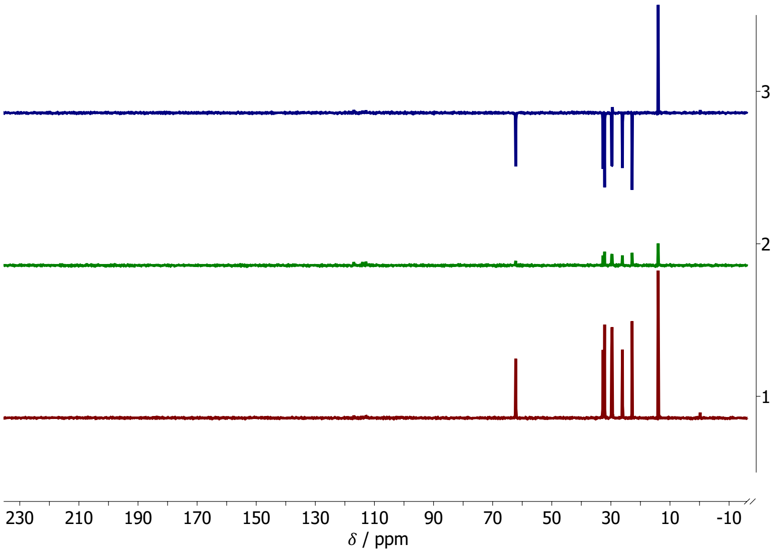


Figure B.9: ^{13}C DEPT NMR spectra of mixture III (1-octanol, tert-butylhydroquinone), cf. Table 5. The peak at 0 ppm in the ^{13}C belongs to TMS, which was used as reference. Top: DEPT spectrum with 135° phase angle. Middle: DEPT spectrum with 90° phase angle. Bottom: DEPT spectrum with 45° phase angle.

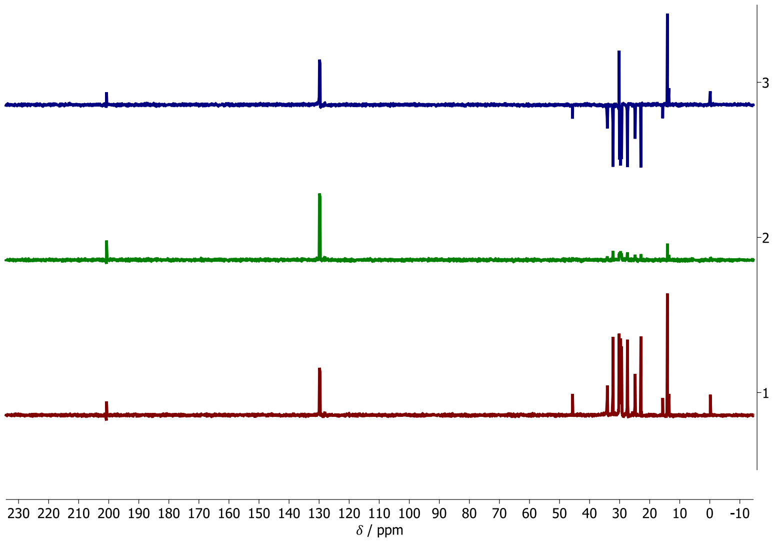


Figure B.10: ^{13}C DEPT NMR spectra of mixture IV (acetone, butanal, oleic acid), cf. Table 5. The peak at 0 ppm in the ^{13}C belongs to TMS, which was used as reference. Top: DEPT spectrum with 135° phase angle. Middle: DEPT spectrum with 90° phase angle. Bottom: DEPT spectrum with 45° phase angle.

B.6 Additional Results for Variant of Method without Information about Labile Protons

B.6.1 Prediction of Structural Groups from Pure-component Spectra

Figure B.11 shows results for the F_1 score for the prediction of structural groups based on pure-component spectra obtained with a variant of the NMR fingerprinting method that does not use prior information about the presence or absence of labile protons. For most group/section combinations, similar results as with the method presented in Chapter 3, which uses information on the presence or absence of labile protons, are obtained; slightly worse results are found for the COOH and CH_xOH groups, which could be expected as those groups contain labile protons.

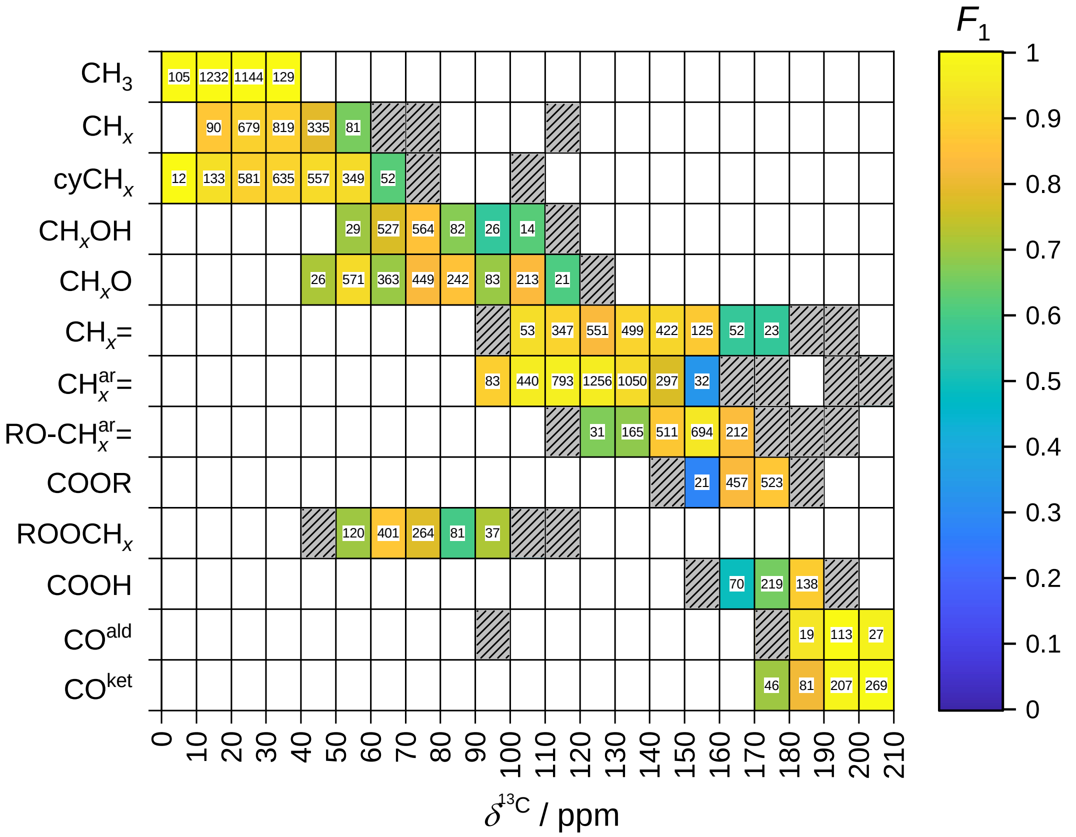


Figure B.11: F_1 test scores (indicated by the color code) of a variant of the NMR fingerprinting method, which does not use prior information about the presence or absence of labile protons, for structural groups g and sections s of the ^{13}C NMR spectra of the pure components in the data set. The numbers inside the cells indicate the number of components N_g^s in the data set that contain the respective structural group g (row) inducing a peak in the respective section s of the ^{13}C NMR spectrum (column). White cells indicate group/section combinations with $N_g^s = 0$, shaded cells with $N_g^s < 10$.

B.6.2 Prediction of Structural Groups from Mixture Spectra

In the following, results for the application of the variant that does not use knowledge about the presence/absence of labile protons on mixture spectra are presented. Consequently, peaks stemming from labile protons were *not* ignored in the input vector here. In Figure B.12, the results for mixture I are shown, which are identical to the results presented in Figure 11. Figure B.13 shows the respective results for mixture II, in contrast to the results shown in Figure 12, the COOH group was misinterpreted as a COOR group. In Figures B.14-B.15, the results for applying the variant to mixtures III and IV are shown, respectively. In contrast to the results for mixture III in Figure 13, here the CH$_x^{ar}$= and RO−CH$_x^{ar}$= group is confused with the CH$_x$= group in all cases, except for the peak at 137.27 ppm. This is presumably because peaks stemming from labile protons are located at 4.99, 5.04, and 5.10 ppm, which fall in the characteristic region of the CH$_x$= group, cf. Figure 10. In mixture IV the COOH group is misinterpreted as CH$_x$= group here, cf. Figure 14.

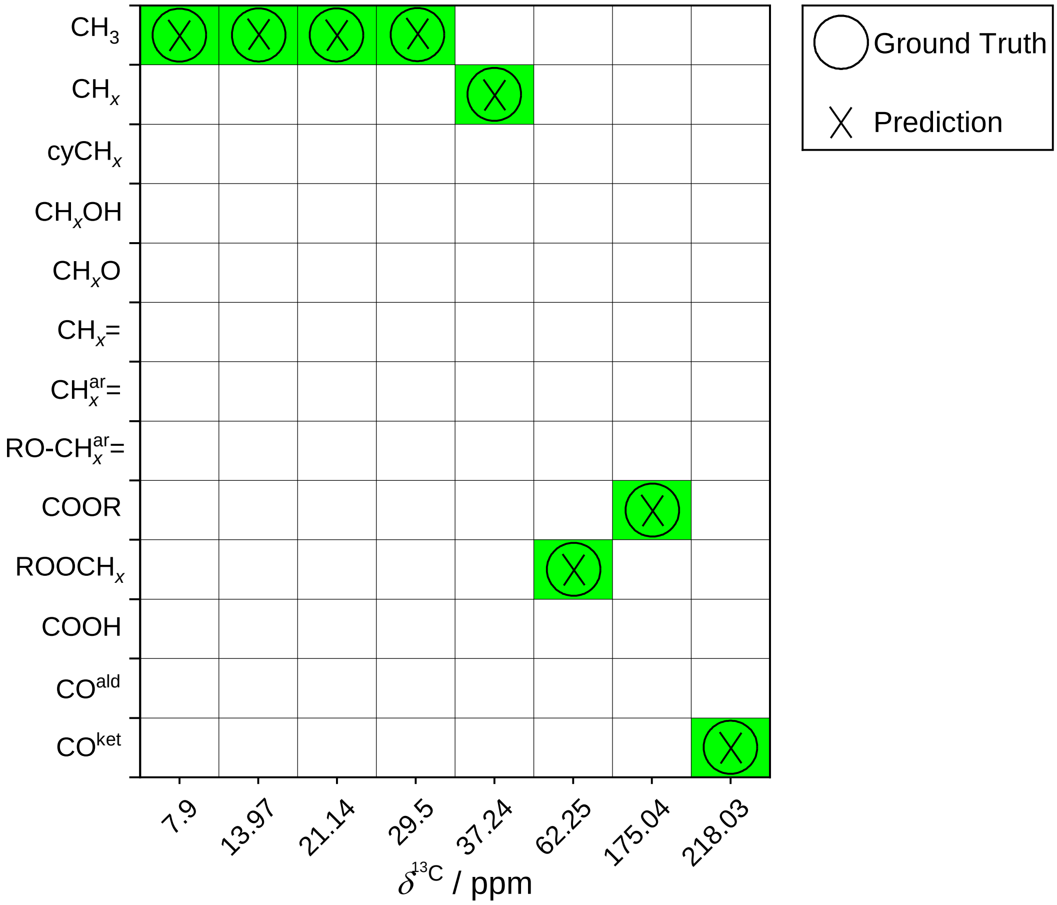


Figure B.12: Results of the application of a variant of the NMR fingerprinting method, which does not use prior information about the presence or absence of labile protons, for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum of mixture I (cf. Table 5). Green color indicates correct predictions. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

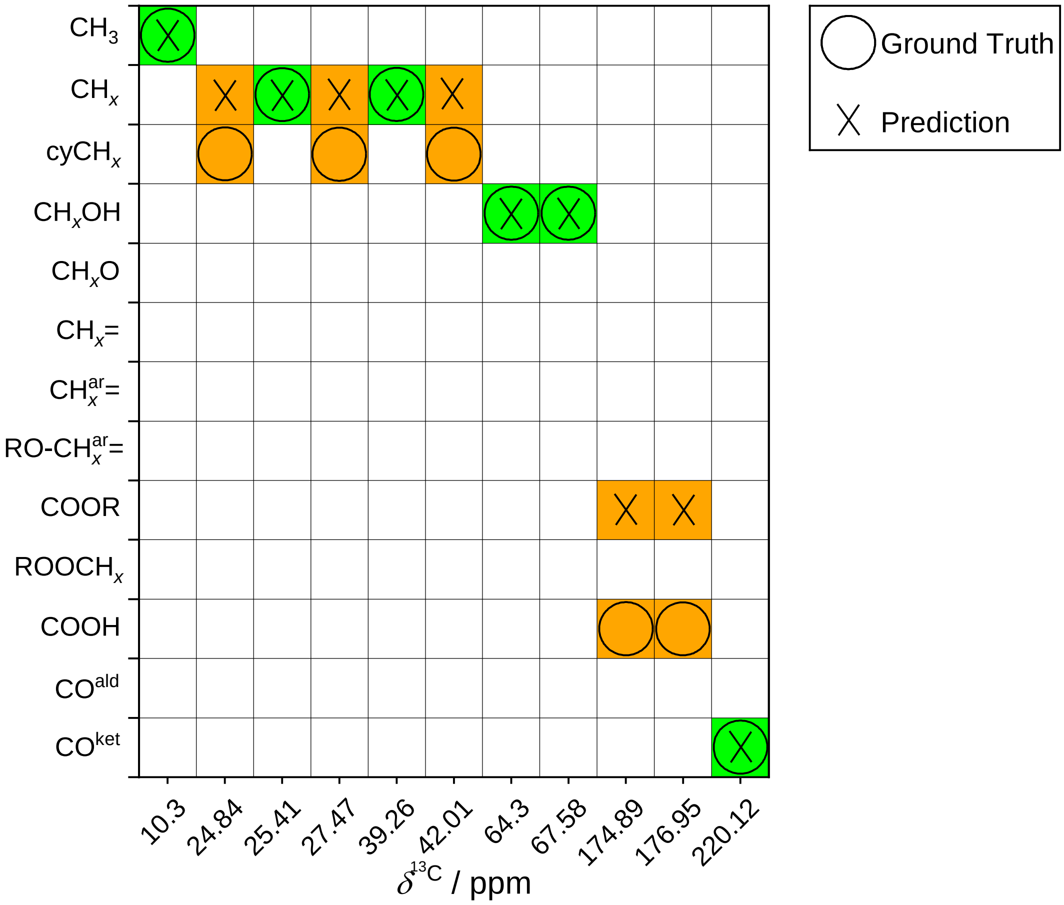


Figure B.13: Results of the application of a variant of the NMR fingerprinting method, which does not use prior information about the presence or absence of labile protons, for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum of mixture II (cf. Table 5). Green color indicates correct predictions and orange color indicates mistakes. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

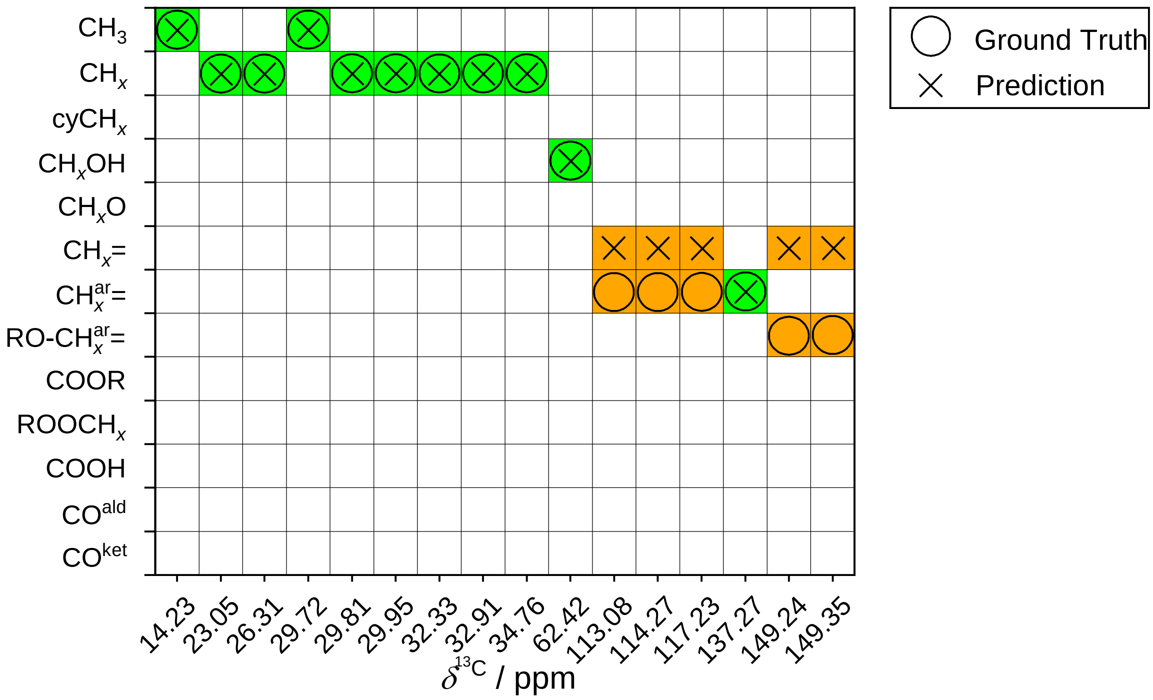


Figure B.14: Results of the application of a variant of the NMR fingerprinting method, which does not use prior information about the presence or absence of labile protons, for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum of mixture III (cf. Table 5). Green color indicates correct predictions and orange color indicates mistakes. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

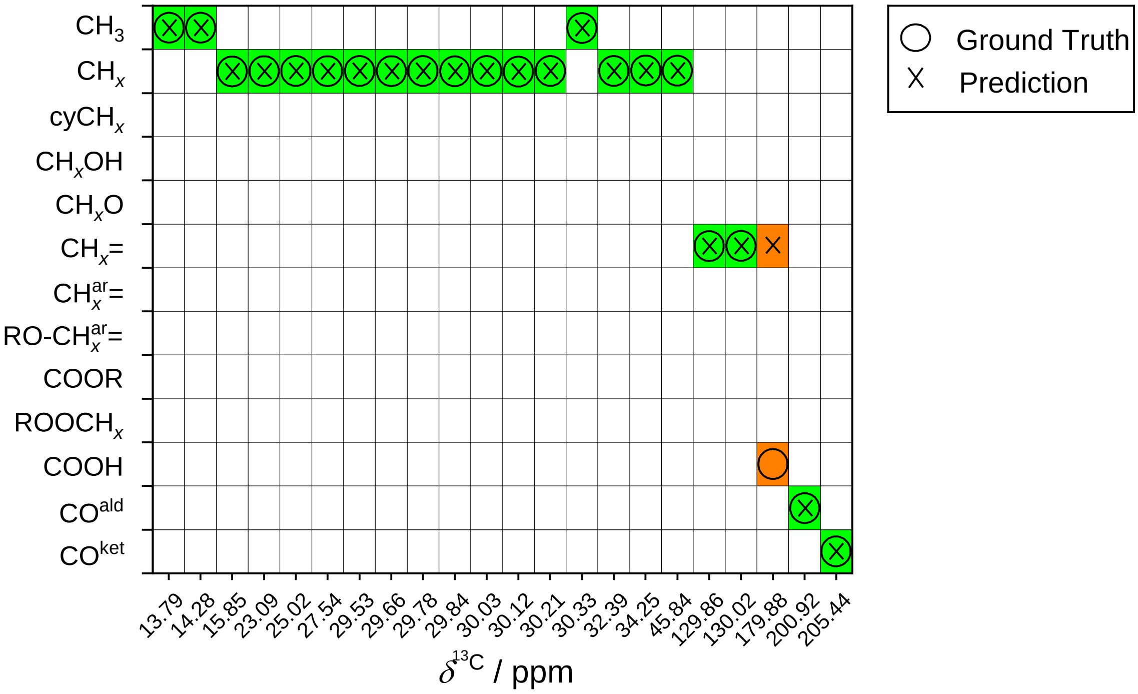


Figure B.15: Results of the application of a variant of the NMR fingerprinting method, which does not use prior information about the presence or absence of labile protons, for the prediction of structural groups and their assignment to peaks in the ^{13}C NMR spectrum of mixture IV (cf. Table 5). Green color indicates correct predictions and orange color indicates mistakes. On the x-axis, the positions of all peaks in the ^{13}C NMR spectrum of the mixture are indicated.

C Supporting Information for Chapter 4

C.1 Experimental Methods

C.1.1 Chemicals

Deionized and purified water, which was used as solvent for all test mixtures studied in Chapter 4, was produced with a purification system of Merck Millipore (Elix Essential 5). In Table C.1, information on the other chemicals used for the preparation of these mixtures is summarized.

Table C.1: Suppliers and purities of the chemicals used in Chapter 4. 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%

C.1.2 NMR Analysis

C.1.2.1 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 $^{1}J_{\mathrm{CH}}$ of 145 Hz that determines the specific delay in the DEPT experiment was chosen (for further details, see, e.g., Ref. [69]). 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%.

C.1.2.2 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 in Refs. [81, 82] was applied. In contrast to Refs. [81, 82], 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 [146]. 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 [147], cf. Eq. (9). 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. (9) were also evaluated by MestReNova.

C.2 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 [69].

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, 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 was used. 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, the ratio of the areas of the respective peaks in the DEPT 90 and the DEPT 135 spectrum was considered, which was well below unity in all cases of a primary carbon.

The peaks of the designated reference component were used 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.

C.3 PFG NMR Spectra and Assignment of Peaks

Figures C.1-C.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. It is noted 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. Therefore, an exponential line broadening of 1 Hz was first carried out, which is a standard processing step of NMR spectra. The great majority of peaks in the ^{13}C PFG NMR spectra recorded here did not show an overlapping with other peaks. Additionally, the peak heights were used to determine the self-diffusion coefficients in all cases to mitigate the effects of overlapping peaks on the evaluation of the diffusion coefficient.

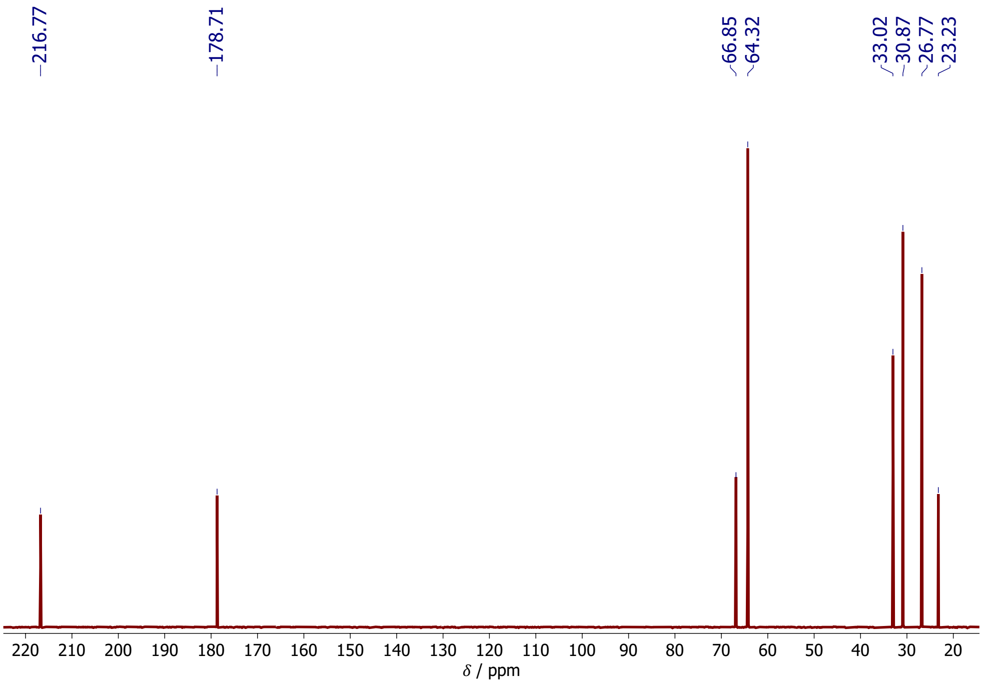


Figure C.1: ^{13}C PFG NMR spectrum of mixture I with gradient strength G = 2.55 G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

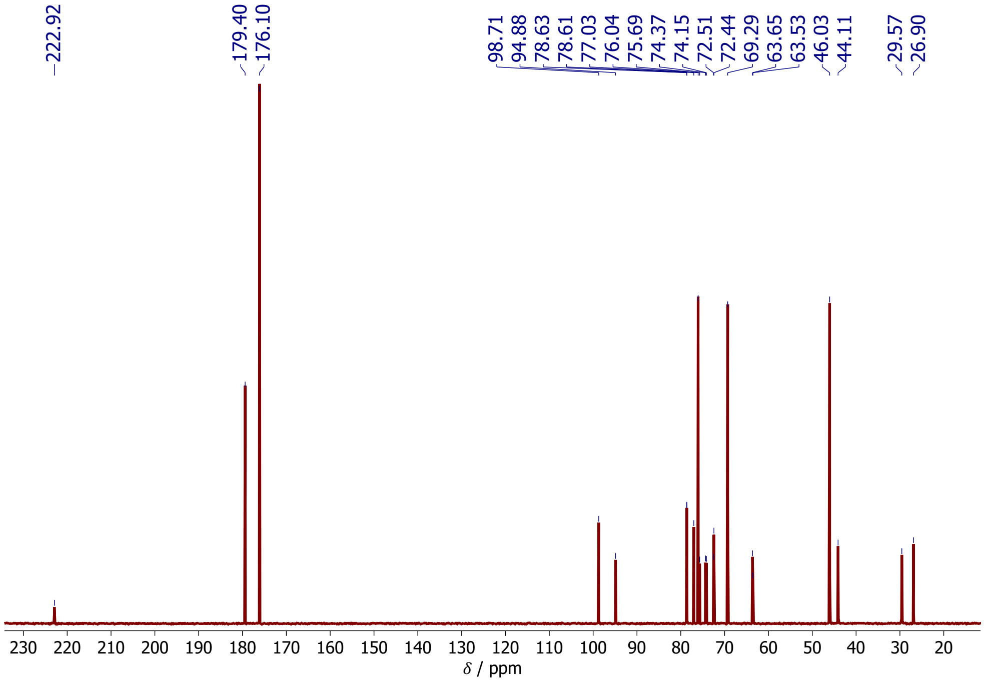


Figure C.2: ^{13}C PFG NMR spectrum of mixture II with gradient strength $G = 2.55$ G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

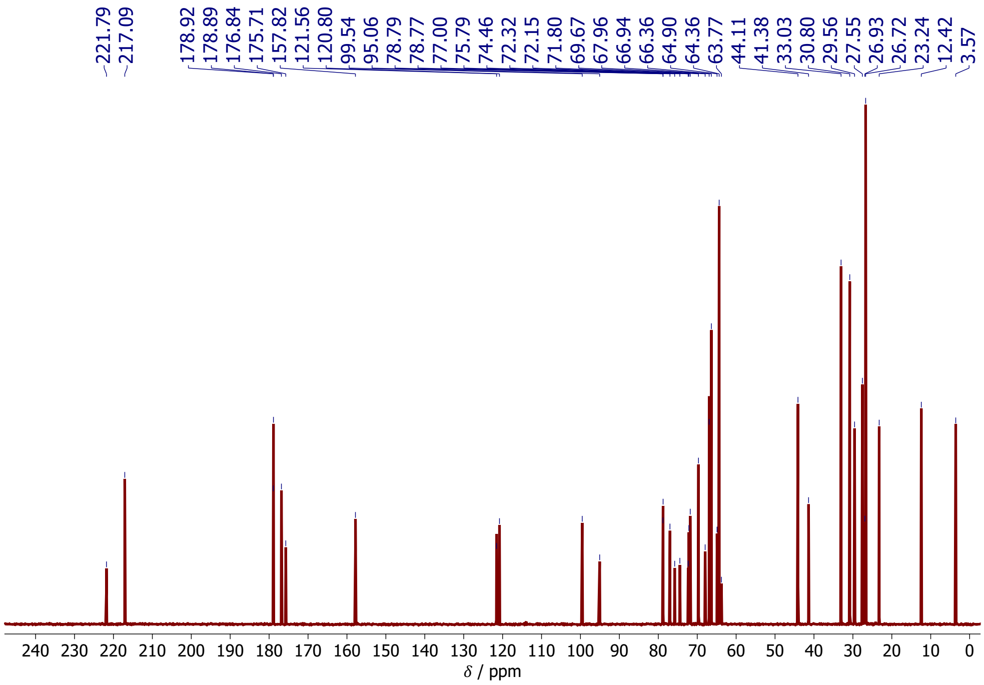


Figure C.3: ^{13}C PFG NMR spectrum of mixture III with gradient strength $G = 2.55$ G cm^{-1}. All distinguished peaks are indicated by their respective chemical shifts.

C.4 K-medians Algorithm and Silhouette Score

For the results presented here and in Chapter 4 it was proposed to use K-medians clustering, which is a variant of the K-means algorithm that is more robust towards outliers [106, 107]. 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:

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

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 Chapter 4, 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 $r_{p,k} = 0$. $\left\| \boldsymbol{x}_p - \boldsymbol{c}_k \right\|_1$ 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 2021 b [104] 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. (C.1), was kept [47, 106] 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, the so-called silhouette score s was used, which is a common metric for automatically selecting the most suitable number of clusters for a given clustering problem [91]. 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:

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

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 $\overline{s}(K)$, 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 $\overline{s}(K)$.

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 $\overline{s}(K)$ was calculated; then, the number of clusters K with the highest $\overline{s}(K)$ was adopted.

C.4.1 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, however, was adapted 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 (here: pseudo-components). To circumvent this issue, the experimental uncertainty $e_{p,95\%}$ of $\boldsymbol{x}_p^*$ for $a(\boldsymbol{x}_p^*)$, i.e., the intra-cluster distance was used, if the respective cluster contains $\boldsymbol{x}_p^*$ as the only data point. Furthermore $b(\boldsymbol{x}_p^*)$ is defined 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 $a(\boldsymbol{x}_p^*)$, but with a large distance to all other data points, i.e., large $b(\boldsymbol{x}_p^*)$, 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 $s(\boldsymbol{x}_p^*)$. On the other hand, a data point with a rather large error bar, i.e., large $a(\boldsymbol{x}_p^*)$, that is close to any other data point, i.e., low $b(\boldsymbol{x}_p^*)$, is not so likely to represent a separate cluster, which should, in this case, result in a small or even negative silhouette score $s(\boldsymbol{x}_p^*)$.

In Figure C.4 it is demonstrated that the default behavior of the MATLAB function "silhouette" would lead to the largest overall silhouette scores $\overline{s}(K)$ 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 C.4 demonstrates this using mixture I from Chapter 4 as an example, where the overall silhouette score $\overline{s}(K)$ continuously increases with increasing K.

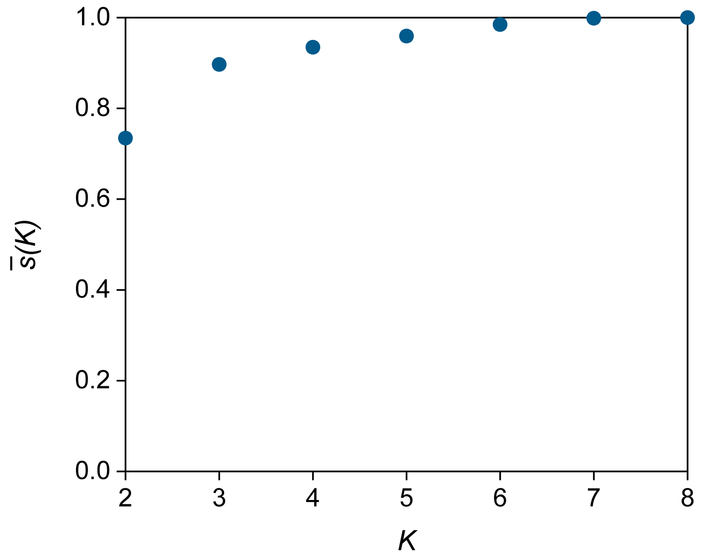


Figure C.4: Overall silhouette score $\overline{s}(K)$ for the clustering of peaks in the ^{13}C NMR spectrum of mixture I with the K-medians algorithm for different numbers of clusters K as calculated with the *default* MATLAB setting.

C.5 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. [148] gives a good overview. Therefore, only those concepts that are relevant to the development of the method are briefly recapitulated 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 [148, 149]. Of course, this requires that the reference components, among other things, are ideally inert and sufficiently soluble in the studied solvent [149]; and it requires the addition of reference components to the mixture of interest. Therefore, in Ref. [108] 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 diffusion coefficients":

$$\log(D_{\mathrm{x,norm}}) = \log(D_{\mathrm{ref,fix}}) - \log(D_{\mathrm{ref}}) + \log(D_{\mathrm{x}}) \tag{C.3}$$

where $\log(D_{\mathrm{x,norm}})$ is the normalized self-diffusion coefficient of the unknown component (labeled "x" here), D_{ref} and D_{x} are the measured self-diffusion coefficients of the reference and unknown component in the sample, respectively, and $D_{\mathrm{ref,fix}}$ is the known value of the self-diffusion coefficient of the reference component that was determined by measuring only the reference component in the same solvent. It is noted that $D_{\mathrm{ref,fix}}$ only has to be determined once for each reference component in a specific solvent and then can be used for the determination of molar masses of unknown components.

Solvent-specific power-laws (for different shapes of unknown components) are then fitted to the normalized diffusion coefficients of a large number of components. In consequence, Eq. (C.3) can be seen as a method to link the measured self-diffusion coefficient of an unknown component (in the actual sample) to a hypothetical sample to which the power-law was fitted, which then enables a good prediction of molar masses without requiring several reference components.

In the following, it is shown that the concept of normalized diffusion coefficients is similar to Eq. (11), where it is assumed that the ratio of the self-diffusion coefficients of an unknown component to that of a reference component in a mixture is the same as their ratio at infinite dilution in the solvent. It is shown in the following that both approaches are directly linked to the concept of relative diffusion coefficients (cf. Eq. (10)).

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

$$\log\left(\frac{D_{\mathrm{x}}}{D_{\mathrm{ref}}}\right) = \log\left(\frac{D_{\mathrm{x,norm}}}{D_{\mathrm{ref,fix}}}\right) \tag{C.4}$$

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

$$\frac{D_{\mathrm{x}}}{D_{\mathrm{ref}}} = \frac{D_{\mathrm{x,norm}}}{D_{\mathrm{ref,fix}}} \tag{C.5}$$

In the following, it is assumed that the reference state is at infinite dilution and switch indices to the notation used in the thesis ($\mathrm{x} \rightarrow \tilde{\mathrm{U}}$):

$$\frac{D_{\tilde{\mathrm{U}}}}{D_{\mathrm{ref}}} = \frac{D^{\infty}_{\tilde{\mathrm{U}},\mathrm{norm}}}{D^{\infty}_{\mathrm{ref,fix}}} \tag{C.6}$$

The resulting Eq. (C.6) is equivalent to the concept of relative diffusion (Eq. (10)) at two different concentrations, or to Eq. (11). In contrast to Ref. [108], no solvent-specific power-law was fitted; instead, the SEGWE [77, 78] model was directly applied, which was developed for describing diffusion coefficients at infinite dilution. Furthermore, the SEGWE model has been demonstrated to perform reasonably well using just one universal fit parameter for different solvents [77, 78, 109].

C.6 Concentration Dependence of Relative Diffusion Coefficients

To verify the validity of Eq. (11), i.e., that the ratio of the diffusion coefficients of two components (a known reference component and a pseudo-component $\tilde{\mathrm{U}}$ here) is approximately constant for different compositions, two aqueous systems were studied here as examples. Table C.2 gives an overview of these systems and specifies 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 diffusion coefficient at infinite dilution in water at 298.15 K of $D^{\infty}_{\mathrm{ref}} = 0.99 \cdot 10^{-9}\mathrm{m}^2\mathrm{s}^{-1}$ was taken from Ref. [115] (as in Chapter 4). In system B, acetone was chosen as reference component, for which $D^{\infty}_{\mathrm{ref}} = 1.3 \cdot 10^{-9}\mathrm{m}^2\mathrm{s}^{-1}$ in water at 298.15 K was taken from Ref. [150]. Figure C.5 shows the ratio $\frac{D_{\tilde{\mathrm{U}}}}{D_{\mathrm{ref}}}$ for the two systems measured by PFG NMR (cf. Section C.1.2.2). For all components, the arithmetic mean of the self-diffusion coefficients of the respective peaks of the components were taken.

Table C.2: Overview of the studied aqueous mixtures for verifying Eq. (11). 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.

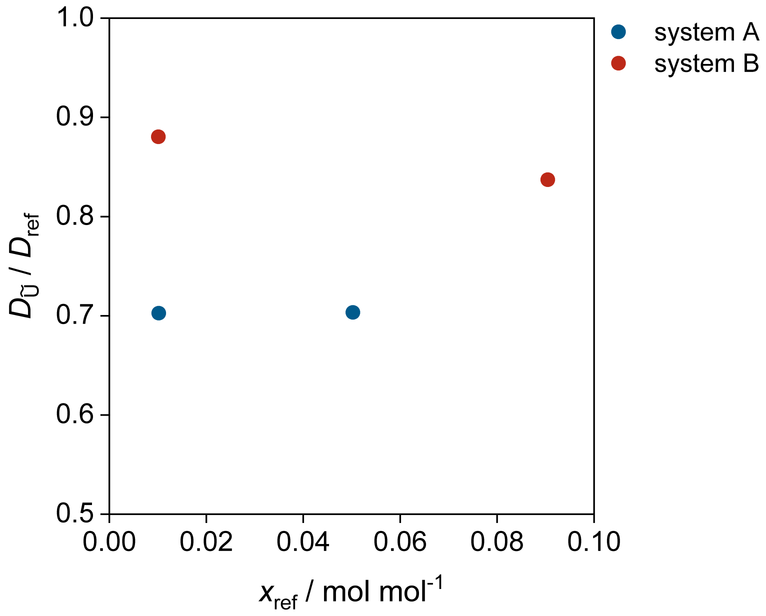


Figure C.5: Measured dependence of the ratio $\frac{D_{\tilde{\mathrm{U}}}}{D_{\mathrm{ref}}}$ on the mole fraction of the reference component, cf. Table C.2 and Eq. (11).

C.7 Identification and Quantification of Structural Groups

Most of the considered structural groups, cf. Table 6, contain only one carbon nucleus that shows a peak in the respective region of the ^{13}C NMR spectrum, which is denoted 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 >180 ppm for each pseudo-component. Also note that if a 'CH3' group is detected in a pseudo-component, 'CH3CO' is chosen, otherwise 'CH2CO'.

C.8 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 ν_k in each pseudo-component:

$$\nu_k = \frac{M_k}{\sum_{g=1}^{G} x_{g,k} M_g} \tag{C.7}$$

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

Table C.3: Molar mass M_g of all considered groups in Chapter 4, cf. Table 6.

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:

$$\nu_{g,k} = x_{g,k}\nu_k \tag{C.8}$$

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:

$$z_k = \sum_{g=1}^{G} \nu_{g,k} z_g^* \tag{C.9}$$

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:

$$x_k^* = \frac{\frac{\sum_{g=1}^{G} A_{g,k}}{z_k}}{\sum_{k=1}^{K}\left(\frac{\sum_{g=1}^{G} A_{g,k}}{z_k}\right)} \tag{C.10}$$

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

C.9 Structural Group Composition

C.9.1 Composition of True Components

Table C.4 shows the composition of all components studied in Chapter 4 regarding the groups of original UNIFAC [31, 129]. Note that the UNIFAC nomenclature uses ’THF,’ [129] as an abbreviation for cyclic ether groups. Since it is found misleading, ’cy-CH2O’ is used as abbreviation instead.

Table C.4: Components considered in Chapter 4 and their composition regarding groups from the UNIFAC table [31, 129]. 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)

C.9.2 Predicted Molar Masses and Composition of Pseudo-components

Tables C.5-C.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 7, 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 ($\tilde{\mathrm{U}}_1$), which was needed for the determination of the stoichiometry of the other pseudo-components, is not included.

Table C.5: Absolute numbers $\nu_{g,k}$ of structural groups g according to UNIFAC [31, 129] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of defined pseudo-components for test mixture I, cf. Table 7.

	$\tilde{U}_2$	$\tilde{U}_3$	$\tilde{U}_4$
M_k	48.98	145.21	72.11
$\nu_{\mathrm{CH3},k}$	0.782	-	1.147
$\nu_{\mathrm{CH2},k}$	-	6.464	-
$\nu_{\mathrm{CH},k}$	-	-	-
$\nu_{\mathrm{C},k}$	-	-	-
$\nu_{\mathrm{OH},k}$	-	3.206	-
$\nu_{\mathrm{CH=CH},k}$	-	-	-
$\nu_{\mathrm{C=C},k}$	-	-	-
$\nu_{\mathrm{COOH},k}$	-	-	1.219
$\nu_{\mathrm{CHO},k}$	-	-	-
$\nu_{\mathrm{CH3CO},k}$	0.865	-	-
$\nu_{\mathrm{CH2CO},k}$	-	-	-

Table C.6: Absolute numbers $\nu_{g,k}$ of structural groups g according to UNIFAC [31, 129] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of defined pseudo-components for test mixture II, cf. Table 7.

	$\tilde{U}_2$	$\tilde{U}_3$
M_k	94.47	212.08
$\nu_{\mathrm{CH3},k}$	-	-
$\nu_{\mathrm{CH2},k}$	3.879	1.898
$\nu_{\mathrm{CH},k}$	-	1.680
$\nu_{\mathrm{C},k}$	-	0.759
$\nu_{\mathrm{OH},k}$	-	2.846
$\nu_{\mathrm{CH=CH},k}$	-	0.203
$\nu_{\mathrm{C=C},k}$	-	-
$\nu_{\mathrm{COOH},k}$	-	2.238
$\nu_{\mathrm{CHO},k}$	-	-
$\nu_{\mathrm{CH3CO},k}$	-	-
$\nu_{\mathrm{CH2CO},k}$	0.953	-

Table C.7: Absolute numbers $\nu_{g,k}$ of structural groups g according to UNIFAC [31, 129] in pseudo-components k and predicted molar masses M_k (g mol^{-1}) of defined pseudo-components for test mixture III, cf. Table 7.

	$\tilde{\mathrm{U}}_2$	$\tilde{\mathrm{U}}_3$	$\tilde{\mathrm{U}}_4$	$\tilde{\mathrm{U}}_5$	$\tilde{\mathrm{U}}_6$	$\tilde{\mathrm{U}}_7$	$\tilde{\mathrm{U}}_8$
M_k	51.71	71.06	85.26	115.96	148.90	248.36	313.85
$\nu_{\mathrm{CH3},k}$	0.853	1.111	2.111	-	-	-	-
$\nu_{\mathrm{CH2},k}$	-	-	1.409	4.688	6.625	1.812	1.617
$\nu_{\mathrm{CH},k}$	-	-	0.712	-	-	3.681	3.135
$\nu_{\mathrm{C},k}$	-	-	-	-	-	-	-
$\nu_{\mathrm{OH},k}$	-	-	1.440	-	3.291	4.577	4.752
$\nu_{\mathrm{CH=CH},k}$	-	-	-	-	-	0.435	-
$\nu_{\mathrm{C=C},k}$	-	-	-	-	-	-	0.858
$\nu_{\mathrm{COOH},k}$	-	1.208	-	-	-	1.907	3.308
$\nu_{\mathrm{CHO},k}$	-	-	-	-	-	-	-
$\nu_{\mathrm{CH3CO},k}$	0.903	-	-	-	-	-	-
$\nu_{\mathrm{CH2CO},k}$	-	-	-	1.194	-	-	-

C.9.3 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:

1. 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.

2. 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 C.1.2.1, and, thus, of only minor influence on the results of Chapter 4.

3. 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 16, 18, and 20.

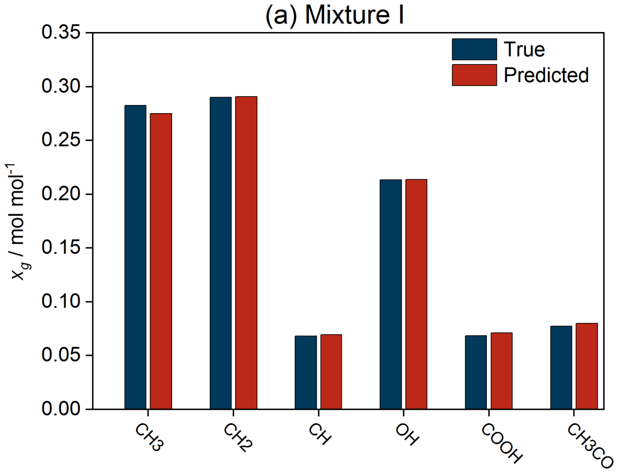
4. Errors introduced by the SEGWE model
   Errors introduced by the SEGWE model have a direct influence on the molar masses predicted from the measured diffusion coefficients. In the original paper [78], the authors reported a root-mean-square deviation in the order of 15% for predicted diffusion coefficients. The rather large expected errors comply with the observations in Figures 22 – 24. It can be assumed, that the errors introduced by the SEGWE model are the main source of error for the results.

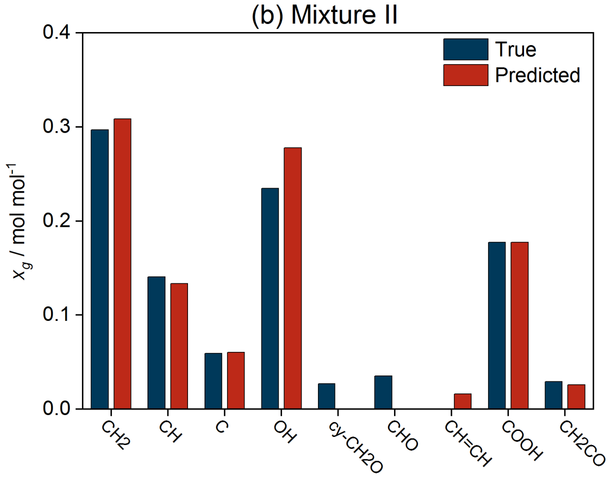
5. Experimental error of the diffusion coefficient of the defined reference component
   For the application of the proposed method, also the diffusion coefficient of a known reference component at infinite dilution in the solvent of the poorly specified mixture is required. The respective experimental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

C.10 Additional Results

C.10.1 NMR Fingerprinting

Figure C.6 shows the results of the NMR fingerprinting in the form of group mole fractions x_g. In mixture I (Figure C.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 C.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 the method. In mixture III (Figure C.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.


(a) Mixture I


(b) Mixture II

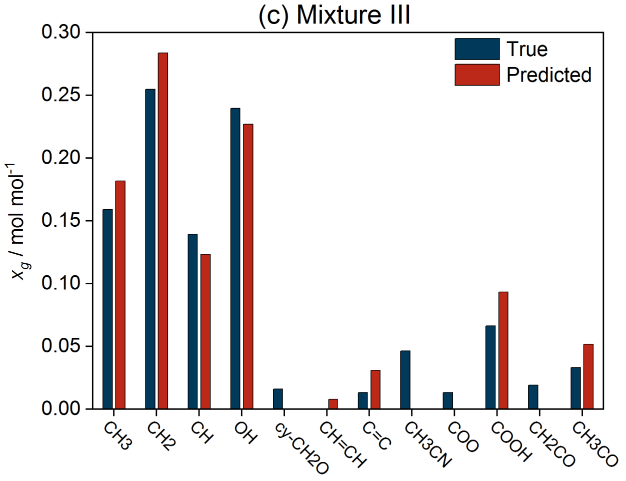

(c) Mixture III

Figure C.6: Prediction of structural groups in test mixtures, cf. Table 7.

C.10.2 Clustering with Prior Information

Figure C.7 shows the results of the clustering with the K-medians algorithm for mixture II from Chapter 4, 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.

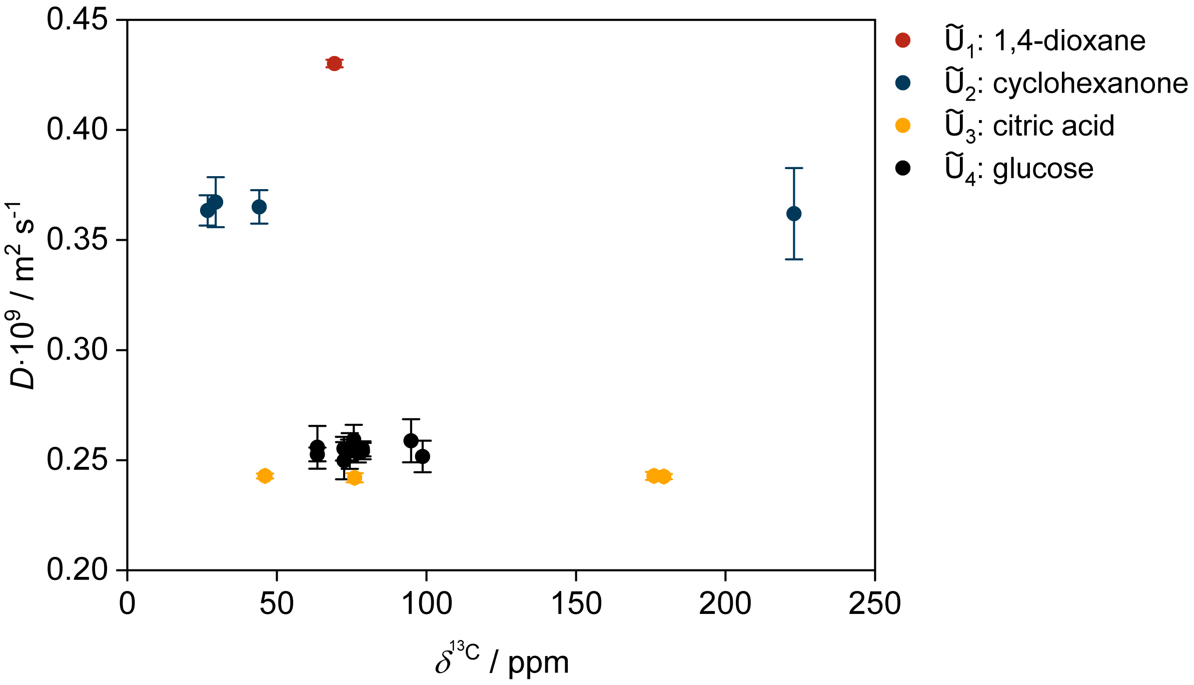


Figure C.7: DOSY map of mixture II with the result of the clustering of peaks (structural groups) by the K-medians algorithm and setting the number of clusters to $K = 4$. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars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

In Figure C.8 results of the clustering with the K-medians algorithm for mixture III from Chapter 4 are shown but here by fixing $K = 10$, which is the true number of components (except water and neglecting the anomers of xylose) in the mixture. By using this prior knowledge, the clustering algorithm correctly assigns all peaks (structural) groups to the different pseudo-components.

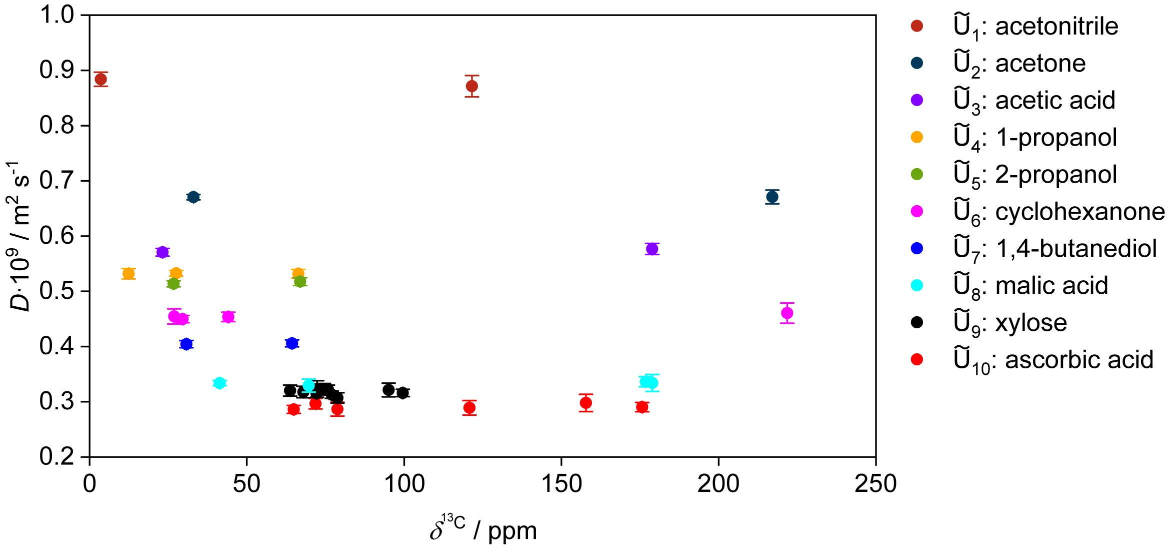


Figure C.8: DOSY map of mixture III with the result of the clustering of peaks (structural groups) by the K-medians algorithm and setting the number of clusters to $K = 10$. Different clusters are indicated by different colors and the respective true components are denoted in the legend. The error bars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

C.10.3 Influence of Reference Component on Predicted Molar Masses

In Figure C.9, the prediction of the molar masses of the pseudo-components defined by the K-medians algorithm in mixture III, cf. Figure 20, with the SEGWE model, is shown. In contrast to Figure 24, 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 $D^{\infty}_{\text{ref}} = 7.495 \cdot 10^{-10} \text{m}^2\text{s}^{-1}$ was adopted from Ref. [151] and used in Eq. (11).

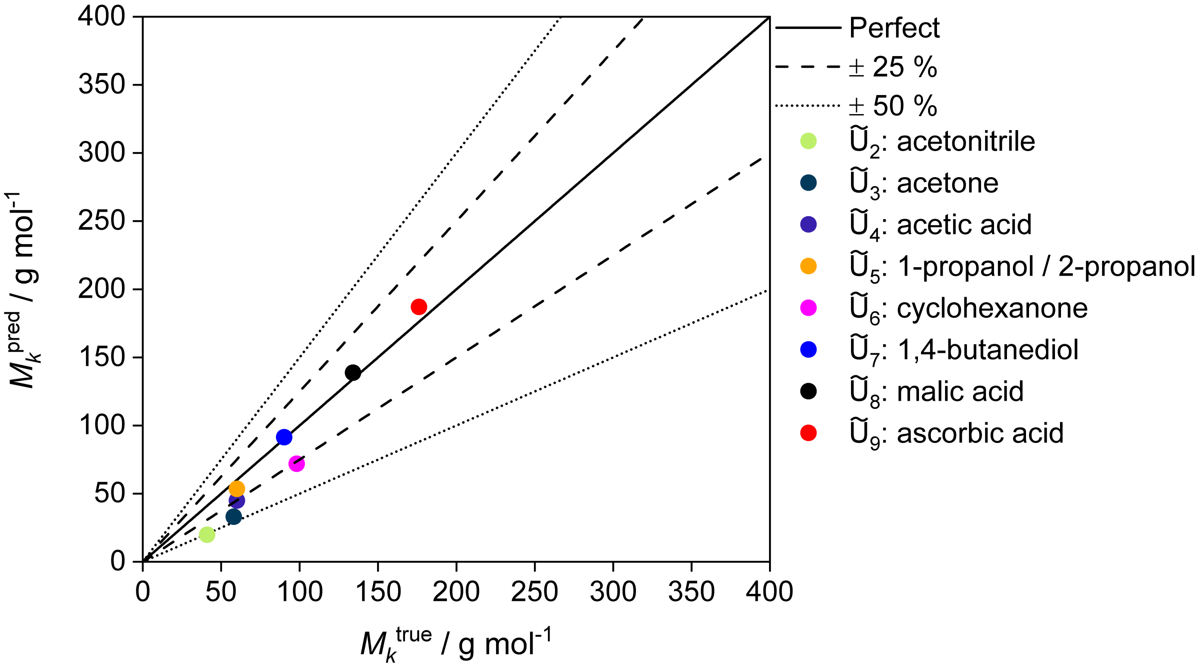


Figure C.9: Prediction of molar masses of pseudo-components in mixture III by considering xylose as reference component.

If the results shown here are compared to the results in Figure 24, an improved prediction of the molar masses of 1,4-butanediol, malic acid, and ascorbic acid in Figure C.9 can be observed. 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 24. These findings can be assigned to the fact that a constant ratio $\frac{D_{\tilde{\text{U}}}}{D_{\text{ref}}}$ for the extrapolation from finite concentrations to infinite dilution for all pseudo-components $\tilde{\text{U}}$ was assumed. The results indicate that a (slightly) different ratio for the different components could improve the results. It can be speculated 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 the method) in future work.

D Supporting Information for Chapter 5

D.1 Information on Water Mass Fraction

In the following, a variant of NEAT that is not only based on information on the mass fraction of the target component T in the studied poorly specified mixture, but also on the mass fraction of water W is described.

The total mass of the mixture is known in any case. If the mass fractions of the target component T and water W are assumed to be known, the mass of the pseudo-component Ũ can be calculated from the mass balance, denoted here as $m_{\tilde{\mathrm{U}}}^{\mathrm{MB}}$. As described in the main text, the NMR analysis yields an estimation of the total mass of the sum of all identified groups related to the unknown components, which are lumped to the pseudo-component Ũ. This mass is denoted as $m_{\tilde{\mathrm{U}}}^{\mathrm{NMR}}$ here. The difference $\Delta m = m_{\tilde{\mathrm{U}}}^{\mathrm{MB}} - m_{\tilde{\mathrm{U}}}^{\mathrm{NMR}}$ is expected to be positive in cases in which not all groups yield a signal in the NMR spectrum [60]. In this case, Δm was used to determine the number of additional OH(P) groups, which do not show a signal in ^{13}C NMR spectroscopy (cf. Tables D.1 and D.2). If Δm was negative, the mole numbers of the determined groups of the pseudo-component were reduced to fulfill the mass balance. The ratios of the mole numbers of the groups among each other were thereby kept constant. In Tables D.1 and D.2 the assignment of chemical groups to ^{13}C NMR chemical shift regions for known water mass fraction for versions of NEAT based on GC-COSMO-RS (OL) and UNIFAC (DO) is given, respectively.

Table D.1: Assignment of chemical groups from GC-COSMO-RS (OL) to ^{13}C NMR chemical shift regions used for the predictions with NEAT with known water mass fraction in this thesis. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

^{13}C NMR chemical shift region / ppm	chemical group name	GC-COSMO-RS (OL) label
0 - 30	methyl group	'CH3' (1)
30 - 60	methylene group	'CH2' (4)
60 - 90	alcohol group	'CH2' (7)[a]+'OH(P)' (35)
90 - 150	alkenyl group	'CH=CH' (58)
150 - 180	carboxyl group	'COOH' (44)
>180	carbonyl group	'CO' (51)
n.a.	hydroxyl group	'OH(P)' (35)

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.2: Assignment of chemical groups from the UNIFAC (DO) table to ^{13}C NMR chemical shift regions used for the predictions with NEAT with known water mass fraction in this thesis. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

^{13}C NMR chemical shift region / ppm	chemical group name	UNIFAC (DO) label
0 - 30	methyl group	'CH3' (1,1)
30 - 60	methylene group	'CH2' (2,1)
60 - 90	alcohol group	'CH2' (2,1)+'OH(P)' (14,5)
90 - 150	alkenyl group	'CH=CH' (6,2)
150 - 180	carboxyl group	'COOH' (42,20)
>180	carbonyl group	'CH2CO' (19,9)
n.a.	hydroxyl group	'OH(P)' (14,5)

In the following, the results of the above-described versions of NEAT for the systems considered in Chapter 5 are given. The presentation of the results is the same as in Chapter 5.

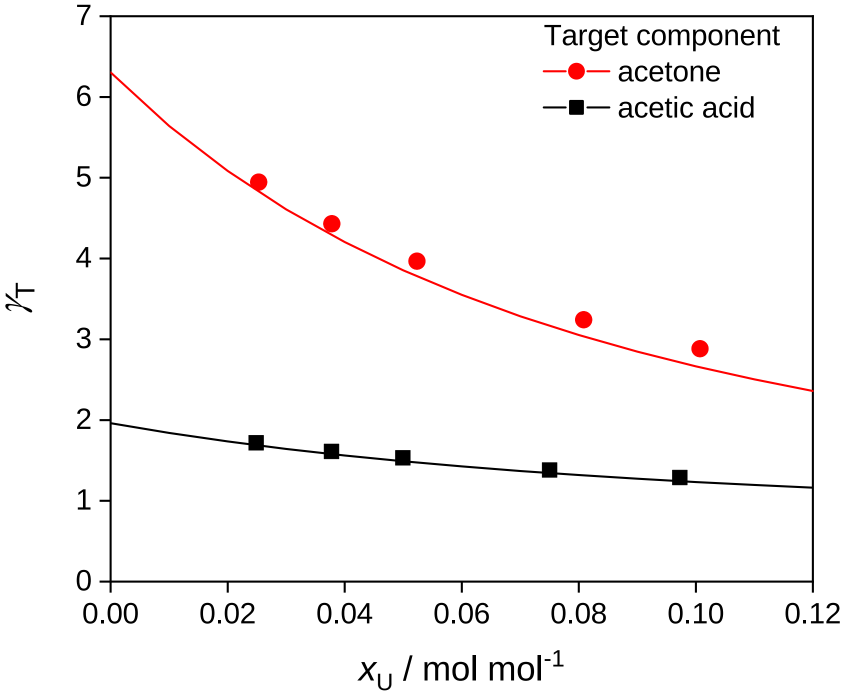


Figure D.1: Activity coefficient γ_T of target components (T = acetone or T = acetic acid) in ternary mixtures of system I and II, cf. Table 11, at 298 K. Lines: results from GC-COSMO-RS (OL) for the fully specified mixtures. Symbols: predictions with NEAT using information on the water mass fraction. No information on the unknown component U was used in NEAT.

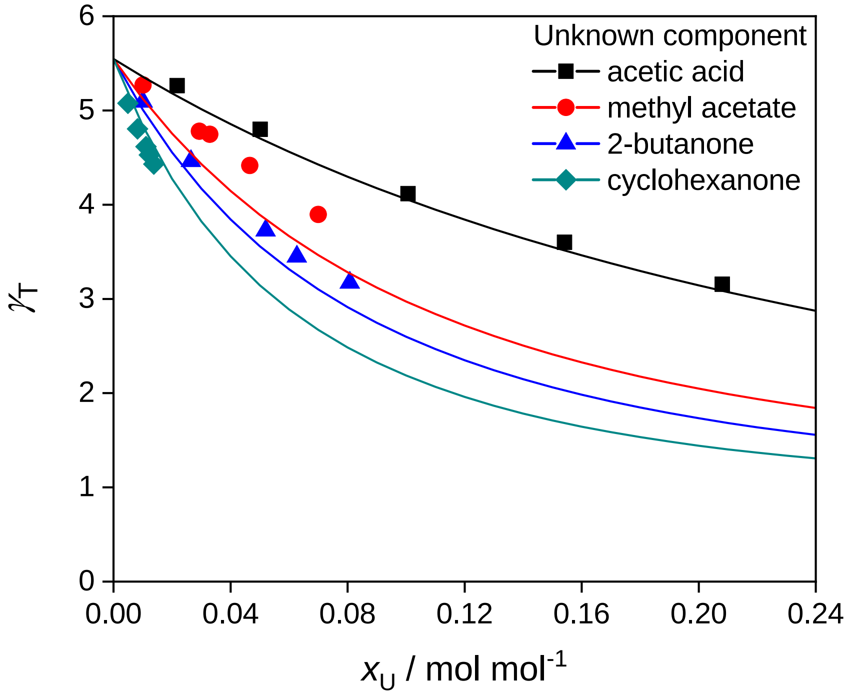


Figure D.2: Activity coefficient γ_T of target component T = ethanol in ternary mixtures of system III-VI, cf. Table 11, at 298 K. Lines: results from GC-COSMO-RS (OL) for the fully specified mixtures. Symbols: predictions with NEAT using information on the water mass fraction. No information on the unknown component U was used in NEAT.

For D-xylose, GC-COSMO-RS (OL) in its current version cannot be used since the required parameters are not available. Therefore, for obtaining the results of system VIII, quantum-chemical obtained σ-profiles, cavity surface areas A and the cavity volume V out of DDB [125] (cf. Section 5.2) were used for D-xylose for the fully specified mixture.

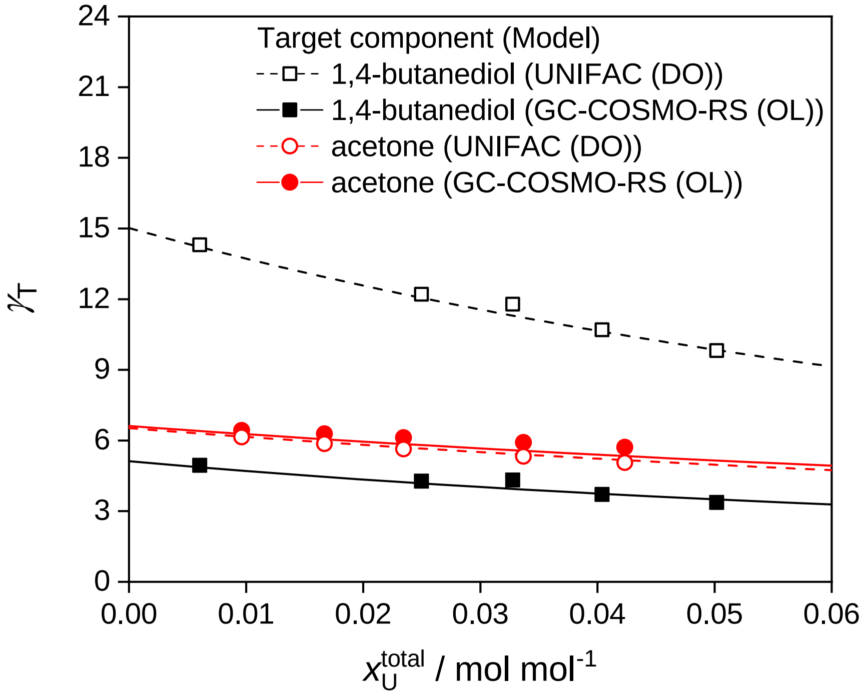


Figure D.3: Activity coefficient γ_T of target components (T = 1,4-butanediol or T = acetone) in five-component mixtures of system VII and VIII, cf. Table 11, at 298 K. Lines: results for fully specified mixtures. Symbols: predictions with NEAT using information on the water mass fraction. No information on the unknown components U was used in NEAT.

D.2 Group Assignment

In Table D.3, the group assignment for the components used in Chapter 5 according to UNIFAC (DO) and GC-COSMO-RS (OL) for the fully specified mixtures are given.

Table D.3: Assignment of UNIFAC (DO) and GC-COSMO-RS (OL) groups to the components of Chapter 5. For UNIFAC (DO), the numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118]. For GC-COSMO-RS (OL) the abbreviated group labels are introduced for clarity. For GC-COSMO-RS (OL) the numbers in parentheses correspond to the group identifiers in the original paper [124].

Component	UNIFAC (DO) groups	GC-COSMO-RS (OL) groups
acetic acid	1 x ’CH3’ (1,1), 1 x ’COOH’ (42,20)	1 x ’CH3’ (1), 1 x ’COOH’ (44)
acetone	1 x ’CH3’ (1,1), 1 x ’CH3CO’ (18,9)	2 x ’CH3’ (1), 1 x ’CO’ (51)
acetonitrile	1 x ’CH3CN’ (40,19)	1 x ’CN’ (57), 1x ’CH3’ (1)
1,4-butanediol	4 x ’CH2’ (2,1), 2 x ’OH(P)’ (14,5)	2 x ’CH2’ (4), 2 x ’CH2’ (7)[a], 2 x ’OH(P)’ (36), 2 x ’OH-OH’ (135)
2-butanone	1 x ’CH3’ (1,1), 1 x ’CH2’ (2,1), 1 x ’CH3CO’ (18,9)	2 x ’CH3 (1), 1 x ’CH2’ (4), 1 x ’CO’ (51)
cyclohexanone	1 x ’CH2CO’ (19,9), 4 x ’CY-CH2’ (78,42)	1 x ’CO’ (51), 5 x ’CY-CH2’ (9)
ethanol	1 x ’CH3’ (1,1), 1 x ’CH2’ (2,1), 1 x ’OH(P)’ (14,5)	1 x ’CH3’ (1), 1 x ’CH2’ (7)[a], 1 x ’OH(P)’ (36)
methyl acetate	1 x ’CH3COO’ (21,11), 1 x ’CH3’ (1,1)	1 x ’CH3’ (1), 1 x ’CH3’ (2)[a], 1 x ’COO’ (45)
2-propanol	2 x ’CH3’ (1,1), 1 x ’CH’ (3,1) , 1 x ’OH(S)’ (81,5)	2 x ’CH3’ (1), 1 x ’CH2’ (7)[a, b], 1 x ’OH(S)’ (34)
water	1 x ’H2O’ (16,7)	n.a.[c]
D-xylose	1 x ’THF’ (27,43), 4 x ’OH(S)’ (81,5), 3 x ’CY-CH’ (79,42)	n.a.[d]

[a] In GC-COSMO-RS (OL), a ’CH2’/’CH3’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’/’CH3’ group.

[b] In GC-COSMO-RS (OL), group 7 represents ’CH2’/’CH’/’C’ in a chain attached to F/Cl/O/N.

[c] No group assignment available for water. The quantum-chemical obtained σ-profiles, cavity surface areas A, and the cavity volume V from DDB were used.

[d] For D-xylose, only the σ-profiles, cavity surface areas A, and the cavity volume V for one anomeric form (β-D-xylopyranose) was available in the DDB and used here.

D.3 Influence of the Molar Mass $M_{\tilde{U}}$ of the Pseudo-Component $\tilde{U}$

Figure D.4 shows the influence of the molar mass $M_{\tilde{U}}$ of the pseudo-component $\tilde{U}$ on the activity coefficient γ_T in a ternary mixture with T = acetic acid and 2-propanol as unknown U in system II. If no unreasonably small values for $M_{\tilde{U}}$ ($M_{\tilde{U}} < 50$ g mol^{-1}) are chosen the influence on the predictions is negligible. This was found for all studied mixtures. As in Refs. [59, 60], $M_{\tilde{U}} = 150$ g mol^{-1} was used in Chapter 5.

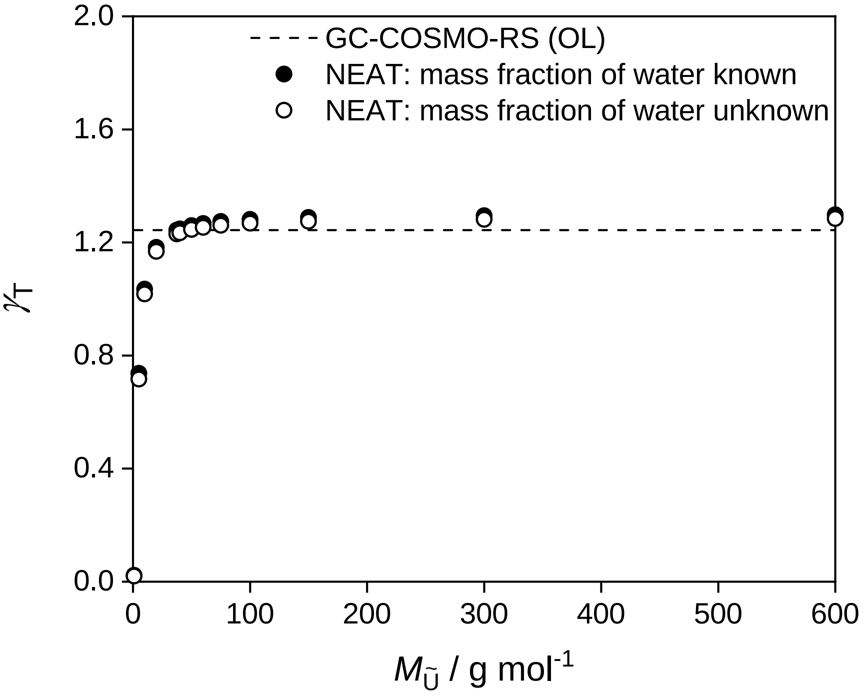


Figure D.4: Activity coefficient γ_T of target component T = acetic acid in a ternary mixture of system II. The composition of the mixture is x_T = 0.043 mol mol^{-1}, x_U = 0.097 mol mol^{-1}. Dashed line: results from GC-COSMO-RS (OL) for fully specified mixture. Symbols: predictions with NEAT for different assumed values $M_{\tilde{U}}$. No information on the unknown component U was used in the NEAT.

D.4 Stoichiometry of the Pseudo-Component $\tilde{\mathrm{U}}$ and Estimated Composition

In the following, the estimated composition of the poorly specified mixtures and the estimated stoichiometry of the pseudo-component $\tilde{\mathrm{U}}$ as calculated with NEAT for all studied mixtures are summarized. The section is divided according to the different versions of NEAT: first, the variant in which only information on the target component T was used, and second, the version in which information on target component T and the mass fraction of water W was used. The mass fraction of the target component T is always the same as for the fully specified mixture.

D.4.1 Unknown Water Mass Fraction

Table D.4: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system I, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 25. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g⁻¹		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.105	0.075	4.800	0.000	2.508	2.508	0.000	0.000	0.000
2	0.101	0.110	4.869	0.000	2.475	2.475	0.000	0.000	0.000
3	0.097	0.146	4.796	0.031	2.496	2.496	0.000	0.000	0.000
4	0.089	0.213	4.889	0.000	2.451	2.451	0.000	0.000	0.015
5	0.084	0.256	4.848	0.000	2.484	2.484	0.000	0.000	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.5: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system II, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 25. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.133	0.076	4.694	0.000	2.338	2.338	0.000	0.152	0.000
2	0.128	0.112	4.824	0.000	2.382	2.382	0.000	0.068	0.017
3	0.123	0.141	4.783	0.000	2.422	2.422	0.000	0.020	0.072
4	0.115	0.199	4.938	0.025	2.411	2.411	0.000	0.013	0.000
5	0.108	0.251	4.901	0.000	2.449	2.449	0.000	0.006	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.6: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system III, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 27. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.094	0.061	2.386	0.000	0.000	0.000	0.000	2.535	0.000
2	0.086	0.139	2.498	0.000	0.000	0.000	0.000	2.498	0.000
3	0.074	0.261	2.496	0.009	0.000	0.000	0.000	2.496	0.000
4	0.064	0.376	2.473	0.000	0.000	0.000	0.000	2.506	0.000
5	0.055	0.487	2.487	0.000	0.000	0.000	0.000	2.502	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.7: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system IV, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 27. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.097	0.034	1.934	2.118	0.000	0.000	0.000	2.026	0.000
2	0.090	0.094	2.057	2.150	0.062	0.062	0.000	1.932	0.000
3	0.088	0.100	2.120	2.263	0.000	0.000	0.000	1.919	0.000
4	0.084	0.142	2.096	2.174	0.019	0.019	0.000	1.941	0.000
5	0.078	0.208	2.101	2.126	0.000	0.000	0.000	1.968	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.8: Estimated compositions in the component space target component (T + pseudo-component Ũ + water W for the poorly specified mixtures of system V, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 27. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of Ũ						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.098	0.032	2.307	4.615	0.000	0.000	0.000	0.000	1.806
2	0.092	0.086	2.117	4.407	0.000	0.000	0.000	0.000	2.012
3	0.084	0.162	2.116	4.334	0.000	0.000	0.000	0.000	2.049
4	0.081	0.191	2.153	4.362	0.000	0.000	0.000	0.000	2.015
5	0.076	0.233	2.078	4.359	0.000	0.000	0.000	0.000	2.057

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.9: Estimated compositions in the component space target component (T + pseudo-component Ũ + water W for the poorly specified mixtures of system VI, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 27. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of Ũ						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.099	0.025	4.558	2.826	0.038	0.038	0.000	0.000	1.451
2	0.097	0.041	4.527	2.974	0.016	0.016	0.031	0.000	1.390
3	0.096	0.055	4.439	2.903	0.045	0.045	0.000	0.000	1.469
4	0.095	0.060	4.489	2.915	0.047	0.047	0.000	0.000	1.434
5	0.094	0.069	4.484	2.876	0.072	0.072	0.000	0.000	1.429

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.10: Estimated compositions in the component space target component T + pseudo-component Ũ + water W for the poorly specified mixtures of system VII, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 29. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(m)}$ / g g^{-1}		Stoichiometry of Ũ						
No.	$x_{\mathrm{T}}^{(m)}$	$x_{\tilde{\mathrm{U}}}^{(m)}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.096	0.013	4.046	2.023	0.000	0.000	0.646	0.658	0.512
2	0.090	0.058	3.876	2.011	0.000	0.000	0.689	0.725	0.461
3	0.088	0.066	3.922	1.939	0.000	0.000	0.683	0.749	0.441
4	0.086	0.097	3.879	1.963	0.000	0.000	0.652	0.764	0.456
5	0.083	0.124	3.840	1.920	0.000	0.000	0.660	0.757	0.502

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.11: Estimated compositions in the component space target component T + pseudo-component Ũ + water W for the poorly specified mixtures of system VII, cf. Table 11, obtained from NEAT based on UNIFAC (DO). The respective plot of the predictions with NEAT is given in Figure 29. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of Ũ					
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1,1)	'CH2' (2,1)	'OH(P)' (14,5)	'CH=CH' (6,2)	'COOH' (42,20)	'CH2CO' (19,9)
1	0.096	0.013	4.046	1.511	0.000	0.646	0.658	0.512
2	0.090	0.058	3.876	1.549	0.000	0.689	0.725	0.461
3	0.088	0.066	3.922	1.498	0.000	0.683	0.749	0.441
4	0.086	0.097	3.879	1.508	0.000	0.652	0.764	0.456
5	0.083	0.124	3.840	1.418	0.000	0.660	0.757	0.502

Table D.12: Estimated compositions in the component space target component T + pseudo-component Ũ + water W for the poorly specified mixtures of system VIII, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure 29. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of Ũ						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.096	0.032	1.721	0.626	1.251	1.251	0.235	1.564	0.000
2	0.093	0.057	1.523	0.761	1.523	1.523	0.169	1.438	0.000
3	0.091	0.077	1.655	0.736	1.410	1.410	0.184	1.471	0.000
4	0.087	0.109	1.569	0.743	1.322	1.322	0.165	1.569	0.000
5	0.084	0.135	1.622	0.746	1.362	1.362	0.162	1.525	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.13: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system VIII, cf. Table 11, obtained from NEAT based on UNIFAC (DO). The respective plot of the predictions with NEAT is given in Figure 29. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$					
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1,1)	'CH2' (2,1)	'OH(P)' (14,5)	'CH=CH' (6,2)	'COOH' (42,20)	'CH2CO' (19,9)
1	0.096	0.032	1.721	1.877	1.251	0.235	1.564	0.000
2	0.093	0.057	1.523	2.284	1.523	0.169	1.438	0.000
3	0.091	0.077	1.655	2.146	1.410	0.184	1.471	0.000
4	0.087	0.109	1.569	2.065	1.322	0.165	1.569	0.000
5	0.084	0.135	1.622	2.109	1.362	0.162	1.525	0.000

D.4.2 Known Water Mass Fraction

Table D.14: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system I, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.1. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.105	0.074	4.800	0.000	2.508	2.508	0.000	0.000	0.000
2	0.101	0.108	4.869	0.000	2.475	2.475	0.000	0.000	0.000
3	0.097	0.145	4.796	0.031	2.496	2.496	0.000	0.000	0.000
4	0.089	0.213	4.889	0.000	2.451	2.451	0.000	0.000	0.015
5	0.084	0.256	4.848	0.000	2.484	2.484	0.000	0.000	0.000

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.15: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system II, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.1. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.133	0.071	4.694	0.000	2.338	2.338	0.000	0.152	0.000
2	0.128	0.105	4.824	0.000	2.382	2.382	0.000	0.068	0.017
3	0.123	0.136	4.783	0.000	2.422	2.422	0.000	0.020	0.072
4	0.115	0.196	4.938	0.025	2.411	2.411	0.000	0.013	0.000
5	0.108	0.244	4.901	0.000	2.449	2.449	0.000	0.006	0.000

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.16: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system III, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.2. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.094	0.065	2.245	0.000	0.000	0.522	0.000	2.385	0.000
2	0.086	0.142	2.445	0.000	0.000	0.188	0.000	2.445	0.000
3	0.074	0.259	2.496	0.009	0.000	0.000	0.000	2.496	0.000
4	0.064	0.363	2.473	0.000	0.000	0.000	0.000	2.506	0.000
5	0.055	0.451	2.487	0.000	0.000	0.000	0.000	2.502	0.000

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.17: Estimated compositions in the component space target component (T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system IV, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.2. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.097	0.038	1.737	1.903	0.000	0.897	0.000	1.820	0.000
2	0.090	0.105	1.845	1.929	0.056	0.965	0.000	1.733	0.000
3	0.088	0.116	1.830	1.953	0.000	1.206	0.000	1.657	0.000
4	0.084	0.159	1.869	1.938	0.017	0.975	0.000	1.730	0.000
5	0.078	0.225	1.945	1.968	0.000	0.655	0.000	1.822	0.000

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.18: Estimated compositions in the component space target component (T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system V, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.2. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.098	0.037	2.003	4.005	0.000	1.165	0.000	0.000	1.567
2	0.092	0.093	1.970	4.102	0.000	0.609	0.000	0.000	1.873
3	0.084	0.171	2.002	4.099	0.000	0.478	0.000	0.000	1.938
4	0.081	0.201	2.049	4.150	0.000	0.427	0.000	0.000	1.918
5	0.076	0.248	1.949	4.088	0.000	0.548	0.000	0.000	1.929

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.19: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system VI, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.2. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.099	0.025	4.558	2.826	0.038	0.038	0.000	0.000	1.451
2	0.097	0.041	4.527	2.974	0.016	0.016	0.031	0.000	1.390
3	0.096	0.054	4.439	2.903	0.045	0.045	0.000	0.000	1.469
4	0.095	0.060	4.486	2.913	0.047	0.052	0.000	0.000	1.433
5	0.094	0.067	4.484	2.876	0.072	0.072	0.000	0.000	1.429

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.20: Estimated compositions in the component space target component T + pseudo-component $\tilde{\text{U}}$ + water W for the poorly specified mixtures of system VII, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.3. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\text{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1)	'CH2' (4)	'CH2' (7)[a]	'OH(P)' (35)	'CH=CH' (58)	'COOH' (44)	'CO' (51)
1	0.096	0.019	2.807	1.404	0.000	2.701	0.448	0.457	0.355
2	0.090	0.075	3.003	1.557	0.000	1.987	0.534	0.562	0.357
3	0.088	0.096	2.701	1.335	0.000	2.744	0.470	0.516	0.303
4	0.086	0.117	3.224	1.631	0.000	1.490	0.542	0.635	0.379
5	0.083	0.143	3.336	1.668	0.000	1.158	0.573	0.658	0.436

[a] In GC-COSMO-RS (OL), a 'CH2' group attached to F/Cl/O/N is distinguished from a custom 'CH2' group.

Table D.21: Estimated compositions in the component space target component T + pseudo-component Ũ + water W for the poorly specified mixtures of system VII, cf. Table 11, obtained from NEAT based on UNIFAC (DO). The respective plot of the predictions with NEAT is given in Figure D.3. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$					
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1,1)	’CH2’ (2,1)	’OH(P)’ (14,5)	’CH=CH’ (6,2)	’COOH’ (42,20)	’CH2CO’ (19,9)
1	0.096	0.019	2.807	1.048	2.701	0.448	0.457	0.355
2	0.090	0.075	3.003	1.200	1.987	0.534	0.562	0.357
3	0.088	0.096	2.701	1.032	2.744	0.470	0.516	0.303
4	0.086	0.117	3.224	1.253	1.490	0.542	0.635	0.379
5	0.083	0.143	3.336	1.232	1.158	0.573	0.658	0.436

Table D.22: Estimated compositions in the component space target component T + pseudo-component Ũ + water W for the poorly specified mixtures of system VIII, cf. Table 11, obtained from NEAT based on GC-COSMO-RS (OL). The respective plot of the predictions with NEAT is given in Figure D.3. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$						
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	’CH3’ (1)	’CH2’ (4)	’CH2’ (7)[a]	’OH(P)’ (35)	’CH=CH’ (58)	’COOH’ (44)	’CO’ (51)
1	0.096	0.038	1.423	0.517	1.035	2.560	0.194	1.294	0.000
2	0.093	0.069	1.260	0.630	1.260	2.783	0.140	1.190	0.000
3	0.091	0.092	1.384	0.615	1.179	2.626	0.154	1.230	0.000
4	0.087	0.128	1.344	0.637	1.132	2.397	0.142	1.344	0.000
5	0.084	0.157	1.391	0.640	1.168	2.425	0.139	1.307	0.000

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.23: Estimated compositions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W for the poorly specified mixtures of system VIII, cf. Table 11, obtained from NEAT based on UNIFAC (DO). The respective plot of the predictions with NEAT is given in Figure D.3. The numbers in parentheses are the identifiers for the sub-group and the corresponding main-group from the original papers [32, 118].

	$x_i^{(\mathrm{m})}$ / g g^{-1}		Stoichiometry of $\tilde{\mathrm{U}}$					
No.	$x_{\mathrm{T}}^{(\mathrm{m})}$	$x_{\tilde{\mathrm{U}}}^{(\mathrm{m})}$	'CH3' (1,1)	'CH2' (2,1)	'OH(P)' (14,5)	'CH=CH' (6,2)	'COOH' (42,20)	'CH2CO' (19,9)
1	0.096	0.038	1.423	1.552	2.560	0.194	1.294	0.000
2	0.093	0.069	1.260	1.890	2.783	0.140	1.190	0.000
3	0.091	0.092	1.384	1.794	2.626	0.154	1.230	0.000
4	0.087	0.128	1.344	1.769	2.397	0.142	1.344	0.000
5	0.084	0.157	1.391	1.808	2.425	0.139	1.307	0.000

D.5 Averaging of σ-profiles

Besides the cavity volume $V_{\tilde{\mathrm{U}}}$, two cavity surface areas $A_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}$ and $A_{\tilde{\mathrm{U}}}^{\mathrm{hb}}$ and two sigma-profiles $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$ are obtained for the pseudo-component Ũ from GC-COSMO-RS (OL). $A_{\tilde{\mathrm{U}}}$ is the sum of the two cavity surface areas, i.e. the total cavity surface area:

$$A_{\tilde{\mathrm{U}}} = A_{\tilde{\mathrm{U}}}^{\mathrm{nhb}} + A_{\tilde{\mathrm{U}}}^{\mathrm{hb}} \tag{D.1}$$

$p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum [123, 152] of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$:

$$p_{\tilde{\mathrm{U}}}(\sigma) = \frac{p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma) A_{\tilde{\mathrm{U}}}^{\mathrm{nhb}} + p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma) A_{\tilde{\mathrm{U}}}^{\mathrm{hb}}}{A_{\tilde{\mathrm{U}}}} \tag{D.2}$$

In analogy to the procedure described above, the respective numbers can be calculated for the unknown components U_i in the fully specified mixture:

$$A_{\mathrm{U}_i} = A_{\mathrm{U}_i}^{\mathrm{nhb}} + A_{\mathrm{U}_i}^{\mathrm{hb}} \tag{D.3}$$

$$p_{\mathrm{U}_i}(\sigma) = \frac{p_{\mathrm{U}_i}^{\mathrm{nhb}}(\sigma) A_{\mathrm{U}_i}^{\mathrm{nhb}} + p_{\mathrm{U}_i}^{\mathrm{hb}}(\sigma) A_{\mathrm{U}_i}^{\mathrm{hb}}}{A_{\mathrm{U}_i}} \tag{D.4}$$

For comparison, $p_{\tilde{\mathrm{U}}}(\sigma)$ for the fully specified mixture is calculated by summing $p_{\mathrm{U}_i}(\sigma)$ for all unknown components U_i in the mixture and weighting with the cavity surface area and the mole fraction of the respective components [123, 152]:

$$p_{\tilde{\mathrm{U}}}(\sigma) = \frac{\sum_i x_{\mathrm{U}_i} A_{\mathrm{U}_i} p_{\mathrm{U}_i}(\sigma)}{\sum_i x_{\mathrm{U}_i} A_{\mathrm{U}_i}} \tag{D.5}$$

where x_{U_i} is the mole fraction of U_i in a mixture that contains only the unknown components.

D.6 Additional σ-profiles

In Chapter 5, σ-profiles of the pseudo-components / unknown components in mixtures of several systems are shown. In the following, σ-profiles for the pseudo-components / unknown components in mixtures of all systems for which this is not the case are reported.

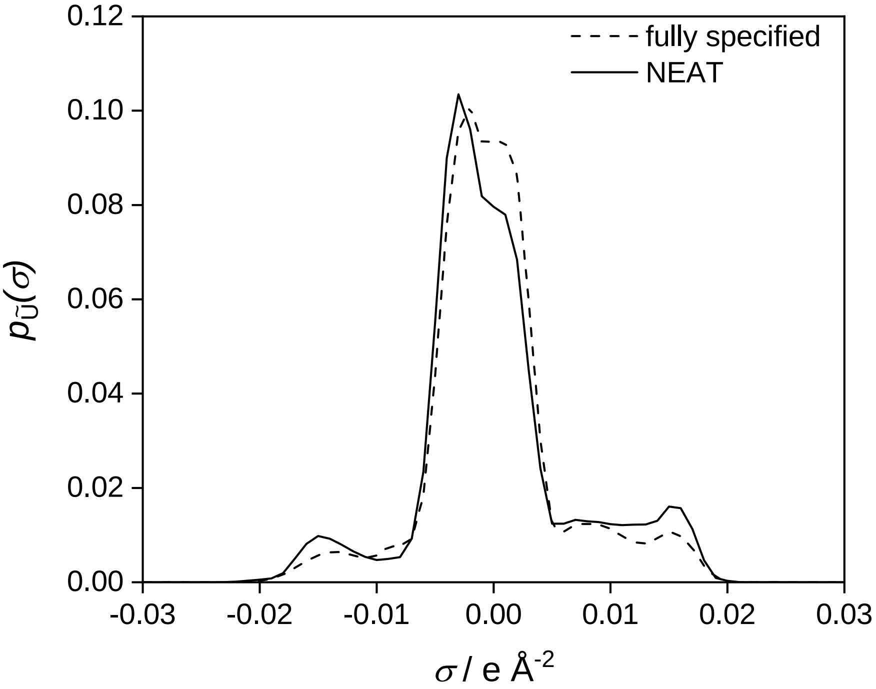


Figure D.5: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system II as predicted with NEAT and the σ-profile of U = 2-propanol. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Averaging of σ-profiles. Dashed line: results from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.046 mol mol^{-1}, x_{U} = 0.025 mol mol^{-1}.

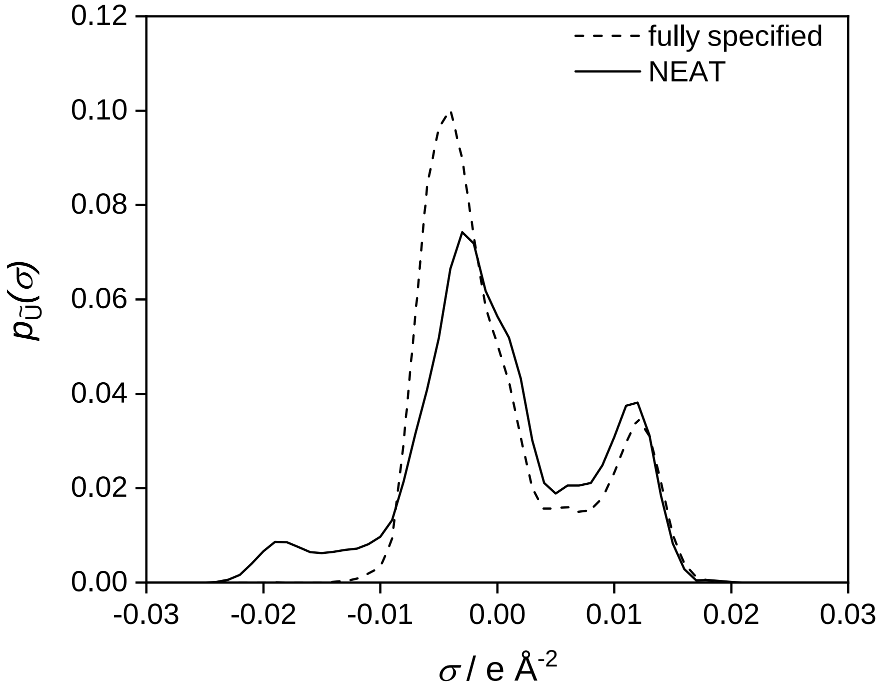


Figure D.6: σ-profile of the pseudo-component $\tilde{U}$ in a mixture of system IV as predicted with NEAT and the σ-profile of U = methyl acetate. $p_{\tilde{U}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{U}}^{\text{nhb}}(\sigma)$ and $p_{\tilde{U}}^{\text{hb}}(\sigma)$, cf. Averaging of σ-profiles. Dashed line: results from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_T = 0.041 mol mol^{-1}, x_U = 0.010 mol mol^{-1}.

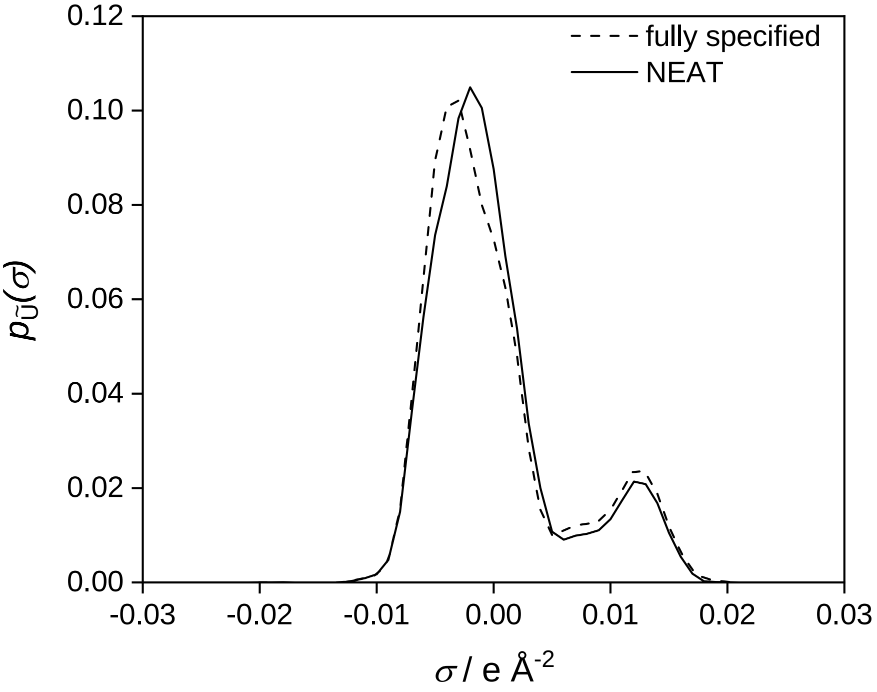


Figure D.7: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system V as predicted with NEAT and the σ-profile of U = 2-butanone. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Averaging of σ-profiles. Dashed line: results from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.042 mol mol^{-1}, x_{U} = 0.010 mol mol^{-1}.

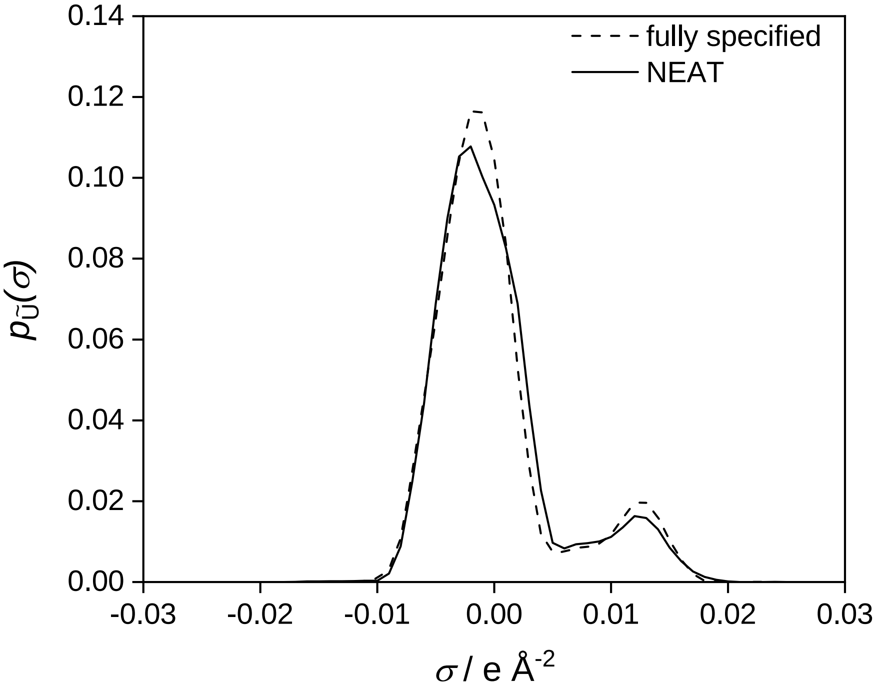


Figure D.8: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system VI as predicted with NEAT and the σ-profile of U = cyclohexanone. $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Averaging of σ-profiles. Dashed line: results from GC-COSMO-RS (OL) for U in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.042 mol mol^{-1}, x_{U} = 0.005 mol mol^{-1}.

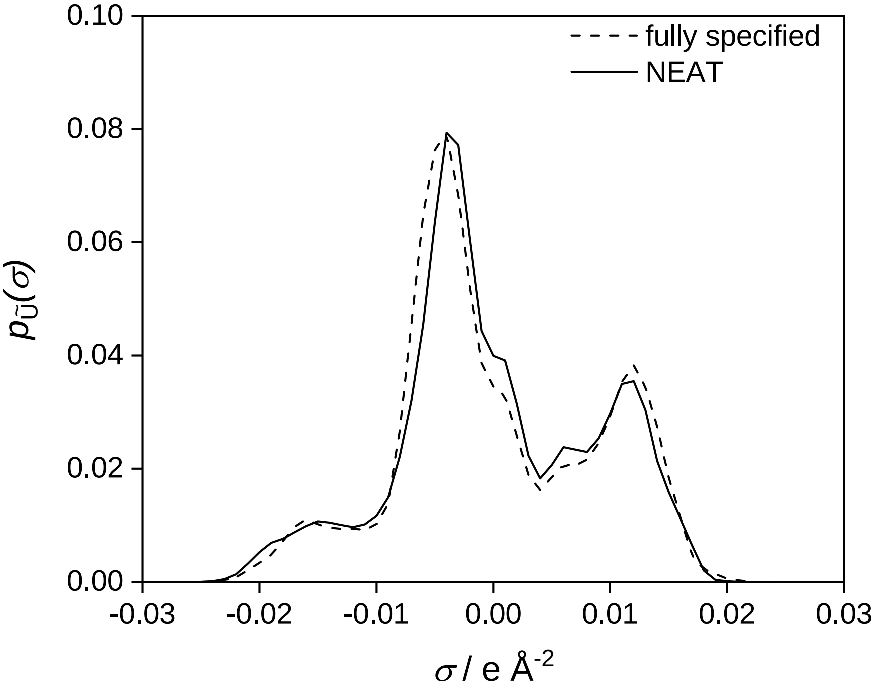


Figure D.9: σ-profile of the pseudo-component $\tilde{\mathrm{U}}$ in a mixture of system VIII as predicted with NEAT and the mixed σ-profile of the unknown components (U_1 = D-xylose, U_2 = acetic acid, U_3 = methyl acetate). $p_{\tilde{\mathrm{U}}}(\sigma)$ is the cavity surface area-weighted sum of $p_{\tilde{\mathrm{U}}}^{\mathrm{nhb}}(\sigma)$ and $p_{\tilde{\mathrm{U}}}^{\mathrm{hb}}(\sigma)$, cf. Averaging of σ-profiles. Dashed line: results from GC-COSMO-RS (OL) for the mixture of all U_i in the fully specified mixture. Solid line: prediction with NEAT. The composition of the mixture is x_{T} = 0.033 mol mol^{-1}, $x_{\mathrm{U}}^{\mathrm{total}}$ = 0.010 mol mol^{-1}.

D.7 Computational Details

The implementation of COSMO-RS (OL) in the DDB was used. In contrast to the work of Ref. [123], Eq. (D.6) is used for the calculation of the exponent in the combinatorial part of the activity coefficient in DDB:

$$r_i^{\#} = \frac{\sum\limits_{j \neq i} x_j r_j}{\sum\limits_{j \neq i} x_j} \tag{D.6}$$

As described in Ref. [124], small negative values of $p(\sigma)$ can occur. To keep consistency, the same algorithm for the treatment of negative $p(\sigma)$ is used here in NEAT for poorly specified mixtures.

D.8 NMR Spectra of the Systems

In the following the NMR spectrum of mixture one of each system is shown. The NMR spectra of the different mixtures of the same system differ only in the intensity of the peaks.

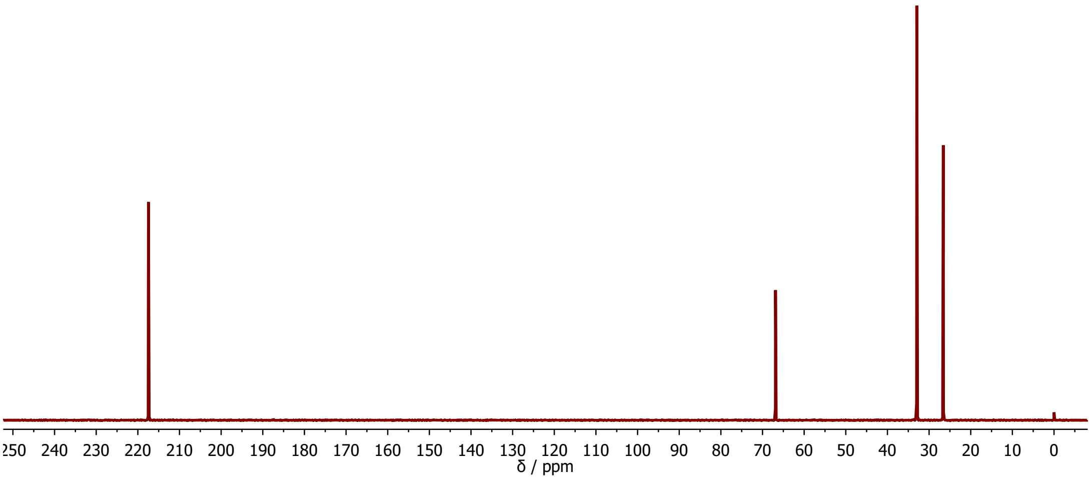


Figure D.10: ^{13}C NMR spectra of mixture 1 of system I with T = acetone, U = 2-propanol.

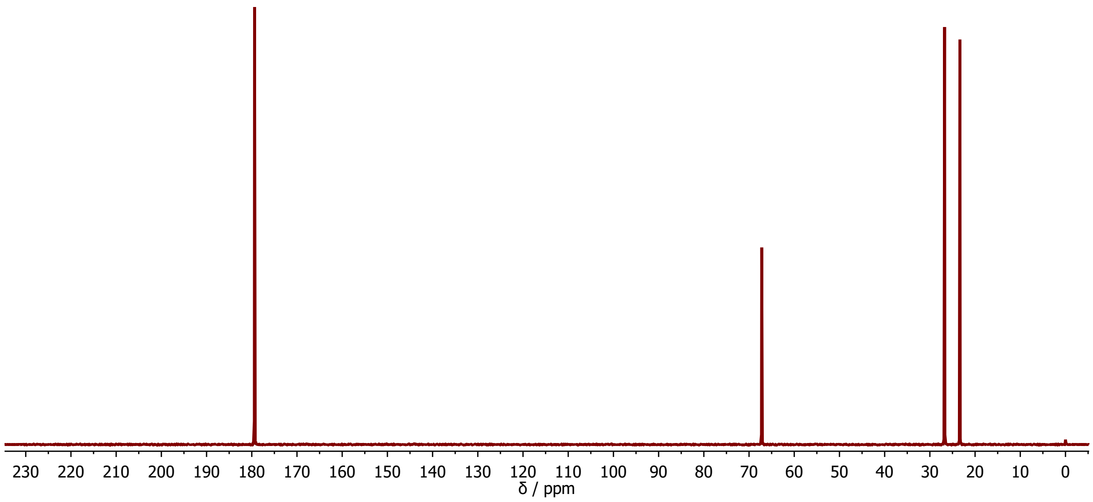


Figure D.11: ^{13}C NMR spectra of mixture 1 of system II with T = acetic acid, U = 2-propanol.

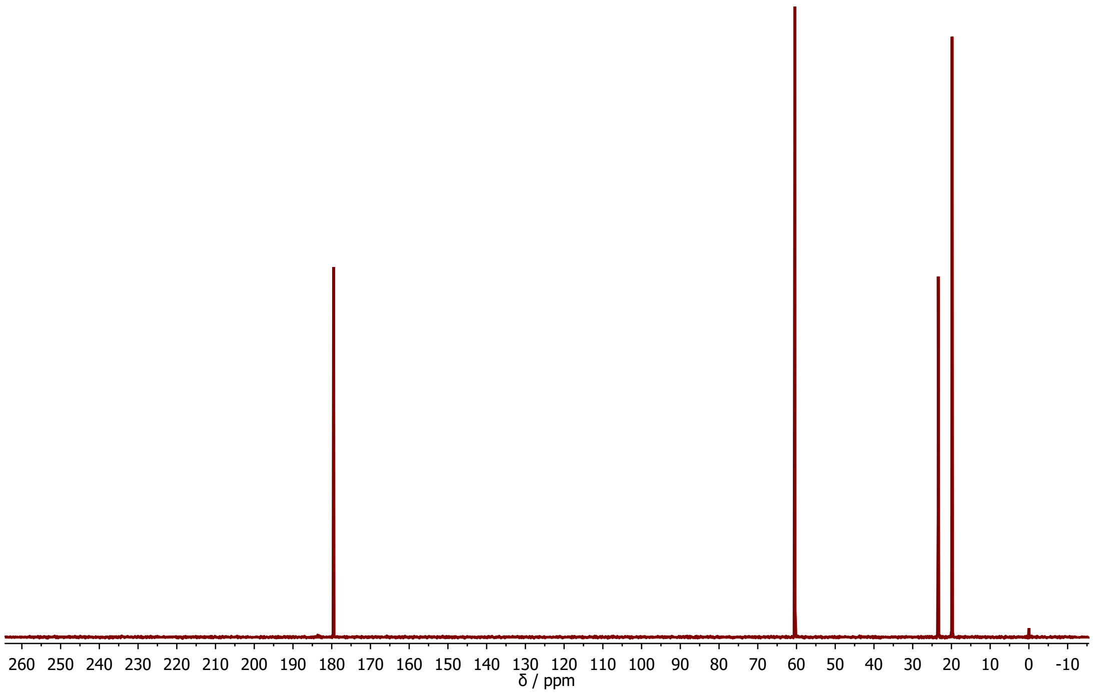


Figure D.12: ^{13}C NMR spectra of mixture 1 of system III with T = ethanol, U = acetic acid.

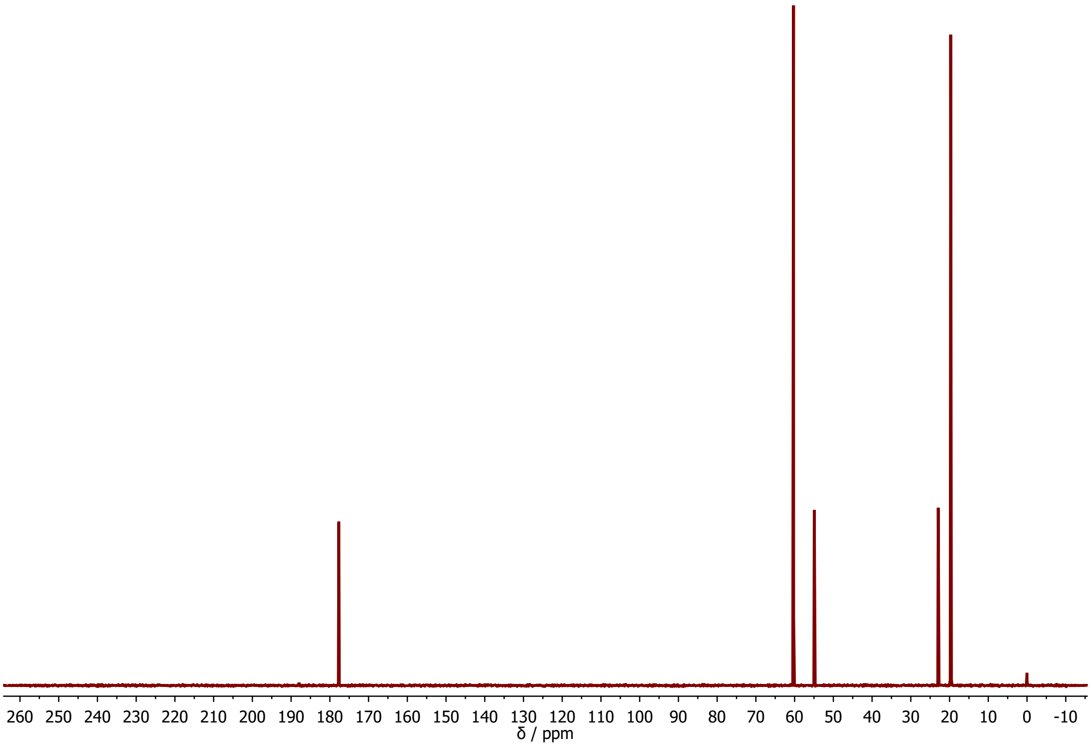


Figure D.13: ^{13}C NMR spectra of mixture 1 of system IV with T = ethanol, U = methyl acetate.

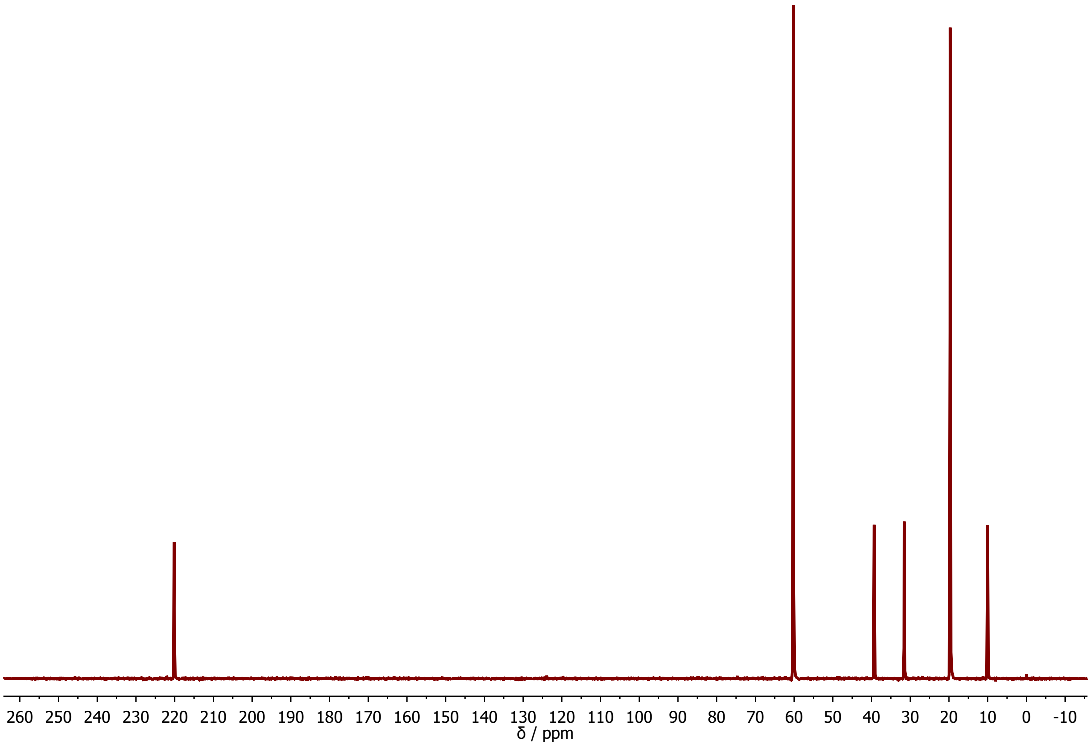


Figure D.14: ^{13}C NMR spectra of mixture 1 of system V with T = ethanol, U = 2-butanone.

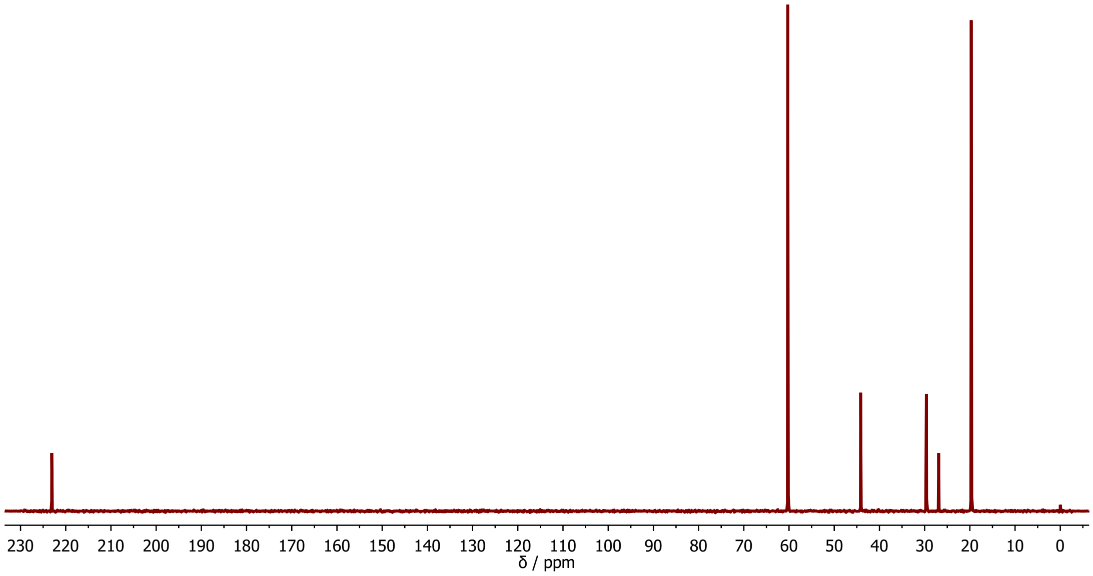


Figure D.15: ^{13}C NMR spectra of mixture 1 of system VI with T = ethanol, U = cyclohexanone.

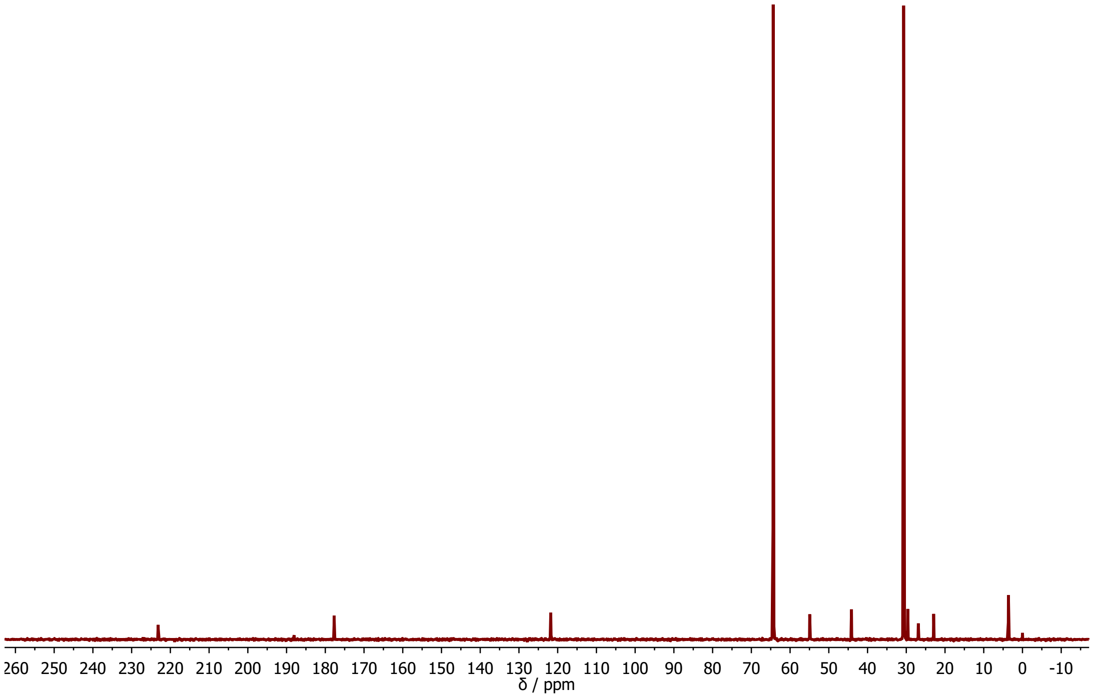


Figure D.16: ^{13}C NMR spectra of mixture 1 of system VII with T = 1,4-butanediol, U_1 = cyclohexanone, U_2 = acetonitrile, U_3 = methyl acetate.

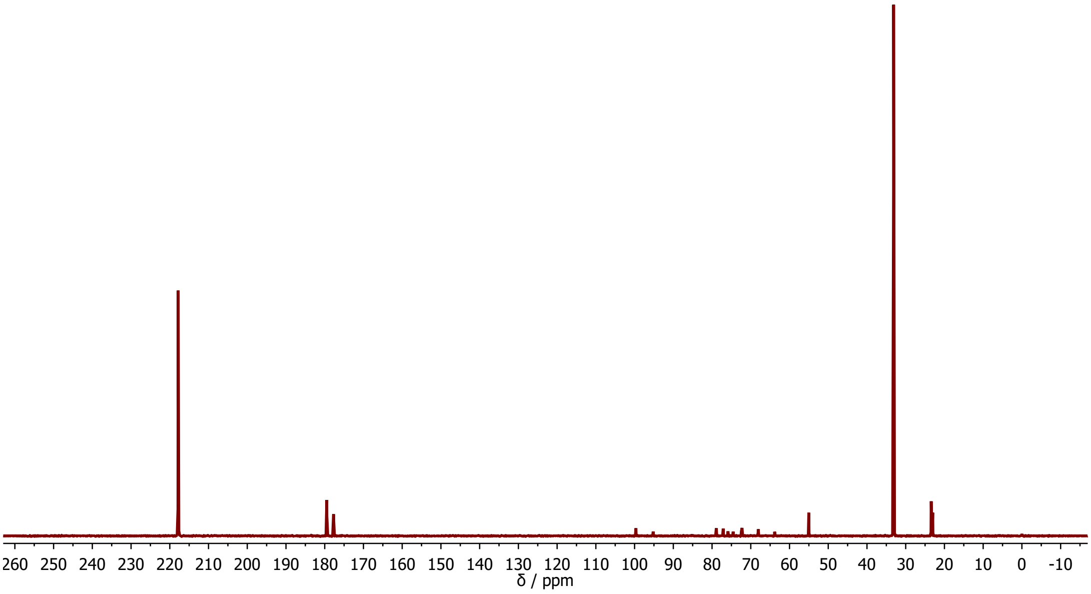


Figure D.17: ^{13}C NMR spectra of mixture 1 of system VIII with T = acetone, U_1 = D-xylose, U_2 = acetic acid, U_3 = methyl acetate.

D.9 Step-by-step Example for NEAT Application

For the prediction of the activity coefficient of a target component in a mixture with NEAT, mole fractions in the component space (target component T + pseudo-component $\tilde{U}$ + water W) have to be calculated. In the following, the calculation of the activity coefficient of the target component T = ethanol with U = 2-butanone and water W (system V, mixture 1) with NEAT is described as an example. A sample of 1 g of this mixture was considered besides a quantitative ^{13}C NMR spectrum of the mixture. The mass of T is known: m_T = 0.098 g (cf. Table D.8). The different regions of chemical shift in the ^{13}C NMR spectrum of the mixture are integrated as described in Table 9. Table D.24 shows the integration results for the example mixture. For simplicity, the peak area of ethanol at about 60 ppm was set to 1. As water shows no signal in ^{13}C NMR spectroscopy, the peak areas can only result from T (A_T) or U (A_U). The first step in the evaluation of the NMR spectra is the calculation of the contributions of the target component T to the peak areas in the different regions of the spectrum from the concentration of T. As the nature of the target component T is assumed to be known, the respective peaks in the spectrum can be assigned easily. The target component ethanol shows a peak at about 60 ppm, which was used as reference as described above, and a second peak at about 18 ppm. It is assumed that there is a proportionality of all signals of T to the concentration of T in the sample and that the proportionality constant is the same for all signals after taking into account the number of carbon atoms in the corresponding groups. This simplification yields good results as shown in [60], but calibration experiments could also be carried out [60]. Using the information on the target component T, the areas belonging to the unknown component U can be calculated as also shown in Table D.24. In the next step, peak areas associated to U have to be assigned to GC-COSMO-RS (OL) groups. Since the mass and molar mass of the target component T is known, the mole number of the target component can be calculated. All obtained groups of the unknown component are lumped to a pseudo-component $\tilde{\mathrm{U}}$ as described in Chapter 5. The mole numbers of these groups of the pseudo-component $\tilde{\mathrm{U}}$ are calculated using once more the assumption that the proportionality constant between the peak areas and the mole numbers is the same for all peaks using the target component area and corresponding mole number as reference. Since the molar mass of all chemical groups is known, the mass of the unknown groups in the pseudo-component $\tilde{\mathrm{U}}$ can also be calculated (cf. Table D.25). A summation of all groups yields the total mass associated to the pseudo-component $\tilde{\mathrm{U}}$: $m_{\tilde{\mathrm{U}}} = 0.032$ g (cf. Table D.8). From the mass balance, the mass of water W is calculated as $m_{\mathrm{W}} = 0.870$ g. Additionally, the group mole fractions in the pseudo-component $\tilde{\mathrm{U}}$ can be calculated, as shown in Table D.26. As described in Chapter 5, a molar mass of the pseudo-component $\tilde{\mathrm{U}}$ of $M_{\tilde{\mathrm{U}}} = 150$ g mol^{-1} was used in Chapter 5 as in Refs. [59, 60]. Hence, an estimated stochiometry of the pseudo-component was obtained (cf. Table D.8). Now, the σ-profile (cf. Figure D.7), cavity surface areas A, and the cavity volume V of the pseudo-component $\tilde{\mathrm{U}}$ can be calculated. In a final step the mole fractions in the component space target component T + pseudo-component $\tilde{\mathrm{U}}$ + water W are calculated. The activity coefficient of the target component is calculated using the implementation of the GC-COSMO-RS (OL) equations [124] in the DDB [125].

Table D.24: Integration results from the ^{13}C NMR spectrum of mixture 1 of system V studied as example here.

^{13}C NMR chemical shift region / ppm	Total area of peaks in region	A_{T}	A_{U}
0 - 30	1.23	1.00	0.23
30 - 60	0.46	0.00	0.46
60 - 90	1.00	1.00	0.00
90 - 150	0.00	0.00	0.00
150 - 180	0.00	0.00	0.00
>180	0.18	0.00	0.18

Table D.25: Assignment of GC-COSMO-RS (OL) groups associated to the pseudo-component $\tilde{\text{U}}$ and corresponding mole numbers $n_\text{g}^{\tilde{\text{U}}}$ and masses $m_\text{g}^{\tilde{\text{U}}}$. M_g is the molar mass of the group. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

^{13}C NMR chemical shift region / ppm	GC-COSMO-RS (OL) label	M_g/ g mol^{-1}	$n_\text{g}^{\tilde{\text{U}}}$/ mmol	$m_\text{g}^{\tilde{\text{U}}}$/ mg
0 - 30	’CH3’(1)	15.03	0.49	7.33
30 - 60	’CH2’(4)	14.03	0.97	13.67
60 - 90	’CH2’(7)^a+’OH(P)’(35)	14.03 / 17.01	0.00	0.00
90 - 150	’CH=CH’ (58)	26.04	0.00	0.00
150 - 180	’COOH’ (44)	45.02	0.00	0.00
>180	’CO’ (51)	28.01	0.38	10.68

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

Table D.26: GC-COSMO-RS (OL) group mole fraction for the pseudo-component $\tilde{\text{U}}$. The abbreviated group labels are introduced for clarity. The numbers in parentheses correspond to the group identifiers in the original paper [124].

GC-COSMO-RS (OL) label	Group mole fraction in $\tilde{\text{U}}$
’CH3’(1)	0.26
’CH2’(4)	0.53
’CH2’(7)^a+’OH(P)’(35)	0.00
’CH=CH’ (58)	0.00
’COOH’ (44)	0.00
’CO’ (51)	0.21

[a] In GC-COSMO-RS (OL), a ’CH2’ group attached to F/Cl/O/N is distinguished from a custom ’CH2’ group.

D.10 Temperature Dependency of Activity Coefficients

In Figure D.18, the temperature dependence of the activity coefficient γ_T of the target component in mixture 1 of system I is shown as an example. Since GC-COSMO-RS (OL) takes the temperature dependence of activity coefficients into account, NEAT based on GC-COSMO-RS (OL) can also predict the activity coefficients at various temperatures.

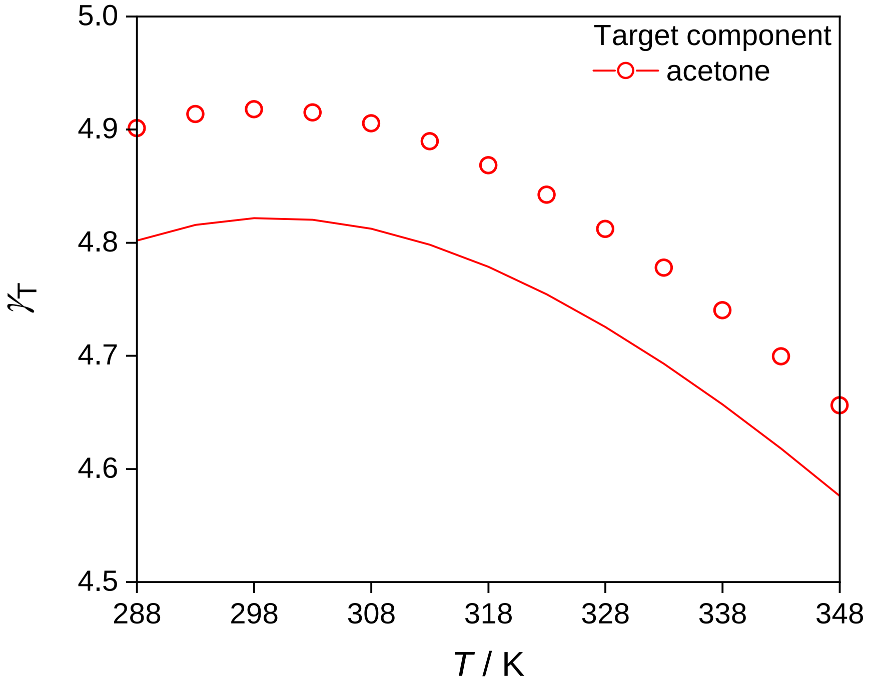


Figure D.18: Temperature dependence of activity coefficient γ_T of target component T = acetone in ternary mixture 1 of system I, cf. Table 11. Lines: results from GC-COSMO-RS (OL) for the fully specified mixtures. Symbols: predictions with NEAT based on an NMR analysis at T = 298 K. No information on the unknown component U was used in NEAT.

E Supporting Information for Chapter 6

E.1 Calculation of Liquid-liquid Equilibria

The partitioning of all (pseudo-)components in thermodynamic equilibrium was calculated by solving the isoactivity criterion together with mass balances and summation conditions formulated in the Rachford-Rice equation [153, 154]:

$$F(\theta) = \sum_{i=1}^{N}(x_i{''} - x_i{'}) = \sum_{i=1}^{N} \frac{x_i^{\text{mix}} K_i}{1 + \theta(K_i - 1)} - \frac{x_i^{\text{mix}}}{1 + \theta(K_i - 1)} \overset{!}{=} 0 \tag{E.1a}$$

$$K_i = \frac{\gamma_i{'}}{\gamma_i{''}}; \quad i = 1, ..., N \tag{E.1b}$$

where $x_i{'}$ and $x_i{''}$ are the mole fraction of (pseudo-)component i in the raffinate (water-rich) and the extract (extracting agent-rich) phase, respectively. $\gamma_i{'}$ and $\gamma_i{''}$ are the activity coefficients in the raffinate and extract phase, respectively, and $K_i = x_i{''}/x_i{'}$ is the partition coefficient of (pseudo-)component i. x_i^{mix} is the lumped mole fraction of i after mixing the feed with the extracting agent E (irrespective of the phase separation), and $\theta = \frac{n''}{n'+n''}$ is the molar extract phase fraction.

Eq. (E.1a) and Eq. (E.1b) were solved in a double-loop approach [154–156] for a constant temperature of $T = 298.15$ K. Therefore, first, the values of K_i for all (pseudo-)components were calculated using UNIFAC [31, 127] for modeling the activity coefficients in both coexisting phases. Then, Eq. (E.1a) was solved with respect to θ using the MATLAB [104] 2021b function “fsolve”, from which updated composition of the extract and raffinate phase and, consequently, updated values for K_i were obtained. This process was repeated until convergence of θ was observed, specifically, until θ changed by less than 10^{-12} between two successive iterations.

E.2 Prediction of Residue Curves

Residue curves were modeled by the Rayleigh-equation [131, 132]:

$$\frac{dx_i}{dn^{\mathrm{L}}} = \frac{y_i - x_i}{n^{\mathrm{L}}}; \quad i = 1 \ldots N-1 \tag{E.2a}$$

$$x_{\mathrm{W}} = 1 - \sum_{i=1}^{N-1} x_i \tag{E.2b}$$

where n^{L} is the total mole number in the liquid phase (which decreases with evaporation over time), x_i and y_i are the mole fractions of all components i except for the solvent water (W) in the liquid and the vapor phase, respectively, in thermodynamic equilibrium. The mole fraction of water in the liquid phase x_{W} was calculated by the summation condition, cf. Eq. (E.2b).

The system of ordinary differential equations defined by Eq. (E.2a) was solved in MATLAB using the composition of the feed as starting composition, i.e., $\boldsymbol{x}(n^{\mathrm{L}} = n^{\mathrm{L},0}) = \boldsymbol{x}_{\mathrm{Feed}}$. The calculation of the residue curves was stopped if the number of moles n^{L} approaches the total amount of non-volatile components (to avoid numerical issues, a small tolerance margin was added).

The vapor-liquid equilibria were modeled by Raoult's law, cf. Eq. (14), whereby the pressure p and all mole fractions x_i in the liquid phase for the respective time step were specified, and the activity coefficients γ_i in the liquid phase were calculated with UNIFAC [31, 127]. Note that for numerical reasons, first, the boiling temperature of the mixture T was calculated by adjusting T until the sum of the partial pressures p_i in Eq. (14) over all N components was equal to the total (specified) pressure p. Subsequently, the mole fractions y_i in the gas phase were calculated.

For the fully specified mixtures, which were considered for comparison only, the vapor pressure of each pure component was calculated with the Antoine equation:

$$\log_{10}\left(750.062 \frac{p_i^{\mathrm{S}}}{\mathrm{bar}}\right) = A_i - \frac{B_i}{\left(\frac{T}{\mathrm{K}} - 273.15\right) + C_i} \tag{E.3}$$

whereby the component-specific constants (A_i, B_i, C_i) were taken from the Dortmund Data Bank (DDB) [133]. For most of the studied components, the DDB reports parameter sets for two different temperature ranges. If only one parameter set was labeled as valid for the temperature of interest here, this one was used; if both parameter sets were labeled as suitable, both sets were used for calculating the vapor pressure p_i^{S}, and the mean value of the two p_i^{S} was subsequently used. In some cases, the vapor pressure calculation was also performed outside the given ranges.

The vapor pressures of the pseudo-components k in the poorly specified mixtures were estimated using group-contribution method from Refs. [27, 29]:

$$\log_{10}\left(\frac{p_k^{\mathrm{S}}}{1.01325\ \mathrm{bar}}\right) = (4.1012 + dB_k)\left(\frac{\frac{T/\mathrm{K}}{T_{\mathrm{b},k}/\mathrm{K}} - 1}{\frac{T/\mathrm{K}}{T_{\mathrm{b},k}/\mathrm{K}} - 0.125}\right) \tag{E.4}$$

where $T_{\mathrm{b},k}$ is the normal boiling point and dB_k is the so-called "slope term" [29] of pseudo-component k. For calculating $T_{\mathrm{b},k}$, the approach from Ref. [27] was used; for calculating dB_k, the approach from Ref. [29] was used. At the heart of both group-contribution approaches, the absolute number of groups in each (pseudo-)component is required, which was directly obtained by the NMR fingerprinting and pseudo-component method described in Ref. [128] and is given in Tables E.4-E.6. The considered structural groups, as well as the assignment procedure, are described in Table 12.

In the group-contribution method of Refs. [27, 29], a so-called group-interaction contribution has to be considered for strongly interacting groups [27, 29], which holds for the hydroxyl ('OH'), carboxylic acid ('COOH'), ketone ('CO') and aldehyde ('CHO') groups here. The calculation of the group-interaction contribution requires that the absolute number of each interactive structural group is an integer in each pseudo-component. Therefore, the absolute number of the interactive groups in each pseudo-component was rounded to the nearest integer only for the calculation of the group-interaction contribution.

E.3 Prediction of Feed Composition

In the following, the calculation of the feed composition on the basis of the NMR fingerprinting and pseudo-component method is described. The feed composition is the basis for the thermodynamic modeling of the mixtures. For predicting the composition of a poorly specified feed with the proposed methodology, the following information was assumed to be available and used:

- the total mass of the poorly specified feed m;
- the nature of the solvent (here: always water W);
- the nature and mass fraction $x_{\mathrm{T}}^{(\mathrm{m})}$ of the target component T, in the feed.

From the assumptions, the amount of the target component was calculated:

$$n_{\mathrm{T}} = \frac{x_{\mathrm{T}}^{(\mathrm{m})} m}{M_{\mathrm{T}}} \tag{E.5}$$

Additionally, a mean area $\bar{A}_\mathrm{T}$ for the target component was calculated:

$$\bar{A}_\mathrm{T} = \frac{A_\mathrm{T}^\mathrm{total}}{z_\mathrm{T}^\mathrm{total}} \tag{E.6}$$

where $z_\mathrm{T}^\mathrm{total}$ is the total number of NMR-active nuclei in the target component and $A_\mathrm{T}^\mathrm{total}$ is the total area of the target component in the corresponding spectrum. Furthermore, a mean area of a structural group g was defined, taken from the UNIFAC table [31, 127], cf. Table 12, in pseudo-component k:

$$\bar{A}_{g,k} = \frac{A_{g,k}}{z_g} \tag{E.7}$$

where z_g is the number of NMR-active nuclei in group g that are expected in the same chemical shift region and $A_{g,k}$ was the total area of group g assigned to pseudo-component k. Note that for the CH3CO/CH2CO group in UNIFAC, one typically expects two peaks in different chemical shift regions. The CH3CO/CH2CO group was therefore only identified if in both regions of the ^{13}C NMR spectrum “0-90 ppm” and “>180 ppm” peaks were observed *and* assigned to the same pseudo-component; as a consequence, 'CH3/CH2' groups were only assigned to the peaks in the region “0-90 ppm” that exceeded the peaks in the region “>180 ppm” for each pseudo-component. From this, the mole number of each structural group g in pseudo-component k was calculated:

$$n_{g,k} = \frac{\bar{A}_{g,k}}{\bar{A}_\mathrm{T}} n_\mathrm{T} \tag{E.8}$$

This yielded the total mass of all groups in pseudo-component k:

$$m_k = \sum_{g=1}^{G} n_{g,k} M_g \tag{E.9}$$

where G is the total number of considered structural groups and M_g the known molar mass of group g. With the known total mass of the pseudo-components, the mass of the target component, and the total mass of the mixture, the mass of the solvent W can be calculated:

$$m_\mathrm{W} = m - m_\mathrm{T} - \sum_{k=1}^{K} m_k \tag{E.10}$$

Using the information on the molar masses of the respective (pseudo-)components, the mole fraction was calculated as follows:

$$x_i = \frac{n_i}{\sum_{i=1}^{N} n_i} \tag{E.11}$$

E.4 Predicted Composition of Pseudo-components

In Tables E.1-E.6, information on the stoichiometry of the pseudo-component as predicted with the approach of the present thesis for all studied mixtures is summarized. For this purpose, the absolute numbers of the structural groups g in each molecule of the defined pseudo-components k, denoted by $\nu_{g,k}$, in the three test mixtures are compiled. Tables E.1-E.3 thereby consider groups from the UNIFAC [31, 127] table, whereas Tables E.4-E.6 consider groups from the table of Refs. [27, 29].

Table E.1: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture I, cf. Table 12 and 13, according to UNIFAC [31, 127, 129]. The numbers in parentheses are the identifiers for the UNIFAC sub-groups.

UNIFAC label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$	$\tilde{\mathrm{U}}_3$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	0.782	-	1.147
CH2 (2)	$\nu_{\mathrm{CH2},k}$	-	6.464	-
CH (3)	$\nu_{\mathrm{CH},k}$	-	-	-
C (4)	$\nu_{\mathrm{C},k}$	-	-	-
OH (14)	$\nu_{\mathrm{OH},k}$	-	3.206	-
CH=CH (6)	$\nu_{\mathrm{CH=CH},k}$	-	-	-
C=C (70)	$\nu_{\mathrm{C=C},k}$	-	-	-
COOH (42)	$\nu_{\mathrm{COOH},k}$	-	-	1.219
CHO (20)	$\nu_{\mathrm{CHO},k}$	-	-	-
CH3CO (18)	$\nu_{\mathrm{CH3CO},k}$	0.865	-	-
CH2CO (19)	$\nu_{\mathrm{CH2CO},k}$	-	-	-

Table E.2: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture II, cf. Table 12 and 13, according to UNIFAC [31, 127, 129]. The numbers in parentheses are the identifiers for the UNIFAC sub-groups.

UNIFAC label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	-	-
CH2 (2)	$\nu_{\mathrm{CH2},k}$	3.879	1.898
CH (3)	$\nu_{\mathrm{CH},k}$	-	1.680
C (4)	$\nu_{\mathrm{C},k}$	-	0.759
OH (14)	$\nu_{\mathrm{OH},k}$	-	2.846
CH=CH (6)	$\nu_{\mathrm{CH=CH},k}$	-	0.203
C=C (70)	$\nu_{\mathrm{C=C},k}$	-	-
COOH (42)	$\nu_{\mathrm{COOH},k}$	-	2.238
CHO (20)	$\nu_{\mathrm{CHO},k}$	-	-
CH3CO (18)	$\nu_{\mathrm{CH3CO},k}$	-	-
CH2CO (19)	$\nu_{\mathrm{CH2CO},k}$	0.953	-

Table E.3: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture III, cf. Table 12 and 13, according to UNIFAC [31, 127, 129]. The numbers in parentheses are the identifiers for the UNIFAC sub-groups.

UNIFAC label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$	$\tilde{\mathrm{U}}_3$	$\tilde{\mathrm{U}}_4$	$\tilde{\mathrm{U}}_5$	$\tilde{\mathrm{U}}_6$	$\tilde{\mathrm{U}}_7$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	0.853	1.111	2.111	-	-	-	-
CH2 (2)	$\nu_{\mathrm{CH2},k}$	-	-	1.409	4.688	6.625	1.812	1.617
CH (3)	$\nu_{\mathrm{CH},k}$	-	-	0.712	-	-	3.681	3.135
C (4)	$\nu_{\mathrm{C},k}$	-	-	-	-	-	-	-
OH (14)	$\nu_{\mathrm{OH},k}$	-	-	1.440	-	3.291	4.577	4.752
CH=CH (6)	$\nu_{\mathrm{CH=CH},k}$	-	-	-	-	-	0.435	-
C=C (70)	$\nu_{\mathrm{C=C},k}$	-	-	-	-	-	-	0.858
COOH (42)	$\nu_{\mathrm{COOH},k}$	-	1.208	-	-	-	1.907	3.308
CHO (20)	$\nu_{\mathrm{CHO},k}$	-	-	-	-	-	-	-
CH3CO (18)	$\nu_{\mathrm{CH3CO},k}$	0.903	-	-	-	-	-	-
CH2CO (19)	$\nu_{\mathrm{CH2CO},k}$	-	-	-	1.194	-	-	-

Table E.4: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture I, cf. Table 12 and 13, according to Refs. [27, 29].

Nannonal label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$	$\tilde{\mathrm{U}}_3$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	1.647	-	1.147
CH2 (4)	$\nu_{\mathrm{CH2},k}$	-	3.258	-
CH (5)	$\nu_{\mathrm{CH},k}$	-	-	-
C (6)	$\nu_{\mathrm{C},k}$	-	-	-
CH2 (7)	$\nu_{\mathrm{CH2},k}$	-	3.206	-
CH (7)	$\nu_{\mathrm{CH},k}$	-	-	-
C (7)	$\nu_{\mathrm{C},k}$	-	-	-
OH(P) (35)	$\nu_{\mathrm{OH(P)},k}$	-	3.206	-
OH(P) (36)	$\nu_{\mathrm{OH(P)},k}$	-	-	-
OH(S) (34)	$\nu_{\mathrm{OH(S)},k}$	-	-	-
OH(T) (33)	$\nu_{\mathrm{OH(T)},k}$	-	-	-
CH=CH (58)	$\nu_{\mathrm{CH=CH},k}$	-	-	-
C=C (58)	$\nu_{\mathrm{C=C},k}$	-	-	-
COOH (44)	$\nu_{\mathrm{COOH},k}$	-	-	1.219
CHO (52)	$\nu_{\mathrm{CHO},k}$	-	-	-
CO (51)	$\nu_{\mathrm{CO},k}$	0.865	-	-

Table E.5: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture II, cf. Table 12 and 13, according to Refs. [27, 29].

Nannonal label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	-	-
CH2 (4)	$\nu_{\mathrm{CH2},k}$	4.832	1.491
CH (5)	$\nu_{\mathrm{CH},k}$	-	-
C (6)	$\nu_{\mathrm{C},k}$	-	-
CH2 (7)	$\nu_{\mathrm{CH2},k}$	-	0.407
CH (7)	$\nu_{\mathrm{CH},k}$	-	1.680
C (7)	$\nu_{\mathrm{C},k}$	-	0.759
OH(P) (35)	$\nu_{\mathrm{OH(P)},k}$	-	0.407
OH(P) (36)	$\nu_{\mathrm{OH(P)},k}$	-	-
OH(S) (34)	$\nu_{\mathrm{OH(S)},k}$	-	1.680
OH(T) (33)	$\nu_{\mathrm{OH(T)},k}$	-	0.759
CH=CH (58)	$\nu_{\mathrm{CH=CH},k}$	-	-
C=C (58)	$\nu_{\mathrm{C=C},k}$	-	-
COOH (44)	$\nu_{\mathrm{COOH},k}$	-	2.238
CHO (52)	$\nu_{\mathrm{CHO},k}$	-	-
CO (51)	$\nu_{\mathrm{CO},k}$	0.953	-

Table E.6: Absolute numbers $\nu_{g,k}$ of structural groups g in pseudo-components k for test mixture III, cf. Table 12 and 13, according to Refs. [27, 29].

Nannonal label	Stoichiometry	$\tilde{\mathrm{U}}_1$	$\tilde{\mathrm{U}}_2$	$\tilde{\mathrm{U}}_3$	$\tilde{\mathrm{U}}_4$	$\tilde{\mathrm{U}}_5$	$\tilde{\mathrm{U}}_6$	$\tilde{\mathrm{U}}_7$
CH3 (1)	$\nu_{\mathrm{CH3},k}$	1.757	1.111	2.111	-	-	-	-
CH2 (4)	$\nu_{\mathrm{CH2},k}$	-	-	0.682	5.883	3.334	0.916	-
CH (5)	$\nu_{\mathrm{CH},k}$	-	-	-	-	-	-	-
C (6)	$\nu_{\mathrm{C},k}$	-	-	-	-	-	-	-
CH2 (7)	$\nu_{\mathrm{CH2},k}$	-	-	0.728		3.291	0.895	1.617
CH (7)	$\nu_{\mathrm{CH},k}$	-	-	0.712	-	-	3.681	3.135
C (7)	$\nu_{\mathrm{C},k}$	-	-	-	-	-	-	-
OH(P) (35)	$\nu_{\mathrm{OH(P)},k}$	-	-	-	-	3.291	0.895	1.617
OH(P) (36)	$\nu_{\mathrm{OH(P)},k}$	-	-	0.728	-	-	-	-
OH(S) (34)	$\nu_{\mathrm{OH(S)},k}$	-	-	0.712	-	-	3.681	3.135
OH(T) (33)	$\nu_{\mathrm{OH(T)},k}$	-	-	-	-	-	-	-
CH=CH (58)	$\nu_{\mathrm{CH=CH},k}$	-	-	-	-	-	0.435	-
C=C (58)	$\nu_{\mathrm{C=C},k}$	-	-	-	-	-	-	0.858
COOH (44)	$\nu_{\mathrm{COOH},k}$	-	1.208	-	-	-	1.907	3.308
CHO (52)	$\nu_{\mathrm{CHO},k}$	-	-	-	-	-	-	-
CO (51)	$\nu_{\mathrm{CO},k}$	0.903	-	-	1.194	-	-	-

Table E.7 shows the composition of all components in the test mixtures regarding the groups of UNIFAC [31, 127, 129]. Table E.8 shows the composition of all extracting agents regarding the groups of UNIFAC [31, 127, 129]. Note that the UNIFAC nomenclature uses 'THF,' [129] as an abbreviation for cyclic ether groups, but since it is seen as misleading 'cy-CH2O' is used as abbreviation instead.

Table E.7: Components considered in the test mixtures, cf. Table 13, and their composition regarding groups from the UNIFAC table [31, 127, 129]. The numbers in parentheses are the identifiers for the UNIFAC sub-groups.

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

Table E.8: Extracting agents considered in Chapter 6 and their composition regarding groups from the UNIFAC table [31, 127, 129]. The numbers in parentheses are the identifiers for the UNIFAC sub-groups.

Component	UNIFAC groups
1-decanol	1 x ’CH3’ (1)
	9 x ’CH2’ (2)
	1 x ’OH’ (14)
decane	2 x ’CH3’ (1)
	8 x ’CH2’ (2)
dipropyl ether	2 x ’CH3’ (1)
	3 x ’CH2’ (2)
	1 x ’CH2O’ (25)
ethyl propionate	2 x ’CH3’ (1)
	1 x ’CH2’ (2)
	1 x ’CH2COO’ (22)
hexane	2 x ’CH3’ (1)
	4 x ’CH2’ (2)
toluene	5 x ’ACH’ (9)
	1 x ’ACCH3’ (11)
1-octanol	1 x ’CH3’ (1)
	7 x ’CH2’ (2)
	1 x ’OH’ (14)
3-octanone	2 x ’CH3’ (1)
	4 x ’CH2’ (2)
	1 x ’CH2CO’ (19)

E.5 Results for Pure-component Vapor Pressures

Figures E.1 to E.3 shows the results for the vapor pressures of the (pseudo-)components over the boiling temperature T. For the true components (solid / dotted lines), the results were obtained with the Antoine equation, cf. Eq. (E.3), for the respective pseudo-components (dashed lines), the group-contribution method of Nannonal [29], cf. Eq. (E.4), was used.

For five of the components used in this study, namely glucose, xylose, citric acid, ascorbic acid, and malic acid, no Antoine-parameters were available in the DDB. However, it can be assumed that the vapor pressure of these components in the considered temperature range is negligible, in particular in comparison to the other considered components; therefore, these components were assumed to remain completely in the liquid phase throughout ($y_i = 0$), and Eq. (14) was not considered here. For (pseudo-)components, for which the vapor pressure was assumed negligible ($p_i^{\mathrm{S}} = 0$), no results are shown here.

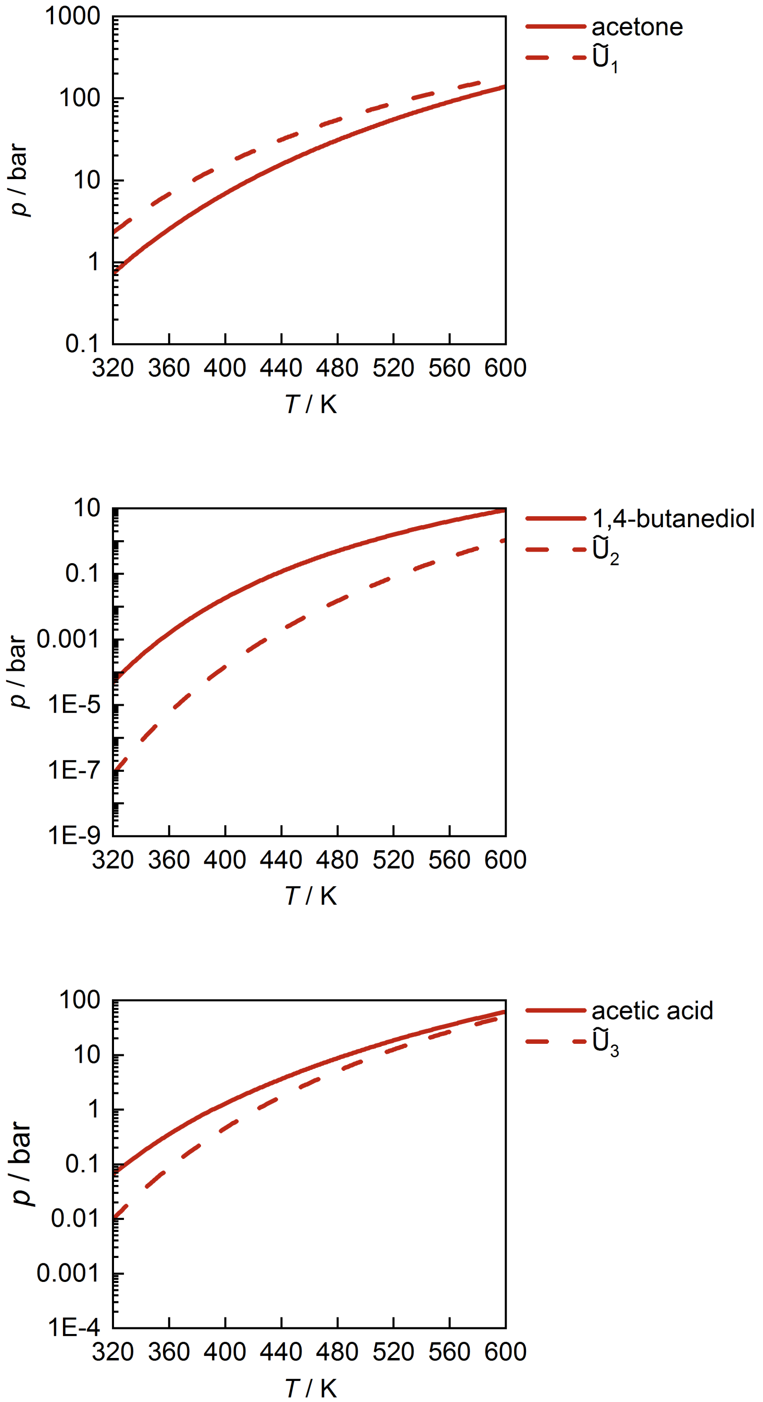


Figure E.1: Vapor pressures p^{S} of pure (pseudo-)components in mixture I. Solid lines: true components, calculated with the Antoine equation, cf. Eq. (E.3). Dashed lines: pseudo-components, estimated with the group-contribution method of Ref. [29], cf. Eq. (E.4).

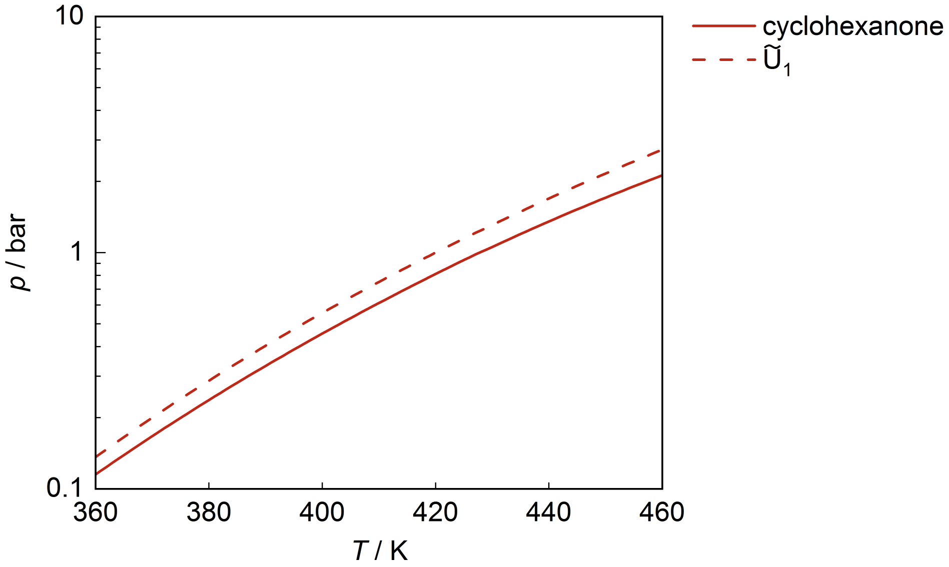


Figure E.2: Vapor pressures p^{S} of pure (pseudo-)components in mixture II. Solid lines: true component, calculated with the Antoine equation, cf. Eq. (E.3). Dashed lines: pseudo-component, estimated with the group-contribution method of Ref. [29], cf. Eq. (E.4).

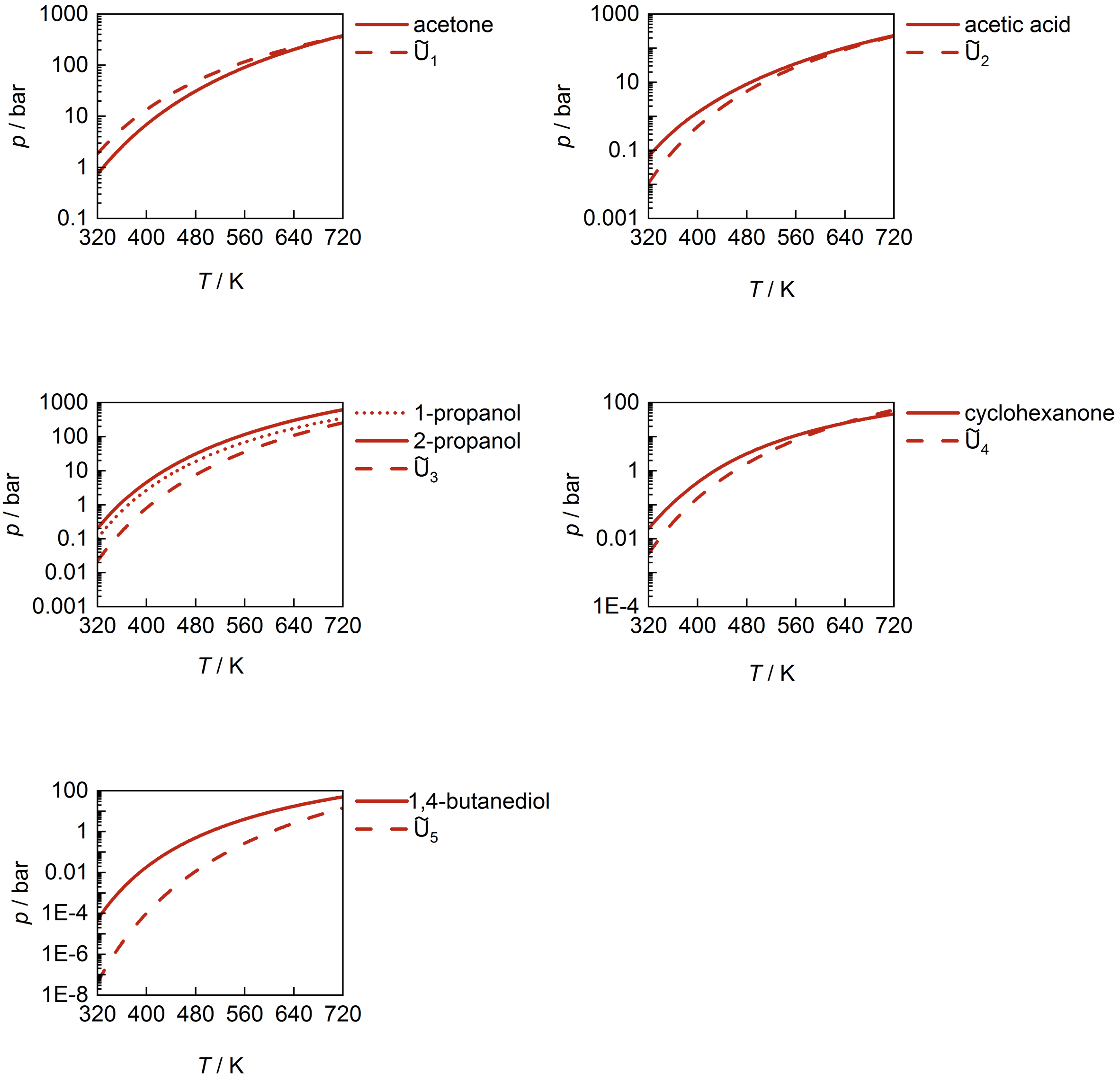


Figure E.3: Vapor pressures p^S of pure (pseudo-)components in mixture III. Solid and dotted lines: true components, calculated with the Antoine equation, cf. Eq. (E.3). Dashed lines: pseudo-components, estimated with the group-contribution method of Ref. [29], cf. Eq. (E.4).

E.6 Additional Results for the Prediction of Residue Curves

In Figures E.4 and E.5, additional results for the residue curves of test mixtures II and III, cf. Table 13, are shown. In contrast to the results shown in Chapter 6, the concentration of all true components is presented separately here, even if they were lumped into pseudo-components by the algorithms.

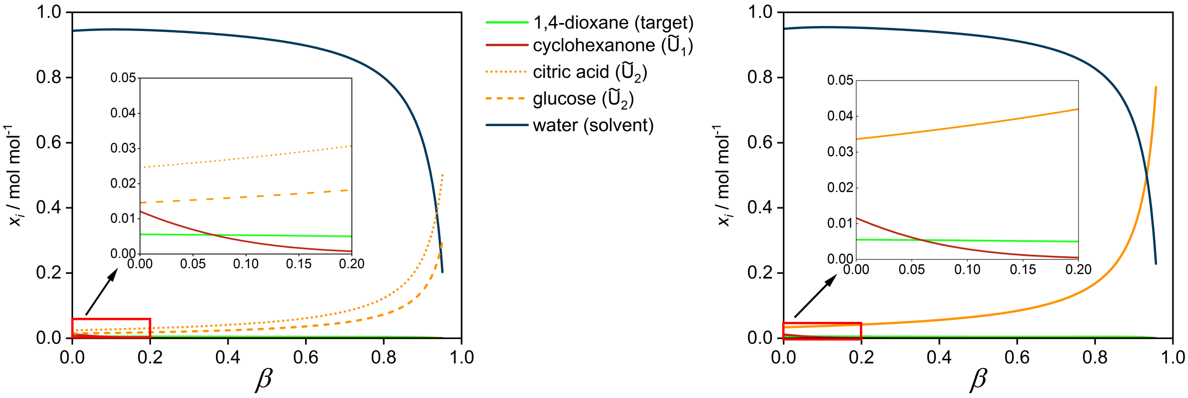


Figure E.4: Residue curves showing the liquid-phase mole fractions as a function of the evaporation ratio β for mixture II, cf. Table 13, at p = 1 bar. Left: results obtained using the full speciation. Right: predictions for the poorly specified feed based on NMR fingerprinting and the pseudo-component method.

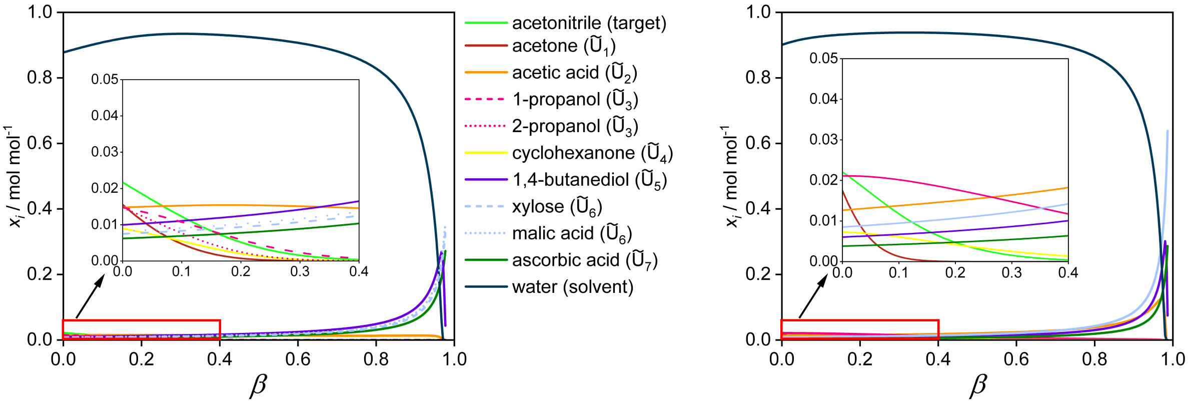


Figure E.5: Residue curves showing the liquid-phase mole fractions as a function of the evaporation ratio β for mixture III, cf. Table 13, at p = 1 bar. Left: results obtained using the full speciation. Right: predictions for the poorly specified feed based on NMR fingerprinting and the pseudo-component method.

Statement on Authorship

This dissertation contains material that has been published previously or that is included in submitted publications. In the following, these publications are listed together with a statement on the contributions of the author of the present dissertation.

- T. Specht, K. Münnemann, F. Jirasek, H. Hasse: Estimating activity coefficients of target components in poorly specified mixtures with NMR spectroscopy and COSMO-RS, Fluid Phase Equilibria 516 (2020) 112604,
  DOI: `https://doi.org/10.1016/j.fluid.2020.112604`.
  The author carried out or supervised the experiments and modeling. The author wrote the manuscript.
- T. Specht, K. Münnemann, H. Hasse, F. 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 (2021) 143-155,
  DOI: `https://doi.org/10.1021/acs.jcim.0c01186`.
  The author carried out or supervised the experiments. The author developed and trained the machine-learning model. The author wrote the manuscript.
- 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: `https://doi.org/10.1039/D3CP00509G`.
  The author carried out or supervised the experiments and modeling. The author wrote the manuscript.
- T. Specht, H. Hasse, F. Jirasek: Predictive thermodynamic modeling of poorly specified mixtures and applications in conceptual fluid separation process design, Industrial & Engineering Chemistry Research 62 (2023) 10657-10667,
  DOI: `https://doi.org/10.1021/acs.iecr.3c01096`.
  The author implemented the models for phase equilibria calculation. The author wrote the manuscript.

- T. Specht, J. Arweiler, J. Stüber, K. Münnemann, H. Hasse, F. Jirasek: Automated nuclear magnetic resonance fingerprinting of mixtures, Magnetic Resonance in Chemistry (2023),
  DOI: `https://doi.org/10.1002/mrc.5381`.
  The author carried out the experiments together with Justus Arweiler. The author developed and trained the machine-learning model. The author wrote the manuscript.

Student Theses

The following student theses were prepared under the supervision of the author of the present doctoral thesis in the frame of his research:

- N. Mollner: NMR spectroscopic method for estimating activity coefficients of target components in poorly specified mixtures with COSMO-RS (Ol). Student thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2019).
- J. Stüber: Determination of chemical groups from ^{13}C NMR spectra with machine learning. Bachelor thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2020).
- J. Arweiler: Determination of the molar mass of unknown components from self-diffusion measurements using NMR spectroscopy. Bachelor thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2020).
- K. Just: Identification of chemical groups from NMR spectra by machine learning. Student thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2021).
- J. Arweiler: Prediction of activity coefficients at infinite dilution using hybrid matrix completion methods with quantum-chemical descriptors. Student thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2022).
- K. Just: Extension of a method for the analysis of structural groups from NMR spectra by machine learning. Bachelor thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2022).
- J. Stüber: Prediction of physicochemical properties with hybrid approaches of matrix completion methods and COSMO-RS. Student thesis, Laboratory of Engineering Thermodynamics (LTD), University of Kaiserslautern (2022).

Curriculum Vitae

Name: Thomas Specht
Geburtsort: Koblenz
Staatsangehörigkeit: deutsch

Schulbildung

1999 – 2003	Grundschule Neuhäusel
2003 – 2012	Goethe-Gymnasium Bad Ems Abschluss: Allgemeine Hochschulreife

Studium

2012 – 2013	Fachhochschule Bingen Studiengang: Biotechnik
2013 – 2016	Technische Universität Kaiserslautern Studiengang: Bio- und Chemieingenieurwissenschaften Abschluss: B.Sc.
2016 – 2018	Technische Universität Kaiserslautern Studiengang: Bio- und Chemieingenieurwissenschaften Abschluss: M.Sc.

Berufliche Tätigkeit

seit 2018	Wissenschaftlicher Mitarbeiter am Lehrstuhl für Thermodynamik Prof. Dr.-Ing. Hans Hasse Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau (bis 2022: Technische Universität Kaiserslautern)