Deep Set Models for Elucidating Unknown Mixtures with NMR Spectroscopy

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

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

Abstract. The elucidation of unknown mixtures is a significant challenge in chemistry and chemical engineering, where accurate specifications are essential for efficient process design and operation. We propose an 'NMR fingerprinting' method to automate the classification of structural groups in mixtures based on standard nuclear magnetic resonance (NMR) experiments and a deep set model (DSM). The DSM is trained on experimental NMR spectra of pure components, augmented with synthetic spectral data, and comprises invariant and equivariant network structures to ensure predictions independent of input size and permutations. When applied to test mixtures, the predictions by NMR fingerprinting agree well with the true mixture compositions.

1 Introduction

Complex mixtures of unknown compositions are ubiquitous in chemistry and chemical engineering and constitute a stiff challenge for designing and optimizing efficient processes. Nuclear magnetic resonance (NMR) spectroscopy is a powerful analytical technique generally suited for this purpose, but for complex mixtures, elucidation is expensive, requires expert knowledge, and often results in ambiguity. Hence, reliable automated methods for elucidating unknown mixtures are precious.

In prior work, we proposed 'NMR fingerprinting' for automatically identifying the structural groups, the building blocks of components, in an unknown mixture based on standard NMR spectra, which is much simpler than obtaining the respective information on the components. From a machine learning (ML) perspective, this is a classification problem, i.e., assigning the correct groups to the signals in the NMR spectra, so a support vector classification (SVC) was developed and trained [1, 2]. Based on the obtained group-specific fingerprints, pseudo-components can be defined [3], which can subsequently be used for thermodynamic modeling of unknown mixtures [4].

However, the SVC-based NMR fingerprinting has significant limitations in its application, mainly due to varying input sizes (number of signals in the NMR spectra) in this application. This work overcomes this limitation by developing a classification model based on a deep-set architecture [5]. Furthermore, the approach is extended by incorporating information from 2D NMR experiments and using data augmentation during training.

2 Methodology

Figure 1 shows an overview of the proposed model architecture. The training

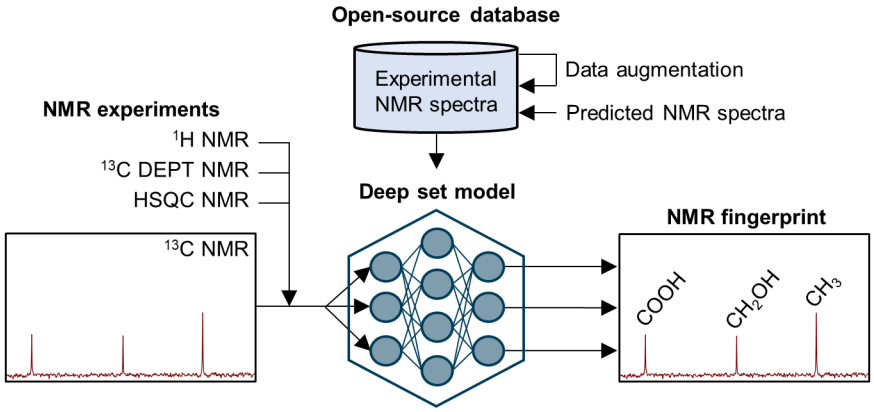

Fig. 1. Overview of the NMR fingerprinting method developed in this work to classify 13 structural groups in NMR spectra of mixtures using a deep set model (DSM) trained on experimental pure-component NMR spectra from the open-source databases BMRB [6] and NMRShiftDB [7].

data for the deep set model DSM was derived from the open-source NMR spectra databases BMRB [6] and NMRShiftDB [7], with missing data augmented from magnetically equivalent nuclei detected using RDKit [8]. Further, missing spectral information was augmented by predicting the NMR spectra of the respective components using the open-source tool NMRium [9].

The model proposed in this work to classify 13 structural groups from NMR spectra combines an invariant and equivariant DSM architecture. In the first step, the information from multiple NMR measurements is processed to obtain an intermediate prediction for each structural group invariant to the permutation of the input order. In the second step, the intermediate predictions are refined in the context of all structural groups in the mixture and assigned to signals in the ^{13}C NMR spectrum, ensuring equivariant classification results.

3 Results

Figure 2 displays the model's performance regarding the F_1-score for predicting the structural groups from 10% of the pure-component NMR spectra, randomly chosen as test data. The model shows overall good performance, reaching a macroscopic F_1-score $F_{1,\mathrm{macro}} = 0.92$. Although the model was trained only on pure-component data, it also applies to NMR spectra of unknown mixtures.

To demonstrate the transferability to mixture spectra, Figure 3a and Figure 3b show the model predictions for a test mixture of diethyl ether, butanal, and butyl acetate as an example. The proposed model correctly identifies all structural groups contained in the mixture, additionally presenting the confidence for classifying the signals into the different structural groups.

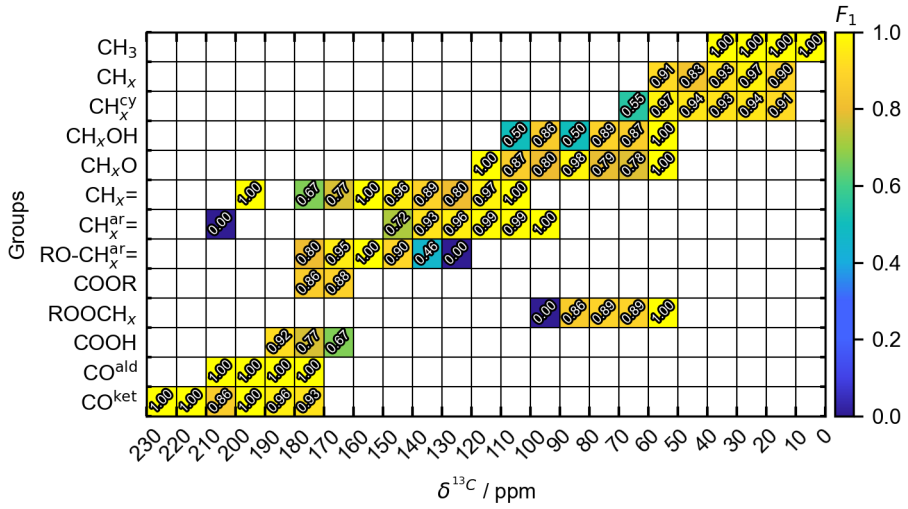

Fig. 2. F_1 test scores (specified and color-coded) of the DSM for the classification of ^{13}C signals in the respective segments of the pure-component ^{13}C spectra.

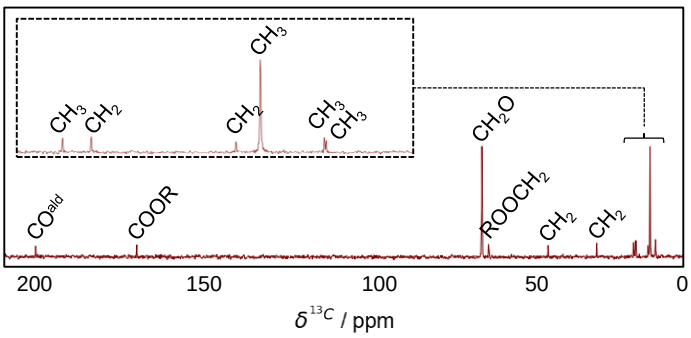

Fig. 3a. Assignment of the predicted structural groups by the DSM to the respective signals in the ^{13}C spectrum of a test mixture of diethyl ether, butanal, and butyl actetate.

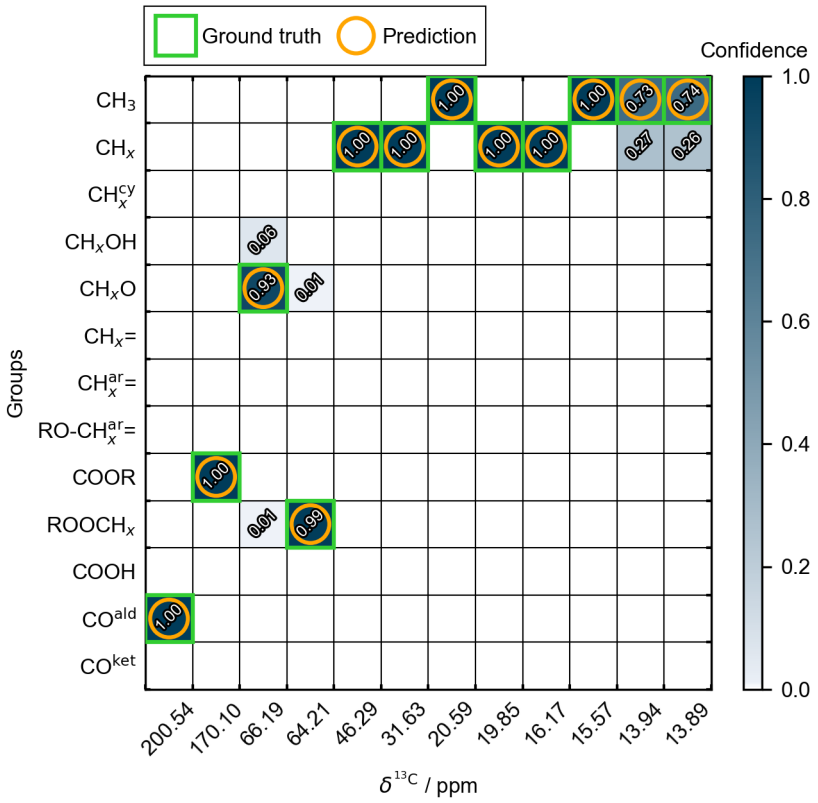

Fig. 3b. Comparison of the predicted structural groups by the DSM to the true mixture composition (ground truth) of a test mixture of diethyl ether, butanal, and butyl actetate.

4 Conclusion

This work introduces a novel fingerprinting method based on a deep set model (DSM) for automatically analyzing unknown mixtures using standard NMR experiments. By incorporating invariant and equivariant network architectures, the DSM ensures prediction results independent of input size and order permutation, making it a versatile tool for different situations. Trained on NMR spectra of pure components, augmented with synthetic spectral data, the model achieves high-performance scores on unseen test data and demonstrates excellent performance in identifying structural groups from NMR spectra of measured test mixtures.

References

1. T. Specht, K. Münnemann, H. Hasse, F. Jirasek, J. Chem. Inf. Model. 61 (2021) 143-155.
2. T. Specht, J. Arweiler, J. Stüber, K. Münnemann, H. Hasse, F. Jirasek, Magn. Reson. Chem. 62 (2024) 286-297.
3. T. Specht, K. Münnemann, H. Hasse, F. Jirasek, Phys. Chem. Chem. Phys. 25 (2023) 10288-10300.
4. T. Specht, H. Hasse, F. Jirasek, Ind. Eng. Chem. Res. 62 (2023) 10657-10667.
5. M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, A. Smola, Adv. Neural Inf. Process. Syst. 30 (2017).
6. E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. K. Wenger, H. Yao, J. L. Markley, Nucleic Acids Res. 36 (2007) D402.
7. S. Kuhn, N. E. Schlörer, Magn. Res. Chem. 53 (2015) 582.
8. RDKit: Open-Source Cheminformatics, https://www.rdkit.org (Last accessed: 04.06.2024).
9. L. Patiny, H. Musallam, A. Bolaños, M. Zasso, J. Wist, M. Karayilan, E. Ziegler, J. C. Liermann, N. E. Schlörer, Beilstein J. Org. Chem. 20 (2024) 25-31.