Supporting Information: Deep Set Model for the Automated NMR Fingerprinting of Unknown Mixtures

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

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

E-mail: fabian.jirasek@rptu.de

Phone: +49 (0)631 - 205 4685

Generation of Synthetic NMR Data

The synthetic NMR spectral data for ^{1}H and ^{13}C nuclei were generated by uploading molfiles of the relevant pure components to NMRium[1] and predicting the corresponding ^{1}H and ^{13}C NMR spectra. In cases where NMRium failed to generate a ^{1}H NMR spectrum due to prediction errors, the molfile was processed using nmrdb.org[2] to obtain the spectrum. The predicted NMR spectra were subsequently used to supplement missing chemical shifts in the experimental pure-component spectra.

Data Distribution in the ^{1}H and ^{13}C Spectrum

Figure S.1 shows the number of structural groups N_g with respective signals in different segments of the ^{1}H NMR spectrum. The segmentation of the NMR spectrum is used solely

for visualization purposes here. It is not required to apply the DSM, which can handle inputs of arbitrary sizes, i.e., containing an arbitrary number of signals.

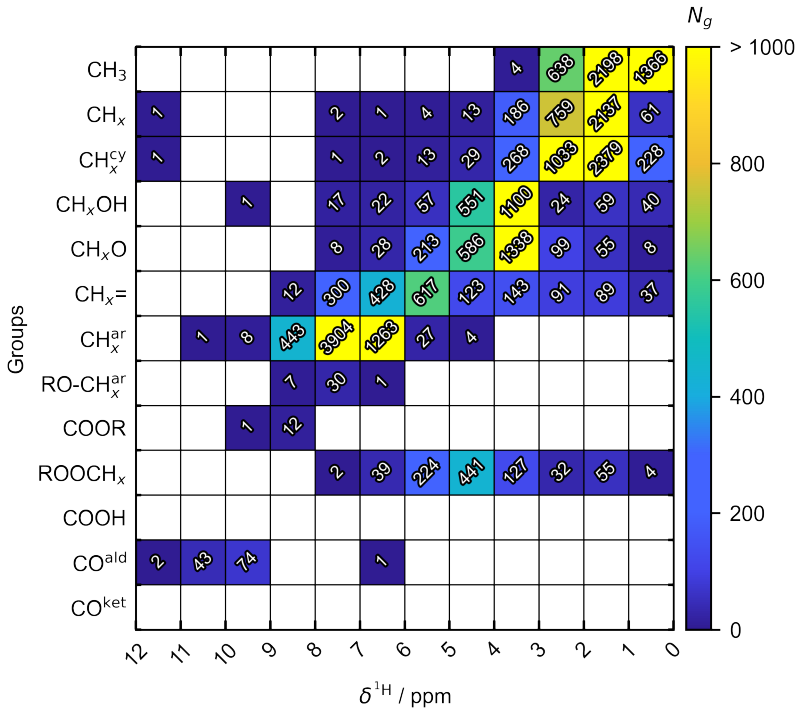

Figure S.1: Distribution of the augmented pure-component data set in the ^{1}H NMR spectrum, with specified and color-coded number of structural groups N_g. The segmentation of the ^{1}H NMR spectrum is solely for visualization purposes and no requirement for the application of the DSM.

In Figure S.2, a separate visualization of the augmented data set in terms of the proportion P_{syn} of structural groups containing synthetic data for ^{1}H or ^{13}C is shown. Figure S.2a provides the proportion P_{syn} of structural groups containing synthetic data for ^{1}H in the ^{1}H NMR spectrum, while Figure S.2b gives the proportion P_{syn} of structural groups containing synthetic data for ^{13}C in the ^{13}C NMR spectrum.

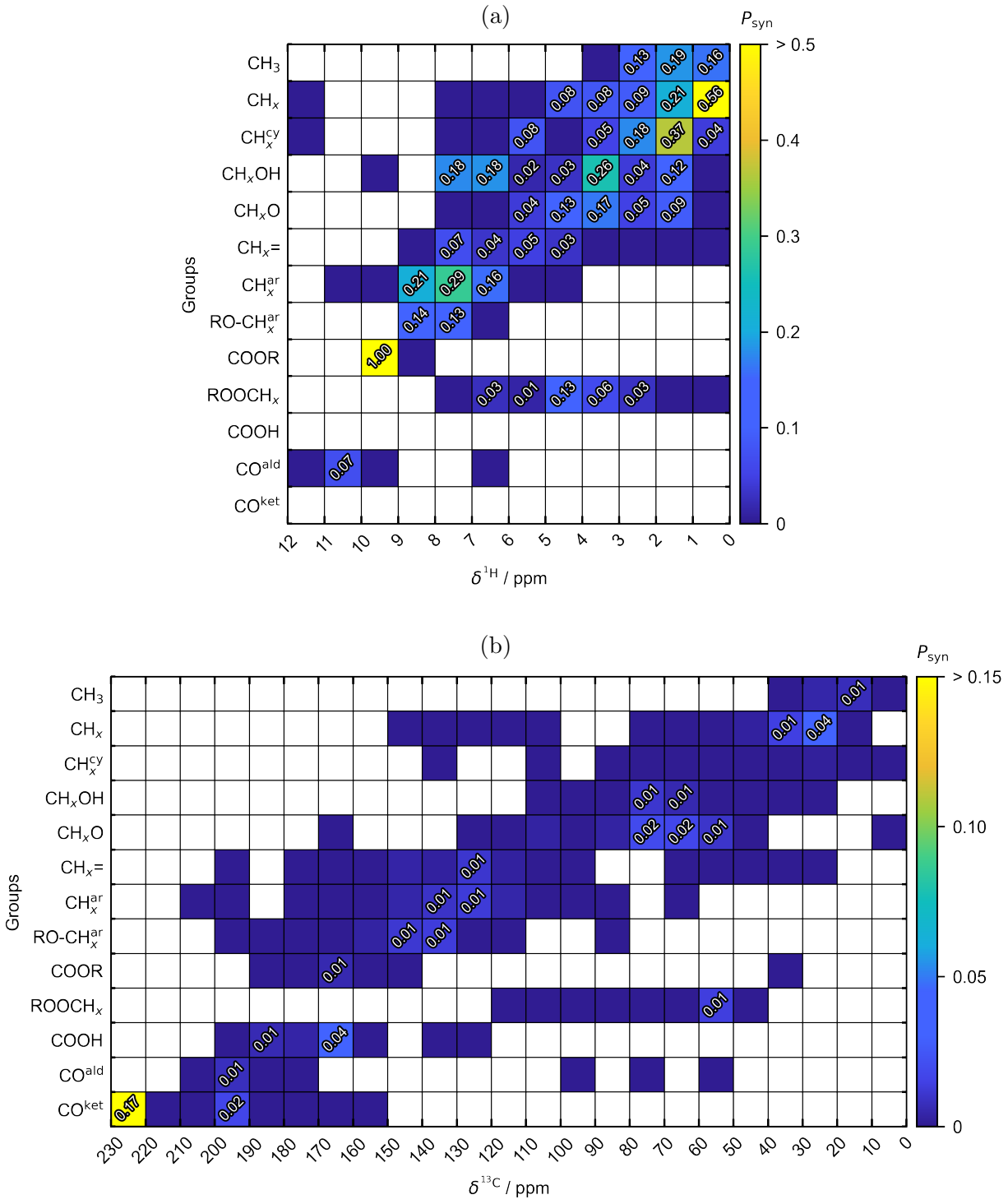

Figure S.2: Distribution of the augmented pure-component data set, with specified (greater than zero) and color-coded proportion P_{syn} of structural groups incorporating synthetic data (a) for ^{1}H in the ^{1}H NMR spectrum (b) for ^{13}C in the ^{13}C NMR spectrum.

Sensitivity Studies

Data Split Sensitivity Analysis

Table S.1 presents the overall F_1 scores achieved by the DSM on test sets generated using different random data-split seeds. The results demonstrate the robustness of the model concerning variations in data partitioning.

Table S.1: F_1 scores on different test sets generated by varying data-split seeds. All manuscript results are based on seed 1.

Seed	F_1 score
1	0.92
2	0.92
3	0.91
4	0.92
5	0.91

Influence of Synthetic NMR Data on Model Training and Performance

To evaluate the influence of synthetic NMR data on the predictive performance of the DSM, we retrained and reevaluated the model using the same procedure, hyperparameter settings, and test data as in the manuscript, but removed all components with synthetic spectral data from the training and validation sets. As a consequence, the number of pure components available for training and validation was reduced by 40.97%. The average F_1 score across all structural groups in the test set decreased from 0.92 (as reported in the manuscript) to 0.89 when excluding the synthetic NMR data. Given the substantial fraction of excluded data, this comparatively small decrease in predictive performance demonstrates the robustness of the DSM with respect to the composition of the training data.

Figure S.3 compares the $F_{1,g}$ scores on the test set plotted against the proportion of structural groups containing synthetic data, $P_{\mathrm{syn},g}$, obtained using the full augmented dataset

($P_{\mathrm{syn},g} > 0$, cf. manuscript) and when excluding components with synthetic NMR data during model training and validation ($P_{\mathrm{syn},g} = 0$) for each structural group g. For the majority of structural groups, the differences in $F_{1,g}$ scores between the two training modes are marginal and lie within the range expected when training separate models with different training data compositions. Importantly, no structural group exhibits a significantly lower $F_{1,g}$ score when synthetic NMR data are excluded. For the structural groups CH_x, ROOCH_x, and $\mathrm{CO}^{\mathrm{ald}}$, a clear improvement in $F_{1,g}$ score is observed when synthetic NMR data are included. Consequently, the inclusion of synthetic NMR data does not introduce systematic biases or degrade predictive accuracy, but instead selectively enhances performance for some structural groups.

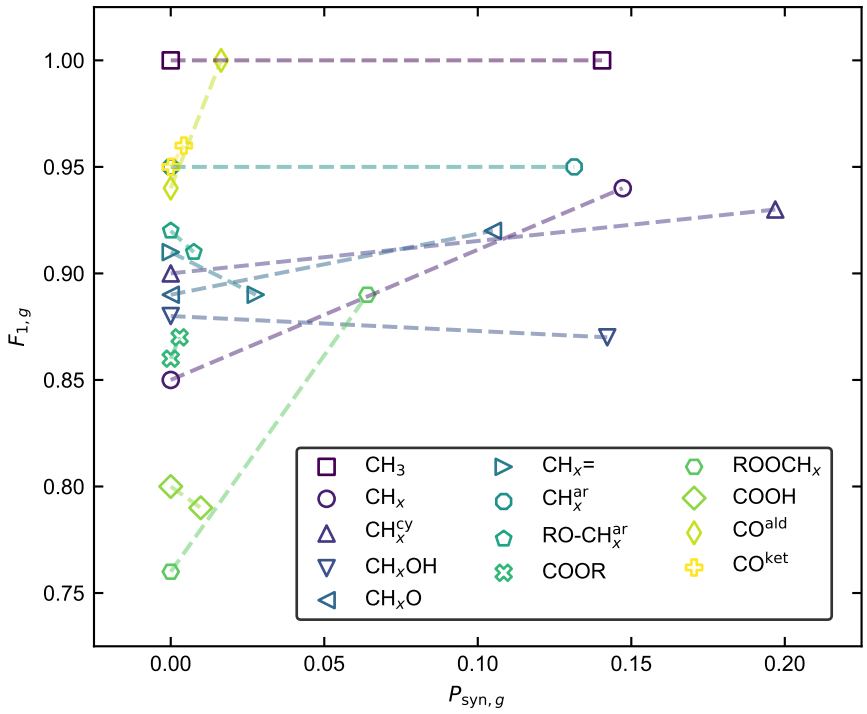

Figure S.3: Comparison of the $F_{1,g}$ scores on the test set for each structural group g, plotted against the proportion of structural groups containing synthetic data, $P_{\mathrm{syn},g}$. Results are shown for the DSM trained with all components containing synthetic NMR data excluded from the data set($P_{\mathrm{syn},g} = 0$) and for the DSM trained using the full data set ($P_{\mathrm{syn},g} > 0$, cf. manuscript). Lines are shown to guide the eye.

Hyperparameter Sensitivity Analysis

Table S.2 presents the results for the sensitivity analysis of the DSM regarding the optimized hyperparameters in terms of the validation loss. In these experiments, the relative dimension of the nuclei-specific networks $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$ were maintained to the dimensions of the networks ϕ and ρ. We defined the hyperparameter settings for the configuration of layers and nodes as follows:

- Layer configuration:
  - Small: 2 layers for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 1 layer for ϕ and ρ.
  - Medium: 3 layers for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 2 layers for ϕ and ρ.
  - Large: 4 layers for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 3 layers for ϕ and ρ.
- Node configuration:
  - Small: 4 nodes for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 128 nodes for ϕ and ρ.
  - Medium: 8 nodes for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 256 nodes for ϕ and ρ.
  - Large 16 nodes for $\phi_{^1\mathrm{H}}$ and $\phi_{^{13}\mathrm{C}}$; 512 nodes for ϕ and ρ.

Table S.2: Results of the hyperparameter sensitivity analysis of the DSM in terms of the validation loss. λ is the weight decay of the Adam optimizer. Model 1 was used throughout the manuscript.

Model No.	λ	Initial learning rate	Batch size	Layer configuration	Node configuration	Validation loss
1	0.0005	0.0001	1	medium	medium	0.03381
2	0.0005	0.0001	1	medium	small	0.04783
3	0.0005	0.0001	1	medium	large	0.04294
4	0.0005	0.0001	1	small	medium	0.05267
5	0.0005	0.0001	1	large	medium	0.03887
6	0.0005	0.0001	4	medium	medium	0.04202
7	0.0005	0.0001	8	medium	medium	0.04352
8	0.0005	0.001	1	medium	medium	0.04773
9	0.0005	0.00001	1	medium	medium	0.04954
10	0.001	0.0001	1	medium	medium	0.04225
11	0.0001	0.0001	1	medium	medium	0.04429

Results of the Support Vector Classification

To compare the performance of the DSM with our previous approach, we retrained the support vector classification (SVC) model from our earlier work[3] using the identical data split as that employed for the DSM and choosing the SVC hyperparameters as in the original work. Figure S.4 presents the predictive performance of the SVC on the test set, measured in terms of the $F_{1,g}$ scores. Because the SVC only processes ^{13}C NMR signals up to 210 ppm, the six CO^{ket} groups in the test set with chemical shifts δ^{13C} above this limit were reassigned to the 200 - 210 ppm segment. The SVC achieved an average F_1 score of 0.85 across all structural groups in the test set.

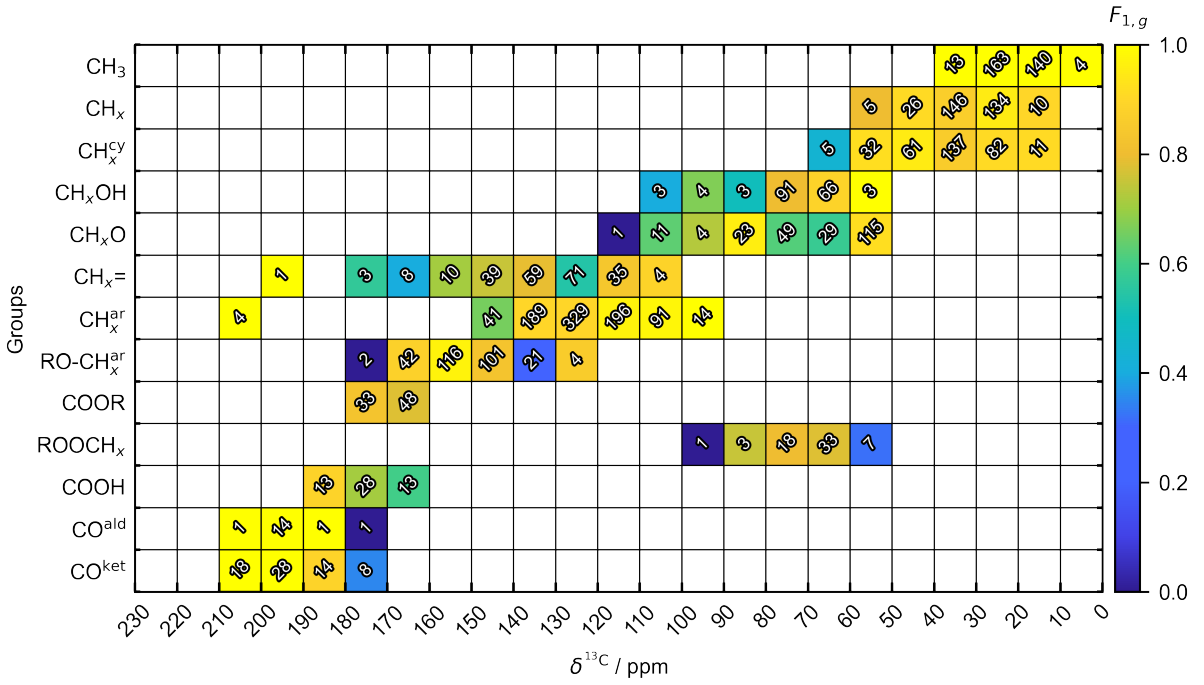

Figure S.4: $F_{1,g}$ scores (indicated by color code) of the SVC from our previous work[3] for predicting the structural groups for the pure components in the test set used in this work. The numbers in the cells indicate the number of structural groups N_g per segment of the ^{13}C NMR spectrum.

Experimental Methods

Chemicals

Table S.3 gives an overview of the chemicals used in preparing the test mixtures in this work. Deionized and purified water was produced with an ultrapure water system (Omnia series, stakpure).

Table S.3: Suppliers and indicated purities of chemicals used in this work.

Chemical	Formula	Supplier	Purity %
Acetone	C_3H_6O	Merck	$\geq$ 99.90
Tartaric acid	$C_4H_6O_6$	Merck	$\geq$ 99.50
1,4-Butanediol	$C_4H_{10}O_2$	Sigma Aldrich	$\geq$ 99.00
Diethyl ether	$C_4H_{10}O$	Merck	$\geq$ 99.70
Butanal	C_4H_8O	Sigma Aldrich	$\geq$ 99.50
Butyl acetate	$C_6H_{12}O_2$	Sigma Aldrich	$\geq$ 99.50
Cyclohexane	C_6H_{12}	Fisher Chemicals	$\geq$ 99.98
1-Hexene	C_6H_{12}	Thermo Scientific	$\geq$ 99.00
Diglyme	$C_6H_{14}O_3$	TCI	$\geq$ 99.00
Anisole	C_7H_8O	Sigma Aldrich	$\geq$ 99.70
1-Octanol	$C_8H_{18}O$	Merck	$\geq$ 99.00
3-Methylbutan-2-one	$C_5H_{10}O$	TCI	$\geq$ 99.00
3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid sodium salt	$C_6H_9D_4O_2SiNa$	Sigma Aldrich	$\geq$ 98.00
Tetramethylsilane	$C_4H_{12}Si$	Sigma Aldrich	$\geq$ 99.00

Sample Preparation and Acquisition of NMR Spectra

Samples of test mixtures ($<$ 20 g) were prepared gravimetrically in glass vessels using a balance (Mettler Toledo) with an accuracy of $\pm$0.001 g. As reference component in ^{1}H and ^{13}C NMR spectroscopy, a small amount of 3-(Trimethylsilyl)propionic-2,2,3,3-d4 acid was added to Mixture I, whereas Tetramethylsilane was used for all other test mixtures. After the preparation, approximately 1 ml of each sample was transferred to a 5 mm NMR tube for the NMR analysis.

All NMR spectra were recorded on a Spinsolve 60 MHz Ultra NMR spectrometer from Magritek (Aachen, Germany) at 299.65 K. ^{1}H NMR spectra were recorded with a flip angle

of 90°, an acquisition time of 6.4 s, and a single scan. ^{1}H decoupled ^{13}C NMR spectra were recorded with a flip angle of 90°, an acquisition time of 3.2 s, a relaxation delay of 60 s, and 128 scans. ^{13}C DEPT NMR spectra were recorded with pulse angles of 90° and 135°, an acquisition time of 3.2 s, a relaxation delay of 30 s, and 64 scans. ^{1}H $-^{13}$ C HSQC NMR spectra were recorded with 128 steps, a repetition time of 5 s, and 64 scans.

Processing of the obtained NMR spectra was performed using MNova. For all spectra, automatic baseline and phase corrections were applied, along with exponential line broadening (0.3 Hz for ^{1}H and 1.0 Hz for ^{13}C and ^{1}H $-^{13}$ C HSQC). Automated peak picking was followed by manual peak integration and intensity determination. The determination of substitution degrees from ^{13}C DEPT NMR spectra and the detection of labile protons from ^{1}H $-^{13}$ C HSQC NMR spectra were conducted as described in previous work.[3]

Regression Model for Chemical Shift Data

To overcome the challenges in the evaluation of the ^{1}H$-^{13}$C HSQC NMR spectra, specifically of non-determinable chemical shifts $\delta_i^{^{1}\mathrm{H}}$ of ^{1}H nuclei bonded to specific ^{13}C nuclei posed by signal overlap in ^{1}H $-^{13}$ C HSQC NMR spectra, we developed a linear regression model to relate the chemical shifts $\delta_i^{^{13}\mathrm{C}}$ of ^{13}C nuclei to those of their directly bonded ^{1}H nuclei $\delta_i^{^{3}\mathrm{H}}$. The model is defined as

$$\delta_i^{^{1}\mathrm{H}} = a\,\delta_i^{^{13}\mathrm{C}} + b \tag{S.1}$$

with the regression parameters a and b determined from our data set being $a = 0.052$ ppm and $b = 0.239$. The regression analysis was performed using the linregress function from the SciPy[4] library.

Figure S.5 shows the relationship of the chemical shifts of ^{13}C nuclei and their bonded ^{1}H nuclei, along with the fitted regression model. Despite the weak correlation, we observed in preliminary studies to this work that leveraging even this modest relationship rather than omitting $\delta_i^{^{1}\mathrm{H}}$ or replacing it with a fixed value significantly improves the predictive

performance of the DSM. Consequently, in cases where the $^{1}\mathrm{H}-^{13}\mathrm{C}$ HSQC signals could not be observed, even though the presence of a $^{1}\mathrm{H}-^{13}\mathrm{C}$ bond was confirmed by the substitution degree determined from ^{13}C DEPT NMR, the regression model was used to estimate the missing chemical shift $\delta_i^{^{1}\mathrm{H}}$ from the corresponding $\delta_i^{^{13}\mathrm{C}}$.

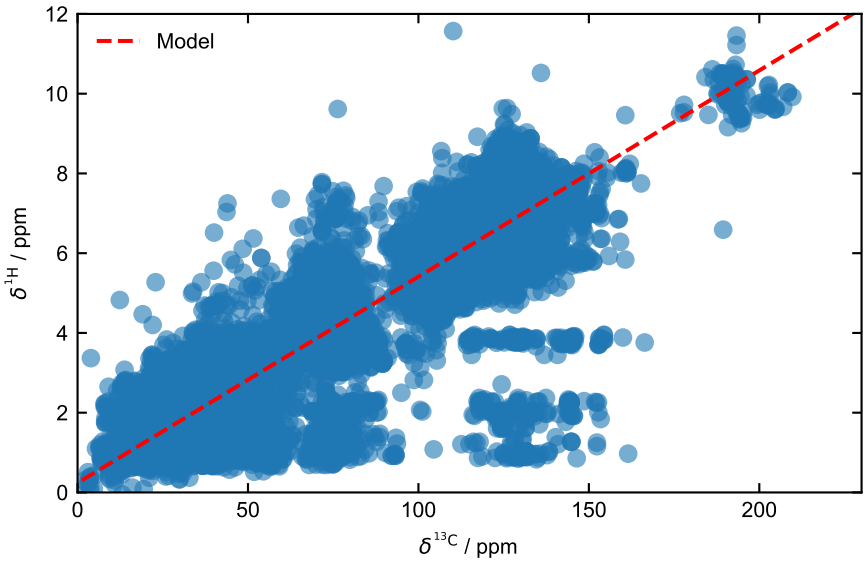

Figure S.5: Scatter plot of the chemical shifts $\delta_i^{^{13}\mathrm{C}}$ of ^{13}C nuclei versus the corresponding chemical shifts $\delta_i^{^{1}\mathrm{H}}$ of directly bonded ^{1}H nuclei, along with the linear regression model from Eq. S.1.

References

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

(2) Banfi, D.; Patiny, L. www.nmrdb.org: Resurrecting and Processing NMR Spectra Online. *CHIMIA* **2008**, *62*, 280.

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

(4) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.