

Automatic Signal Region Detection in ^1H NMR Spectra Using Deep Learning

Federico Paruzzo,¹ Simon Bruderer,¹ Youssef Janjar,² Bjoern Heitmann,¹ and Christine Bolliger.¹

¹Bruker Switzerland AG, 8117 Fällanden, Switzerland.

²École polytechnique fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.



Signal region detection is a routine task in NMR spectroscopy, however, it is one of the few processing steps that is typically performed manually. This is due to the fact that the method implemented in TopSpin to carry out this task automatically often requires fine-tuning of multiple parameters to obtain satisfactory results. This limitation typically prevents the effective integration of multiple, diverse spectra in an automatic way, and favours the use of the manual process. Recently, deep learning techniques have emerged as a powerful tool for recognition and segmentation tasks, proving to be able to reach and surpass human-level performance in a fraction of the time. These techniques could potentially lead to a fully automatic extraction and analysis of the information contained in NMR spectra. Here, we make the first step toward this goal. We introduce *sigreg*, a deep learning algorithm for signal region detection in 1D ^1H NMR spectra. We show that this method is robust and provides better performance than current Bruker solutions, achieving excellent accuracy for spectra obtained in a wide range of base frequencies without requiring any inputs from the user side.

From the early days of nuclear magnetic resonance (NMR) it has been known that NMR signals yield quantitative information if the right experimental conditions are met. [1] [2] [3] Relative concentrations of magnetically distinct nuclei are obtained by integrating the areas under the signals in an NMR spectrum, and absolute concentrations can be calculated by using a standard with a known concentration. This is an important property of NMR, which has contributed to the widespread use that NMR has achieved over the years, and is today at the base of many NMR applications such as qNMR, structure verification, and structure elucidation.

All these processes rely on the accurate determination of the limits of signal regions to integrate, which can be manually defined directly at the spectrometer using the “manual-integration” module available in TopSpin. This module is easy to use and straightforward enough to even allow beginners to obtain integral values. However, the integration of multiple datasets or of complex spectra using this method can be time-consuming. An alternative is the automatic integration of NMR spectra, which can be done in TopSpin using the *int auto* command. However, using this method, the quality of the final result strongly depends on several parameters that need to be tuned by the user. This often prevents users from automatically integrating spectra that have different features using the same set of parameters, and the time required to optimize a new set encourages the use of manual integration.

Over the past year, deep learning has shown to be able to achieve excellent results in recognition and segmentation tasks, and, in particular, in the field of NMR spectroscopy it has been used to support users with spectra processing and interpretation. [4] [5] With applications such as peak picking in 2D NMR spectra [6] or structure verification, [7] deep learning has proven to have the potential to obtain human-like results in the processing and analysis of NMR spectra. Ideally, deep learning could be implemented in the full NMR workflow, from the acquisition of NMR data to the extraction of the results. At Bruker we are committed to bringing this innovation to our customers.

Here, we introduce *sigreg*, the result of our initial steps in this field. *sigreg* is the first deep-learning-based command available in TopSpin, and it performs fully automatic, parameter-free, signal region detection in 1D ^1H NMR spectra. This is also one of the first deep learning applications that focus on signal region detection in NMR spectra. We show that *sigreg* provides significantly better results than *int auto* for ^1H NMR spectra acquired in a wide range of base frequencies (from 80 to 800 MHz), without requiring the intervention of the user. We test the limits of detection using simple artificial spectra, and we show examples of the performance of this method on experimental datasets with different features (e.g. with high noise, phase distortion, broad peaks, and large solvent signals).

Methods

To develop *sigreg* we have used supervised machine learning. The deep learning algorithm was built from scratch, and it is a fully convolutional neural network inspired by the U-net. [5] To train and validate the algorithm we have used 500,000 artificial spectra generated using a home-built code, carefully developed to provide realistically looking ^1H NMR spectra. Spectra were generated using base frequencies ranging from 80 to 800 MHz. The signal regions were labelled during spectra generation and consist only of the main peaks, leaving the ^{13}C satellites unlabelled. The signal intensity was normalized in each spectrum by setting the intensity of the highest peak equal to 1.

The limits of detection were tested using artificial 400 MHz spectra with one signal (singlet), by evaluating the detection of the signal regions as a function of the signal and noise intensities. The spectra were described by 16,384 points and have a width of 15.5 ppm. The singlet was placed in the middle of the spectrum, and it has a full width at half height of 2 Hz (Gaussian broadening). We have generated the reference spectrum and evaluated the number of points to be labelled as signal ($Area_{true}$). Then, we have varied the signal intensity while keeping the noise constant (and vice versa), and evaluated the number of points detected by *sigreg* as signal ($Area_{sigreg}$). Each spectrum (e.g. each signal-noise combination) was created 100 times, and the ratio $Area_{sigreg}/Area_{true}$ was calculated using the mean $Area_{sigreg}$ for all 100 experiments.

The model performance was tested using 100 experimental ^1H NMR spectra labelled by Bruker NMR experts. The spectra were labelled by several experts, and without giving any pre-set rule (e.g. each expert might have its own labelling style). They should thus represent, on average, the typical result obtained by NMR users. 25 out of 100 spectra were acquired at 80 MHz base frequency, using the Bruker Fourier 80 bench-top NMR instrument. The other spectra were acquired at base frequencies ranging from 300 to 800 MHz, with the majority being acquired at 400 MHz. For these spectra, we have extracted the results obtained using *sigreg* and *int auto* using

TopSpin version 4.0.9. For *int auto*, the standards parameters were used for all spectra. For *sigreg*, the signal intensity was normalized by setting the intensity of the highest peak in each spectrum equal to 1 before signal region detection.

To confirm our finding, we have also tested the performance against 10,000 artificially generated ^1H NMR spectra. In this case, we only have tested *sigreg* and not *int auto*.

To measure the accuracy of the model we have used the F_1 score. [6] This score is based both on precision (how many points in the spectrum detected as signal regions detected are true signal regions) and on recall (how many points of the true signal regions are detected by the model), and it is defined as:

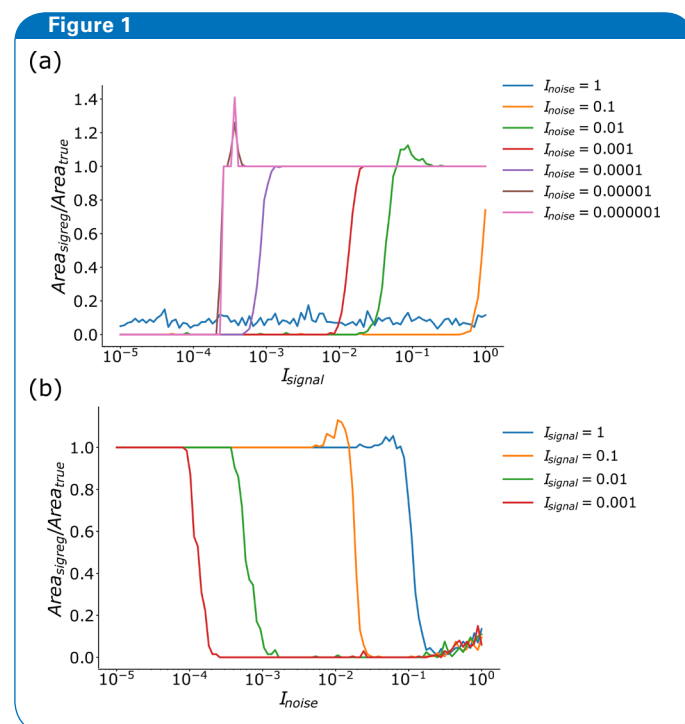
$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

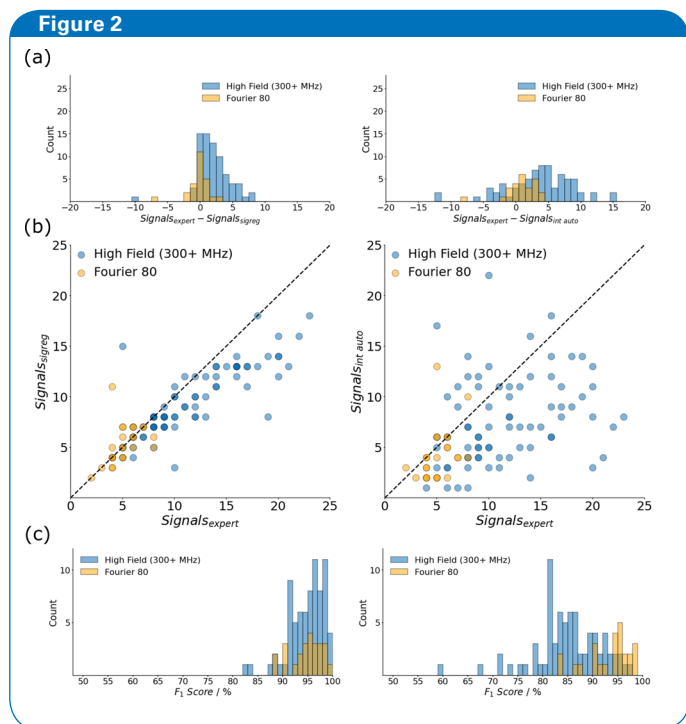
and

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

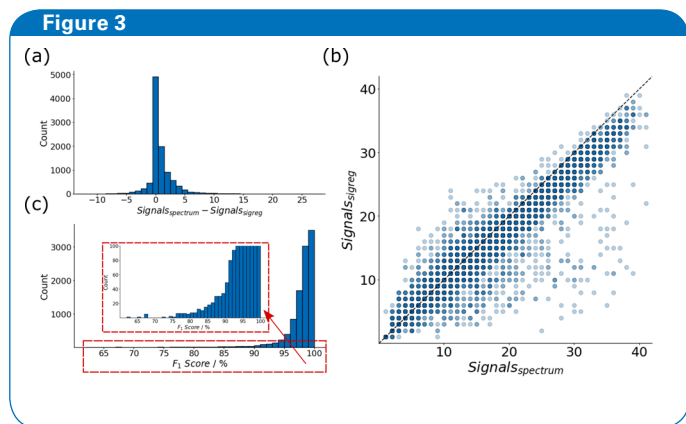


Limits of detection of *sigreg* as a function of signal intensities (a) and noise intensities (b). Values of $Area_{sigreg}/Area_{true}$ close to 1 indicate that *sigreg* fully detects the signals, while values close to 0 indicate that the signal (or at least part of it) is not labelled as signal region.

Results and Discussion



Performance of *sigreg* and *int auto* on 100 experimental ^1H NMR spectra. (a) Histograms showing the difference between the number of signals found by the experts and the ones detected with *sigreg* (left) and *int auto* (right). (b) Scatter plots that shows the correlation between the number of signals found by the experts and the ones detected with *sigreg* (left) and *int auto* (right). The black dashed lines represent a perfect correlation. (c) F_1 scores achieved with *sigreg* (left) and *int auto* (right).



Performance of *sigreg* on 10,000 artificially generated ^1H NMR spectra. (a) Histogram showing the difference between the number of signals in the spectra and the ones detected with *sigreg*. (b) Scatter plot that shows the correlation between the number of signals in the spectra and the ones detected with *sigreg*. The black dashed line represents a perfect correlation. (c) F_1 scores achieved with *sigreg*.

Figure 1 shows the limits of detection of *sigreg* tested using the single-peak spectra described in the methods section. In **Figure 1a**, the fraction of the signal detected by *sigreg* as signal is plotted as a function of the signal intensity for several noise values. This test shows for a given noise level, how intense should the signal be to be detected by *sigreg*. For high noise values ($I_{noise} = 1$), the area detected by *sigreg* is much smaller than $Area_{true}$, showing that we do not correctly detect the signal in those spectra. This is expected, considering that with the highest signal intensity ($I_{signal} = 1$), the noise and the signal are of the same order of magnitude. With lower noise values, *sigreg* starts to correctly detect the signal. Overall, we observe that, in these test-spectra, *sigreg* finds signals that have intensities at least one order of magnitude larger than the noise value. This is true up to a noise intensity equal to 10^{-5} , below which the limit of detection seems to be due to the signal intensity rather than the SNR. The lowest signal intensity detected in this model spectra was about $2 \cdot 10^{-4}$ relative to the highest peak in the spectrum, suggesting that this is the limit of detection of the current *sigreg* version.

Figure 1b shows the area detected by *sigreg* as a function of the noise level, for several signal intensities. We can again see that we are able to successfully detect signals that are about 1 order of magnitude larger than the noise level, with intensities as low as 0.001. Thus, this test suggests that *sigreg* should not detect signals that are less than 10 times stronger than the noise level. In these test-spectra, signals with higher intensities are detected by *sigreg*, even when these signals are 1000 times less intense than the highest peak in the spectrum. Note that these tests were made on simple artificial spectra, and that in more complex spectra (e.g. composed by an higher number of signal) these numbers might vary.

While this test is a useful indication of the detection limits of the current *sigreg* version, it does not test how the model performs on experimental results. To do so, we have compared the *sigreg* results with the signal regions selected by Bruker NMR experts of 100 ^1H 1D NMR experimental spectra (**Figure 2**). The same comparison was also carried out for *int auto*, to assess the performance of the previously existing methods to detect signal regions.

Figure 2a and **b** compare *sigreg*, *int auto* and expert results in terms of number of signals selected in each spectrum. From the histograms of **Figure 2a** we immediately see that in most of the experimental spectra the number of signals detected using *sigreg* agrees with the one found by the experts, both for the 80 MHz spectra (in orange in the figure) and the high field NMR spectra (300+ MHz). Using *int auto*, instead, the agreement between the number of manually and automatically detected signals is much lower. The average difference between the absolute number of signals found by the experts

and the ones detected automatically is 2.1 for *sigreg* (1 for the 80 MHz and 2.5 for the 300+ MHz) and 4.7 for *int auto* (1.9 for the 80 MHz and 5.6 for the 300+ MHz). The spreading of the *int auto* results is also visibly much higher than the *sigreg* one.

The scatter plots in **Figure 2b** show the correlation of the number of signals found by the experts and the one detected automatically. Again, *sigreg* shows a better correlation compared to *int auto*.

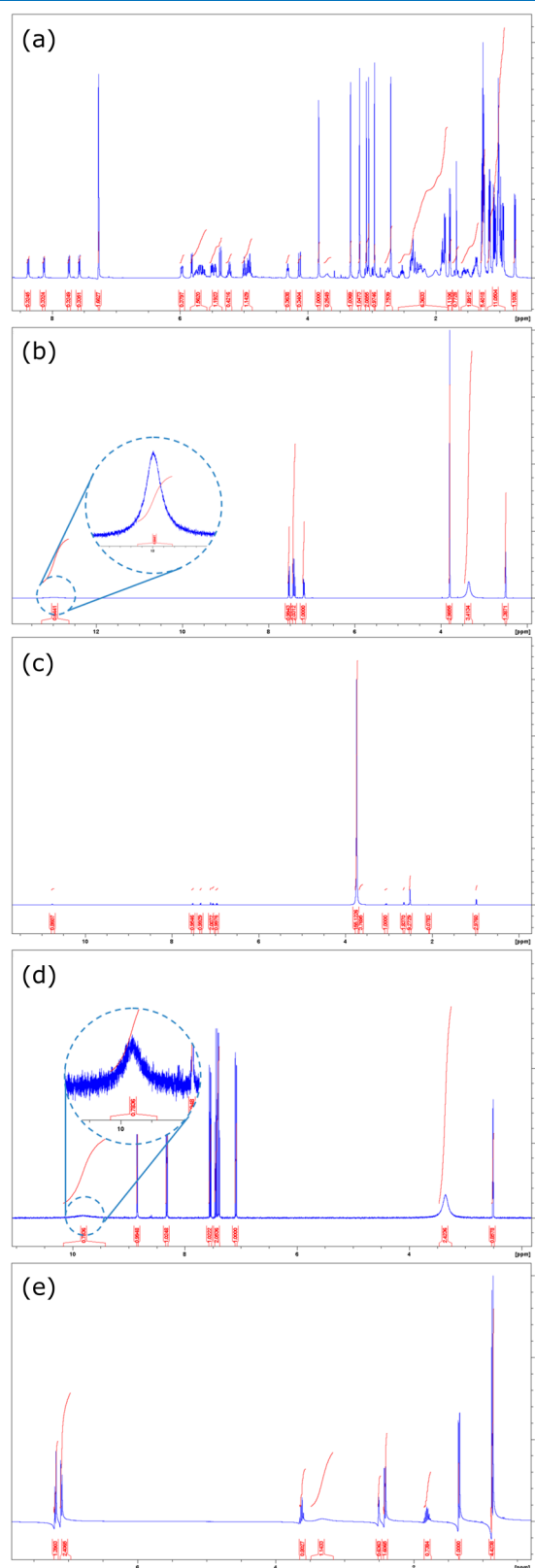
While the number of signals detected by the algorithm is a good indication of how the algorithm works and how it performs compared to NMR users, it does not give us information on how many signals are correctly detected. The algorithms and the users, indeed, might easily provide a different number of signals by splitting a signal region or grouping multiple regions together. Thus, to estimate the accuracy of these models we have calculated the F_1 score between the expert and the automatically detected signal regions.

Figure 2c shows the F_1 scores obtained for the 100 experimental ^1H 1D NMR spectra analyzed here. With an average F_1 score of 94.8% (the average of the 80 MHz and 300+ MHz are both 94.8%) *sigreg* shows an overall higher accuracy and lower field dependence than *int auto*, which has an average F_1 score of 87.1 (93.4% for 80 MHz and 85.0% for 300+ MHz).

The accuracy of *sigreg* was confirmed also on 10,000 artificially generated ^1H 1D NMR spectra, as shown in **Figure 3**. The histogram of **Figure 3a** and the scatterplot of **Figure 3b** show good agreement between the number of signals in the spectra and the ones detected by *sigreg*. The average difference of the number of signals detected is 0.9 (with nearly 5000 spectra with the correct number of signals detected), the R^2 value of the correlation is 0.91, and the average F_1 score is 97.8%. Most of the outliers are spectra where the solvent signal is more than 1000 times more intense than the other signals.

To conclude, we present in **Figure 4** the results obtained by integrating the signal regions of a few ^1H 1D NMR experimental spectra using *sigreg*. **Figure 4a** shows the result obtained on a relatively crowded NMR spectrum, the spectrum of cyclosporin which shows more than 30 signal regions. **Figures 4b-4e** show the results obtained on spectra characterized by having broad peaks (**Figure 4b**), low SNR (**Figure 4c**), high intensity of the solvent signal (**Figure 4d**), and phase distortion (**Figure 4e**). Note that despite the good results achieved with the phase distorted spectrum of **Figure 4e**, *sigreg* was not developed to work on phase distorted spectra, and we suggest to apply phase correction before using this command.

Figure 4



Examples of experimental results obtained with *sigreg* for (a) a spectrum with a relatively high number of signals, (b) a spectrum with broad peaks, (c) a spectrum with high solvent intensity, (d) a spectrum with low signal-to-noise ratio, and (e) a phase distorted spectrum

Conclusions

In summary, we have presented here a new method for signal region detection in ^1H 1D NMR spectra. This method uses deep learning to detect signal regions and is available starting from TopSpin v. 4.0.9. We have shown that this method offers great advantages compared to the previously existing method (*int auto*), such as having higher accuracy and being parameter-free.

Overall, the performance of *sigreg* is good. We have shown on 100 experimental ^1H NMR spectra that the signal regions detected with this method agree well with the ones detected by NMR experts (the F_1 score is 94.8 %). This algorithm has shown to be quite robust, and to perform well on spectra obtained in a wide range of base frequencies (ranging from 80 to 800 MHz), and on spectra with different features (e.g. spectra that have many resonances, broad peaks, low SNR, high intensity of the solvent signal, or phase distortion). While this is only the first release of this algorithm and this method will be developed even further, *sigreg* is already robust enough to be used routinely to integrate phase and baseline corrected NMR spectra and to be the method of choice for the automatic integration of 1D ^1H NMR spectra.

Remarks

Note that *sigreg* can be easily implemented in au programs using the macro 'SIGREG' for a full automatization of your NMR routines.

References

1. R. B. Williams, «Quantitative Determination of Organic Structures by Nuclear Magnetic Resonance: Intensity Measurements,» *Annals of the New York Academy of Sciences*, Bd. 70, Nr. 4, pp. 890-899, 1958.
2. J. L. Jungnickel und J. W. Forbes, «Quantitative Measurement of Hydrogen Types by Intergrated Nuclear Magnetic Resonance Intensities,» *Analytical Chemistry*, Bd. 35, Nr. 8, pp. 938-942, 1 7 1963.
3. D. P. Hollis, «Quantitative Analysis of Aspirin, Phenacetin, and Caffeine Mixtures by Nuclear Magnetic Resonance Spectrometry,» *Analytical Chemistry*, Bd. 35, Nr. 11, pp. 1682-1684, 1963.
4. D. Chen, Z. Wang, D. Guo, V. Orekhov und X. Qu, «Review and Prospect: Deep Learning in Nuclear Magnetic Resonance Spectroscopy,» 13 1 2020.
5. C. Cobas, «NMR signal processing, prediction, and structure verification with machine learning techniques,» *Magnetic Resonance in Chemistry: MRC*, 2020.
6. P. Klukowski, M. Augoff, M. Zięba, M. Drwal, A. Gonczarek und M. J. Walczak, «NMRNet: a deep learning approach to automated peak picking of protein NMR spectra,» *Bioinformatics*, Bd. 34, Nr. 15, pp. 2590-2597, 2018.
7. M. M. Zanardi und A. M. Sarotti, «GIAO C–H COSY simulations merged with artificial neural networks pattern recognition analysis. Pushing the structural validation a step forward,» *The Journal of Organic Chemistry*, Bd. 80, Nr. 19, pp. 9371-9378, 2015.
8. O. Ronneberger, P. Fischer und T. Brox, «U-net: Convolutional networks for biomedical image segmentation,» in *International Conference on Medical image computing and computer-assisted intervention*, 2015.