A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition

GDCh

Chemie

www.angewandte.org

Accepted Article

Title: Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning

Authors: Xiaobo Qu, Yihui Huang, Hengfa Lu, Tianyu Qiu, Di Guo, Tatiana Agback, Vladislav Orekhov, and Zhong Chen

To be cited as: *Angew. Chem. Int. Ed.* 10.1002/anie.201908162
Angew. Chem. 10.1002/ange.201908162

Link to VoR: http://dx.doi.org/10.1002/anie.201908162
http://dx.doi.org/10.1002/ange.201908162

WILEY-VCH

10.1002/anie.201908162

WILEY-VCH

COMMUNICATION

Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning

Xiaobo Qu*[a], Yihui Huang[a], Hengfa Lu[a], Tianyu Qiu[a], Di Guo[b], Tatiana Agback[c], Vladislav Orekhov[d], Zhong Chen*[a]

Abstract: Nuclear magnetic resonance (NMR) spectroscopy serves as an indispensable tool in chemistry and biology but often suffers from long experimental time. We present a proof-of-concept of application of deep learning and neural network for high-quality, reliable, and very fast NMR spectra reconstruction from limited experimental data. We show that the neural network training can be achieved using solely synthetic NMR signal, which lifts the prohibiting demand for a large volume of realistic training data usually required in the deep learning approach.

Nuclear magnetic resonance (NMR) spectroscopy is an invaluable biophysical tool in modern chemistry and life sciences. However, duration of NMR experiments increases rapidly with spectral resolution and dimensionality[1], which often imposes unbearable limitations due to low sample stability and/or excessive costs of NMR measurement time. To accelerate the data acquisition and optimize sensitivity, modern NMR experiments are often acquired using the Non-Uniform Sampling (NUS) approach [1], where only a small fraction of traditional NMR measurements, usually called free induction decay (FID), is performed and, thus, only a fraction of measurement time is spent.

Over the past two decades, several methods have been established in the NMR field to reconstruct high quality spectra from NUS data. In all cases, a prior knowledge or assumption are incorporated in order to compensate for missing information introduced by the NUS scheme. Examples include the maximum entropy[1b], spectrum sparsity in compressed sensing[2], spectral line-shape estimation in SMILE[3], tensor structures in MDD[1a] or Hankel tensors[4], and exponential nature of NMR signal in low rank[4-5]. Although spectra are reconstructed well with these approaches, a number of important practical limitations and conceptual question remain. Thus, despite of varying implementations, algorithms of all these methods are iterative and require lengthy calculations and/or use of super-computers. Pros and cons of applying different prior assumptions are not well understood and combination of the best features, while avoiding the negative sides of different approaches is problematic.

Deep learning (DL) is a representative artificial intelligence technique that uses neural networks. DL has been successfully demonstrated in computer vision[6], medical imaging[7] and biological data analysis[8]. Data analysis by a trained neural network is fast. Furthermore, in contrast to the traditional methods that relies on a prior knowledge or formal assumptions, for instance, sparsity or maximum entropy, the neural network retrieves the essential features embedded in training data and thus does not require any predefined formal priors. In this work, we explore the DL for fast and high-quality reconstruction of NMR spectra from non-uniformly sampled data.

A critical challenge of the DL is that it requires an enormous amount of realistic data at the training stage. Whilst obtaining of such a gigantic data set is practically impossible due to NMR sample and instrument time limitations, our work demonstrates that successful training of the neural network in the DL is possible using solely synthetic data. These are generated using the classic assumption that NMR signal is a superposition of small number of exponential functions[1b, 4-5]. The strategy of using synthetic data for training is beyond the traditional DL approach that usually requires huge volume of practical experimental data. This work exemplifies bridging of the traditional signal modelling to DL that enables smart artificial intelligence computational tools in applications that lack enough practical data to train the neural network. This work can be treated as a proof-of-concept for DL NMR spectroscopy.

Reconstructing a spectrum from NUS data is equivalent to mapping of the input undersampled FID signal to the target spectrum. In the DL NMR, a neural network is trained to perform the mapping as shown in Figure 1. First, the spectrum artifacts introduced by NUS are removed with dense convolutional neural network (CNN)[9] and then intermediately reconstructed spectra are further refined to maintain the data consistency to the sampled signal. Artifacts are gradually removed as the stage of reconstruction increases and the final spectrum is produced after several stages. In our implementation, dense CNN is chosen because it ensures maximum information flow between layers in the neural network[9] while data consistency constraint the reconstruction to the sampled data points[7].

The key step for DL NMR is to learn the mapping. We simulate the fully-sampled time domain NMR signal, from which undersampled NUS signal was obtained using Poisson gap sampling scheme (See Supplement S1.1 for more details). Given the synthetic NUS signal $\boldsymbol{y}$ and the corresponding target spectrum $\boldsymbol{s}$ produced from the fully sampled time domain data, a large number of pairs $(\boldsymbol{y}_k, \boldsymbol{s}_k)$ (k=1, 2, …, K) are fed into the neural network to learn the best network parameters $\boldsymbol{\theta}$ that minimizes the least errors $e(\boldsymbol{\theta}) = \sum_{k=1}^{K}(\boldsymbol{f}(\boldsymbol{y}_k, \boldsymbol{\theta}) - \boldsymbol{s}_k)^2$. Therefore,

[a] Prof. X. Qu, Y. Huang, H. Lu, T. Qiu, Prof. Z. Chen
Department of Electronic Science, Fujian Provincial Key Laboratory of Plasma and Magnetic Resonance, State Key Laboratory of Physical Chemistry of Solid Surfaces, Xiamen University
P.O.Box 979, Xiamen 361005 (China)
E-mail: quxiaobo@xmu.edu.cn & chenz@xmu.edu.cn
[b] Prof. D. Guo
School of Computer and Information Engineering, Xiamen University of Technology
[c] Dr. T. Agback
Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden
[d] Prof. V. Orekhov
Department of Chemistry and Molecular Biology, University of Gothenburg, Box 465, Gothenburg 40530, Sweden.

Supporting information for this article is given via a link at the end of the document.

10.1002/anie.201908162
WILEY-VCH

COMMUNICATION

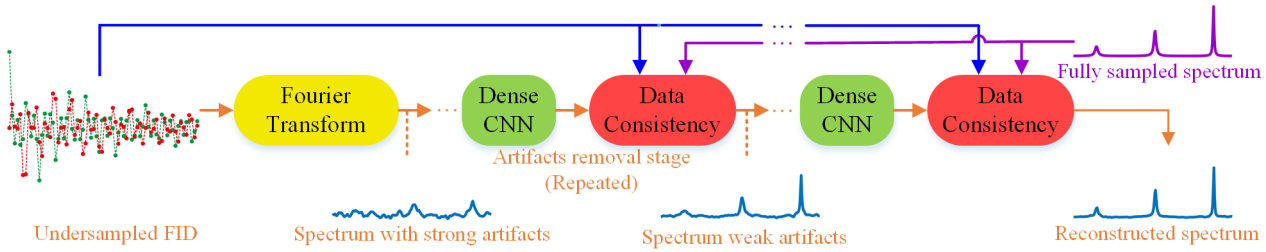

Figure 1. Flowchart of deep learning NMR spectroscopy. Note: Please refer to Supplement S1 for more details.

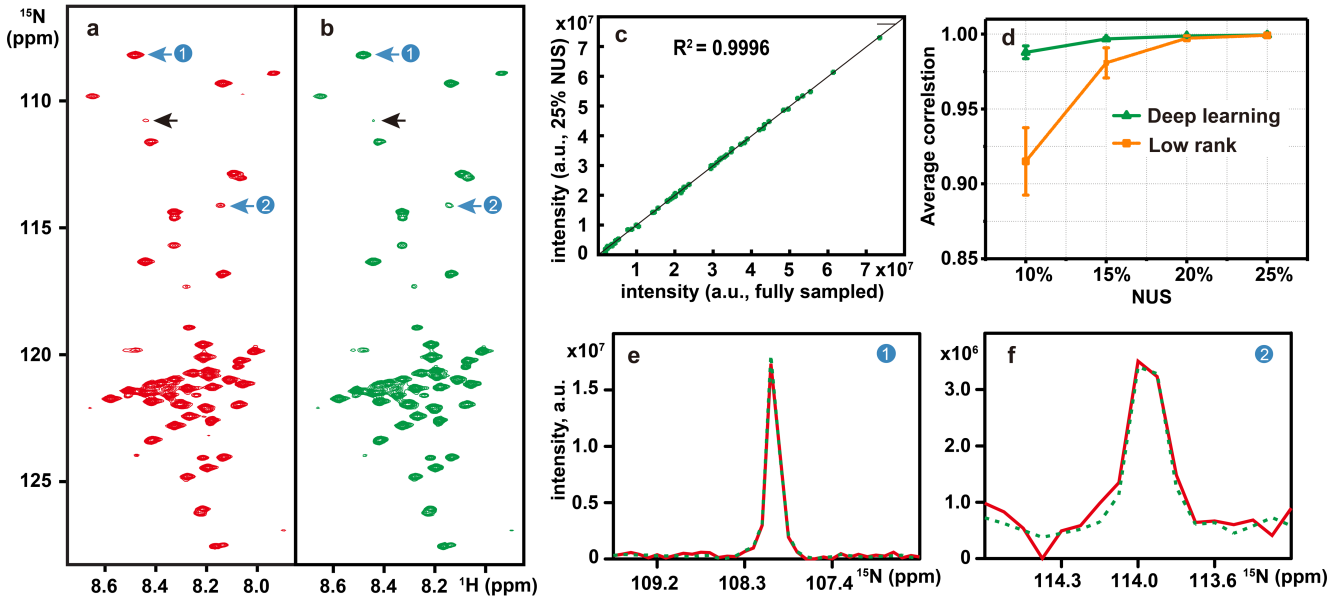

Figure 2. Reconstruction of a 2D ^{1}H–^{15}N HSQC spectrum of the cytosolic domain of CD79b protein from the B-cell receptor. (a) and (b) are the fully sampled spectrum and deep learning NMR reconstruction from 25% NUS data, respectively. (c) Peak intensity correlations between fully sampled spectrum and reconstructed spectrum. (d) denotes the peak intensity correlation obtained with the deep learning and low rank methods under different NUS levels. (e) and (f) are zoomed out 1D ^{15}N traces of (a). Red and green lines represent the reference and the reconstructed spectra, respectively. Note: The R^2 denotes the square of Pearson correlation coefficient. The closer the value of R^2 gets to 1, the stronger the correlation between the reference and the reconstructed spectra is. The average and standard deviations of correlations in (d) are computed over 100 NUS trials. The intensity distortion of small peaks in reconstruction is marked with the black arrow.

DL provides an optimal mapping $\boldsymbol{f}(\boldsymbol{y},\boldsymbol{\theta})$ from the input $\boldsymbol{y}$ to the target spectrum in the sense of least square error for all pairs. Then, for a given undersampled signal $\tilde{\boldsymbol{y}}$ from a NUS experiment, a spectrum $\tilde{\boldsymbol{s}}$ is obtained via $\tilde{\boldsymbol{s}} = \boldsymbol{f}(\tilde{\boldsymbol{y}},\boldsymbol{\theta})$.

To demonstrate the applicability of the DL NMR, we first validate the reconstruction performance on several fully sampled 2D and 3D spectra of small proteins. As shown in Figure 2, DL reconstructs excellent 2D ^{1}H-^{15}N HSQC spectrum from 25% NUS data with correlation of the peak intensity to the fully sampled spectrum reaching 0.9996. Figure 2d indicates that DL is in pair with the state-of-the-art reconstruction techniques[5a] in robustness and spectra quality and may even surpass the other methods at low NUS densities (See Supplement S4.2 for more details). High fidelity of the reconstructed peak shapes is illustrated in Figures 2e and 2f. Using the network with same trained parameters, the correlations greater than 0.98 were also obtained for 2D spectra of three other proteins (See Supplement S4.2). High potential of the DL in reconstructing high-quality multi-dimensional spectra is illustrated in Figure 3, exemplified by 3D HNCO for Azurin (14 kDa protein) and 3D HNCACB spectrum for GB1-HttNTQ7 (10 kDa protein). The peak intensity correlations approaching 0.99 for both 3D spectra (Figures 3e and 3f) indicates excellent fidelity of the DL reconstruction.

Figure 4 illustrates quality and performance of DL for challenging cases of a large protein MALT1 (44 kDa) and an intrinsically disordered protein alpha-synuclein (14.5 kDa) (See Supplement S2.1 for more details). Even with 10% data, DL provides robust high-quality spectra reconstruction, thus allowing up to factor of ten saving of experiment time for these challenging cases.

An important advantage of the DL NMR is fast spectra reconstruction due to harnessing of a non-iterative low-complexity neural network algorithm that allows massive parallelization with graphics processing units. Without compromising the spectra quality (See Supplements S2.2 and S2.3 for detailed comparisons), DL is much faster than other state-of-the-art methods such as low rank[5a] and compressed sensing[2a]. The comparisons, shown in Figure 5, indicate that the computational time of DL is 4%~8% of that needed for low rank for 2D spectra and 12%~22% of that consumed by compressed sensing for 3D spectra. Although the training time is long, which is 5.08 hours for 2D NMR and 31.68 hours for 3D NMR, a unique network can be trained in advance and then applied to reconstruct many spectra that have the same dimensionality (2D or 3D) and do not deviate much in sizes of the spectral dimensions and NUS levels (See Supplement S1.1.5 for more details).

10.1002/anie.201908162

WILEY-VCH

COMMUNICATION

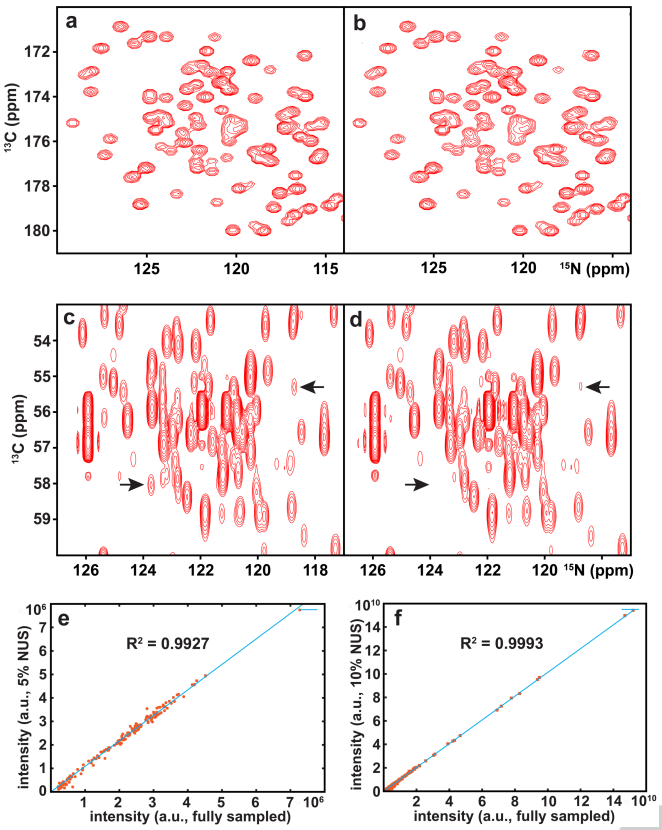

Figure 3. 3D spectra reconstructions for two small proteins, azurin (14 kDa) and GB1-HttNTQ7 (10 kDa). (a-b) Sub-regions of ^{13}C-^{15}N projections from the fully sampled HNCO spectrum of azurin and the deep Learning reconstruction with 5% NUS data. (c-d) Sub-regions of ^{13}C-^{15}N projections from the fully sampled HNCACB spectrum of GB1-HttNTQ7 and the deep Learning reconstruction with 10% NUS data. (e-f) Peak intensity correlations between deep learning reconstructions and fully sampled 3D spectra of azurin and GB1-HttNTQ7. Note: The contours of spectra are at the same level. The distortion and missing of some small peaks in reconstruction are marked with black arrows.

With the afore results, we have demonstrated that DL can be successfully applied to triple resonance experiment commonly used for protein backbone assignment. We anticipate that DL is also applicable to higher dimensional NMR, e.g. 4D or 5D, by synthetizing sufficient training data and carefully handling the Fourier transform on a large data set [10] or other experiment types, e.g. relaxation, diffusion or temperature series, by taking the advantage that network may automatically grasps essential features of the spectra series from the training on the corresponding synthetic data sets.

Limitations of the current DL NMR are similar to other methods used for reconstructing spectra from NUS data: (i) Due to inherent non-linearity of the algorithm, very small peaks may significantly change their amplitudes or disappear as indicated by black arrays in Figures 2 and 3. Thus, usage of DL for experiments with large dynamic range of cross-peak intensities, such as NOESY, should be done with caution; (ii) Unsatisfactory reconstructions under very low NUS density, e.g. 5% for 2D NMR, and/or low signal-to-noise ratios. Even though, in some cases in our tests, DL demonstrated somewhat more robust results than other state-of-the-art methods, e.g. low rank (See Supplement S4).

In summary, we present the proof-of-concept demonstration of application the DL for reconstructing high quality NMR spectra of small, large and disordered proteins from NUS data. This result opens an avenue for application of DL and possibly other artificial intelligence techniques in biological NMR.

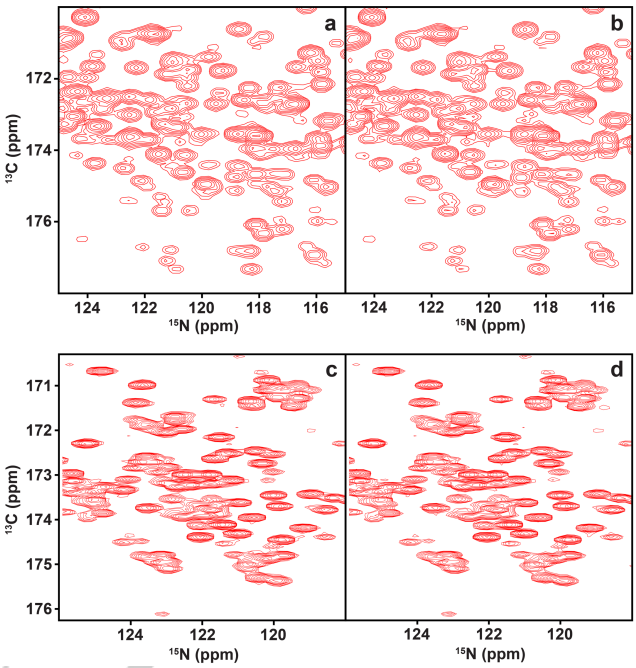

Figure 4. 3D HNCO spectra reconstructions for a large protein, MALT1 (44 kDa), and an intrinsically disordered protein, alpha-synuclein (14.5 kDa). (a-b) Sub-regions of ^{13}C-^{15}N projections from reconstructions of the MALT1 protein by DL with 30% and 10% NUS data, respectively. (c-d) Sub-regions of ^{13}C-^{15}N projections from reconstructions of the alpha-synuclein protein by DL with 15% and 10% NUS data, respectively. Note: The experimental data for MALT1 and alpha-synuclein proteins were acquired under 30% and 15% NUS, respectively. Further randomly under sampling is retrospectively applied to the experimental NUS data to emulate sampling at lower NUS densities. The contours in the pairs of spectra are at the same level.

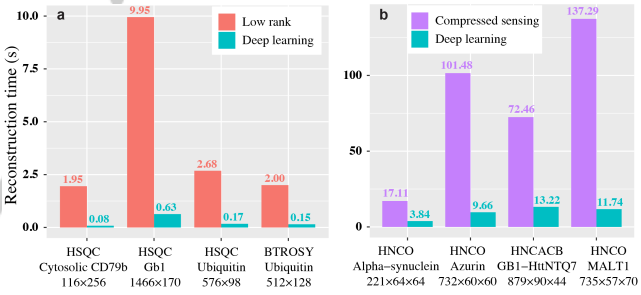

Figure 5. Computational time for the reconstructions of (a) 2D spectra and (b) 3D spectra. The spectra type, its corresponding protein and spectra size after routine processing of the direct dimension are listed below each bar. Details about the time comparisons are found in Supplement S3.

Acknowledgements

Authors thank Marius Clore and Samuel Kotler for providing the 3D HNCACB data; Jinfa Ying for assisting processing and helpful discussion on the 3D HNCACB spectrum; Luke Arbogast and Frank Delaglio for providing the 2D HSQC spectrum of GB1; Esmeralda Woestenenk for help with MALT1 protein production. This work was supported in part by the National Natural Science Foundation of China (NSFC) under grants 61571380, 61971361, 61871341 and U1632274, the Joint NSFC-Swedish Foundation for International Cooperation in Research and Higher Education (STINT) under grant 61811530021, the National Key R&D Program of China under grant 2017YFC0108703, the Natural Science Foundation of Fujian Province of China under grant 2018J06018, the Fundamental Research Funds for the Central

Universities under grant 20720180056, the Science and Technology Program of Xiamen under grant 3502Z20183053, the China Scholarship Council under grants 201806315010 and 201808350010, the Swedish Research Council under grant 2015–04614 and the Swedish Foundation for Strategic Research under grant ITM17-0218.

Conflict of interest

The authors declare no conflict of interest.

Keywords: artificial intelligence • deep learning • NMR spectroscopy • fast sampling

[1] a) V. Jaravine, I. Ibraghimov, V. Yu Orekhov, *Nat. Meth.* **2006**, *3*, 605–607; b) M. Mobli, J. C. Hoch, *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *83*, 21-41.

[2] a) K. Kazimierczuk, V. Y. Orekhov, *Angew. Chem., Int. Ed.* **2011**, *50*, 5556-5559; b) D. J. Holland, M. J. Bostock, L. F. Gladden, D. Nietlispach, *Angew. Chem., Int. Ed.* **2011**, *50*, 6548-6551; c) Y. Shrot, L. Frydman, *J. Magn. Reson.* **2011**, *209*, 352-358; d) M. Mayzel, K. Kazimierczuk, V. Y. Orekhov, *Chem. Commun.* **2014**, *50*, 8947-8950; e) X. Qu, X. Cao, D. Guo, Z. Chen, in *International Society for Magnetic Resonance in Medicine (ISMRM) 18th Scientific Meeting*, Stockholm, Sweden, **2010**, p. 3371; f) X. Qu, D. Guo, X. Cao, S. Cai, Z. Chen, *Sensors* **2011**, *11*, 8888-8909.

[3] J. Ying, F. Delaglio, D. A. Torchia, A. Bax, *J. Biomol. NMR* **2017**, *68*, 101-118.

[4] J. Ying, H. Lu, Q. Wei, J. Cai, D. Guo, J. Wu, Z. Chen, X. Qu, *IEEE Trans. Signal Process.* **2017**, *65*, 3702-3717.

[5] a) X. Qu, M. Mayzel, J.-F. Cai, Z. Chen, V. Orekhov, *Angew. Chem., Int. Ed.* **2015**, *54*, 852-854; b) J. Ying, J. Cai, D. Guo, G. Tang, Z. Chen, X. Qu, *IEEE Trans. Signal Proces.* **2018**, *66*, 5520-5533; c) H. Lu, X. Zhang, T. Qiu, J. Yang, J. Ying, D. Guo, Z. Chen, X. Qu, *IEEE Trans. Biomed. Eng.* **2018**, *65*, 809-820; d) D. Guo, H. Lu, X. Qu, *IEEE Access* **2017**, *5*, 16033-16039; e) D. Guo, X. Qu, *IEEE Access* **2018**, *6*, 4933-4940; f) X. Qu, T. Qiu, D. Guo, H. Lu, J. Ying, M. Shen, B. Hu, V. Orekhov, Z. Chen, *Chem. Commun.* **2018**, *54*, 10958-10961.

[6] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436–444.

[7] a) S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, D. Liang, in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, IEEE, Prague, Czech Republic, **2016**, pp. 514-517; b) J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, D. Rueckert, *IEEE Trans. Med. Imaging* **2018**, *37*, 491-503.

[8] a) S. G. Worswick, J. A. Spencer, G. Jeschke, I. Kuprov, *Science Advances* **2018**, *4*, eaat5218; b) P. Klukowski, M. Augoff, M. Zięba, M. Drwal, A. Gonczarek, M. J. Walczak, *Bioinformatics* **2018**, *34*, 2590-2597.

[9] G. Huang, Z. Liu, L. v. d. Maaten, K. Q. Weinberger, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2017**, pp. 2261-2269.

[10] Y. Pustovalova, M. Mayzel, V. Y. Orekhov, *Angewandte Chemie International Edition* **2018**, *57*, 14043-14045.

10.1002/anie.201908162

WILEY-VCH

COMMUNICATION

X. Qu, Y. Huang, H. Lu, T. Qiu, D. Guo, T. Agback, V. Orekhov, Z. Chen**

Page No. – Page No.

Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning

The first proof-of-concept of application of deep learning, an artificial intelligence technique, for high-quality, reliable, and very fast NMR spectra reconstruction from limited experimental data.