Cobas Carlos (Orcid ID: 0000-0002-7494-7876)

NMR signal processing, prediction and structure verification with Machine Learning techniques

Carlos Cobas

carlos@mestrelab.com

Mestrelab Research, S.L Feliciano Barrera 9B – Bajo, 15706 Santiago de Compostela

(SPAIN)

Introduction

Artificial Intelligence (AI) is increasingly present in our daily routine, drastically shaping our world, sometimes subtly, almost invisibly, sometimes in a much more obvious way. Our phones, smart TVs, spam filters, Google searches, banking transactions, autonomous vehicles, etc. are constantly executing algorithms based on AI techniques, among which Machine and Deep Learning are two of the most eminent representatives and probably the most used catchwords today.

Even though Machine Learning (ML) has generated quite the buzz, it has been around for a long time. It was born from pattern recognition (PR) and the theory that computers can learn without being programmed to perform specific tasks. It is a subset of artificial intelligence which aims at extracting complex patterns and relationships from large data sets, to predict specific properties of the data. The enormous resurging interest in ML is due to growing volumes and varieties of available data, coupled with the rapid development of computer power and affordable data storage. It's a science that's not new, but one that has gained fresh momentum.

NMR spectroscopy has not been oblivious to this technology and a long tradition connects ML to NMR. Indeed, NMR appears to be especially suitable for the application of ML techniques. It is a quantitative, highly reproducible and information-rich technique. NMR spectra can act as "fingerprints" which can be used to compare, discriminate or classify samples. One of the earliest applications of PR to NMR data, demonstrating a direct route from spectrum to molecular structural information, was reported by B. R. Kowalski et. al.[1] already in 1971 using artificial neurons based on Linear Threshold Unit (TLU). In a later article,[2] Kowalski proposed the use of K-Nearest Neighbors algorithm, a popular classifier that still has its place in most of the modern PR and ML toolboxes. At this point it is worth nothing that both PR and ML strongly overlap and there is no sharp distinction. It might be considered that PR is a problem-solving task or tool in ML, and both are core techniques used in chemometrics. In this article no further attempts to distinguish between those concepts will be made and for a more formal discussion the reader is referred to excellent reviews.[3]

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/mrc.4989

Since those pioneering successful applications, numerous ML methods in diverse areas of NMR spectroscopy have been described. Many of them were extensively applied for the analysis of complex mixtures in the "omics" context[4,5] (e.g. metabolomics/metabonomics, lipidomics, foodomics, etc) as well as in clinical applications.[6] More recently, Deep learning (DL) techniques are gaining interest in different fields of NMR, particularly in Magnetic Resonance Imaging (MRI),[7] due to their self-learning and generalization ability over large amounts of data. In this context, compared with conventional ML methods, which require the process of selecting and creating features from the input data, DL allows creation of the most suitable set of features within the process of training, without any design or involvement by the researcher.

In this work, only applications of ML and DL techniques in the areas of NMR signal processing and analysis of small molecules, including automatic structure verification and prediction of NMR observables in solution (e.g. chemical shifts and coupling constants) will be discussed. Applications in other fields such as NMR relaxometry, Magnetic Resonance Spectroscopy for clinical diagnosis, Protein NMR or metabonomics will not be covered.

Signal Processing

ML has not been an easy playground for NMR signal processing applications and only a few of them have been proposed for high resolution NMR applications. ML strongly depends on training data and getting experimental spectra is expensive (specially to get them assigned or labeled) and may result in sparse and unbalanced data sets. An interesting ML exercise providing a valuable lesson on the importance of training data was recently reported by Hoch.[8] They tried to apply a ML algorithm (the details of the actual ML algorithm have not been disclosed) to find optimal sampling schedules for Non-Uniformly Sampled (NUS) NMR spectra. They found that when trained against a single spectrum, the results were much better than randomly choosing a subset. However, the ML-optimized sampling schedule performed worse than random sampling when used to recover a similar spectrum with some of the peaks located at different frequencies. This example raises the question about how much data or coverage is needed in the training set: ideally, the training data should contain examples of the full range of possibilities. This is of course a pipe dream, but it highlights the importance of both the size and diversity of the training set.

In general, DL-based methods require a large amount of experimental data at the training stage. Since obtaining such amount of data is unrealistic, two research groups suggested recently the use of simulated or synthetic data, as opposed to experimental spectra.[9,10] Qu et al simulated 40,000 fully sampled synthetic FIDs following the classical exponential model. For each of these FIDs, corresponding under-sampled FIDs were generated by using a Poisson-gap scheme. On the other hand, Hansen developed a proof-of-concept approach using 1D spectra consisting of ca $8x10^6$ synthetic FIDs. In this work, the author uses long short-term memory (LSTM) networks[11] that have traditionally been used to analyze time series, whereas Qu et al. used an entirely different DL architecture, based on densely connected convolution neural network (CNN). Using this architecture, they show that their DL method not only provides robust high-quality spectra reconstruction but also that DL is much faster than other state-of-the-art methods such compressed sensing or low rank reconstruction.

Another application of DL that takes advantage of using simulated data was recently proposed by Bruker.[12] They simulated 1D 1H spectra for 2 million structures from PubChem. These perfect spectra were then transformed into more realistic spectra by adding noise, impurities, and line shape, phase and baseline distortions. After training the DL on this simulated data, it was able to

detect signal regions (i.e. multiplets) in real experimental NMR spectra that correspond to manually selected signal regions.

CNNs have also been used to automate peak picking of NMR spectra.[13] As CNN were primarily intended for the recognition of natural images, the authors introduced a two-way spectra normalization procedure, where NMR data points are normalized with respect to intensity and resolution. The proposed deep learning-based method, NMRNet, seems to provide high accuracy on datasets of proteins of diverse kind. Application of the method to the NMR spectra of small molecules was not discussed in the original paper.

A totally different application of Neural Networks in the context of NMR signal processing was proposed by Morita[14] for the estimation of the spectral resonances present in an NMR FID. The proposed procedure took advantage of the fact that NMR signals live in the complex domain and hence, they developed a method based on complex-valued Hopfield Networks in which the weights and thresholds for conventional networks are expanded to accommodate complex numbers. It is worth pointing out that, even though it employs Neural Networks, strictly speaking it would not be correct to classify it as a ML method since it does not have a learning process but rather, it exploits the ability of the Neural Networks to find a local minimum solution to determine the NMR parameters. Nevertheless, whilst the method showed promising results with simulated data and, as opposed to other methods, it does not require complicated matrix calculations, it was found that it is still inadequate for a practical application. Some improvements to this method were proposed later by El-Barky and Nikos Mastorakis[15] but still, it has not found widespread use.

NMR Prediction

Resolving a molecular structure on the basis of the measured NMR spectra can be classified as an inverse problem where the goal is to recover 'hidden' information (i.e. the molecule) from 'outside' noisy data (NMR data). This is an ill-posed problem for which unique solutions can only be selected by imposing some constraints. For instance, 2D NMR experiments have been proven to be extremely effective providing unequivocal "hard" proof of through-bond atom connectivities and hence drastically narrowing down the acceptable hypotheses set.[16] However, this information may not lead to an unambiguous structure and more additional knowledge is usually required. An alternative, complementary tool to regularize the problem is given by chemical shift prediction: of all the potential molecules initially compatible with the different available NMR data, only those which can be justified in the light of the predicted chemical shifts are retained.

Predictions are, therefore, very valuable for a variety of applications including structure validation,[17] automatic assignment[18] and compound elucidation.[19] The mere mention of the term prediction points directly to a potential application of ML techniques. Indeed, already in the first years of development of NN techniques, they were applied to the prediction of NMR, initially to specific groups of organic compounds such as alkanes,[20,21] alkenes[22] or benzenes[23] and then generalized to nearly every class of organic molecules for different nuclides, including ^{31}P,[24] ^{13}C[25] and ^{1}H.[26,27] NN-based predictions, available both in commercial[28] and free applications packages were followed by other ML techniques such as Supported Vector Machines (SVM),[29] Random Forests classification[30] and Principal Least Regression (PLS).[31]

ML methods have several advantages over purely database-based predictions such as those based on HOSE codes.[32] First, once the ML prediction model has been created, its execution is very fast, typically orders of magnitude faster than using database predictions. Second, database predictions work by identifying in a database similar structures (i.e. those having matching HOSE codes up to a number of shells) and returning weighted averages of the experimental data corresponding to these

structures. The accuracy of these prediction strongly depends on the similarity between the new and known HOSE codes and on the quality of the database. That is to say, this approach exactly reproduces the contents of the internal database, including every error within that reference collection.

Conversely, ML/DL methods are lauded for their generalization capabilities, that is, going beyond the specific given examples to classify or predict unseen examples. This generalization property, however, must be taken with some caution: in general, ML are usually capable of generalizing a model within the subspace spanned by the training samples, but not outside of it. In other words, they can interpolate data, that is, make predictions about a situation that lies between other, known situations. On the other hand, ML algorithms have huge problems when you move outside of the area populated by the training dataset in the vector space, although there are some novel investigations aimed at addressing this limitation so that ML can make better extrapolations for unknown situations.[33]

Therefore, ML/DL NMR prediction methods are also sensitive to the experimental data set used for model building. Increasing the size of the training set would be the simplest way to improve the performance of the predictor. However, no matter how large the database used, it is going to be extremely tiny compared to the actual chemical space (i.e. the universe of possible compounds). For example, there are more than 166 billion organic small molecules (up to 17 atoms of C, N, O, S, and halogens) compatible with many drugs.[34] And if we consider larger molecules (e.g. MW up to 500), the best guess for the number of plausible compounds is around 10^{60}. There is therefore an effectively endless frontier and having access to experimental data that covers such chemical space is unrealistic. Whilst some of those assigned data sets are publicly available, such as NMRShiftDB2,[35] others are proprietary and cannot be disclosed. However, it is possible to run different prediction algorithms, including ML, HOSE code or even Quantum Chemical-based ones and combine the results of all of them in a way that the final predicted values are both more accurate and precise than those coming from each individual predictor. That is the main concept under the so-called *Ensemble NMR Prediction* developed by Mestrelab Research.[36] It consists of a Bayesian algorithm that uses the information of the individual predicted chemical shifts (such as their distances) as well as their corresponding confidence intervals. Therefore, it is necessary that each predictor not only calculates the chemical shift of each atom but it also has to provide estimates of uncertainty, again, on individual atom basis. Excellent results have been obtained with this method both for 13C and 1H predictions.[37]

This method can *plug* any other prediction method, provided that a confidence interval can be assigned to each predictor value. One DL-based prediction method with that capability has recently been published.[38] Furthermore, Density Functional Theory (DFT) calculations can also be used, in particular in cases when the molecule of interest is somewhat "exotic" and is not properly handled by any of the empirical or ML-derived methods. However, this would require the development of a mechanism that permits the calculation of the prediction error bounds of individual atoms for DFT-based calculations.

Some of these methods, including CSEARCH[39] (available in Mnova software), the more recent Stereo-Aware Extension of HOSE Codes code,[40] the NN method by Aires-de-Sousa et. al. and Mestrelab XGBoost prediction method provide some 3D perception capabilities which can be used for the analysis of the relative 3D configuration (e.g. diastereoisomers). However, the discrimination power is usually quite limited and cannot deal with 3-dimensional conformational analysis.

Over the last few years, quantum chemical calculations, typically based on DFT methods, have become the established approach for modern 3D structure determination.[41] Whilst optimal DFT methods can be accurate to within 0.2/2 ppm for ^{1}H and ^{13}C chemical shifts respectively, prediction times for a single molecule range from several minutes to days, especially when computing scalar coupling constants. In addition, to get accurate values in the case of flexible molecules, exploration of the conformational space becomes necessary, making such calculations even more time consuming and expensive.

In an effort to develop a prediction engine that can be effectively used for 3D structure discrimination studies, Gerrard et al developed a ML method system for the prediction of NMR parameters, trained using large datasets of DFT-computed NMR parameters, such as chemical shifts and scalar couplings, derived from 3D structures.[42] This overcomes the need for the existence of large, structurally diverse, error-free experimental databases. In this work, the authors used a Kernel Ridge Regression framework for model building and FCHL[43] to encode the similarity between chemical environments of each molecular structure. The authors showed that this engine is capable of reproducing DFT-predicted NMR parameters for a range of experimentally relevant systems with high accuracy but in a fraction of the time. Application of this method to the 3D relative structure configuration of several challenging cases was also reported in this work.

Finally, an ingenious alternative to the previous prediction methods is *Ask Erno*.[44] Given that obtaining a sufficiently large training set consisting of assigned NMR spectra is complex and laborious, the authors processed a fully automatic self-learning assignment and prediction system that progressively improves its capabilities as it solves more instances of assignments. Briefly, this system starts by automatically assigning the 1H-NMR spectra using an algorithm that does not require predictions.[45] This initial assignment is used to populate a database of assigned protons that is used by a HOSE code-based chemical shift predictor. Next, these predictions facilitate improvement of the assignment process. Iteration on these steps allows *Ask Ernö* to improve its ability to assign and predict spectra without any prior knowledge or human intervention. The learning algorithm is based on a self-organizing map and consists of a recursive cycle on the training dataset, which is repeated until nothing new is learnt.

Automatic Structure Verification (ASV)

ASV can be defined as the computer algorithms attempting to assess the degree of compatibility between a proposed molecular structure and one or more sets of spectral data.

In general, an ASV system can be classified as a fuzzy logic problem in which many aspects cannot be described in an exact way using deterministic rules. Experimental chemical shifts will vary depending upon the experimental conditions, including solvent, temperature, pH, concentration, salt content, etc. Theoretically, all these parameters can be modeled in such a way that, for instance, NMR predictions can take all those considerations into account. However, this introduces an enormous level of complexity and whilst exist some prediction algorithms that consider some of them, all prediction methods provide values with varying degrees of uncertainty in their estimates.

Furthermore, experimental spectra exhibit a number of features that greatly complicate the fully automatic analysis and contribute to the fuzziness of the verification problem. Examples of those include truncation artefacts, magnetic field inhomogeneity, small impurities, residual solvents, baseline distortions, labile protons, to cite a few. The latter are especially difficult to handle as their chemical shift often fluctuate in a more unpredictable way than non-exchangeable protons. Primary residual solvent peaks can be easily identified when they are isolated from other sample peaks or when they show a clearly characteristic multiplet pattern (e.g. the quintuplet corresponding to

DMSO resonating at c.a. 2.54 ppm), but it becomes more challenging when it is, for example, a singlet in a highly overlapped spectral region (e.g. residual $CHCl_3$ peak at c.a. 7.26 ppm). Another complex issue is the automatic recognition of residual water as a secondary solvent as its chemical shift is quite temperature-dependent and any potential hydrogen bond acceptor will tend to shift it from its expected nominal chemical shift, even by several ppm. Moreover, the water signal does not only show up as a singlet but also as a more complex multiplet pattern. Depending on the experimental conditions, the water peak can appear in different forms, line widths and chemical shifts.

Any ASV system should have the critical ability to deal efficiently with overlapping peaks. This is particularly important when some of the relevant peaks of the compound to be analyzed are very close to other signals such as large solvent peaks. Naturally, there will be cases in which peaks' overlap cannot be disentangled, but this highlights the importance of using NMR analysis procedures that can reliably work with individual peaks rather than with full multiplets quantified by standard NMR integration methods. An example of this is given in Figure 1. The large residual water signal makes the correct quantification of H-13 impossible if traditional integration methods (e.g. running sum) are used. Global Spectral Deconvolution (GSD),[46] properly identifies and quantifies the triplet and therefore, permits the full assignment and validation of the molecular structure.

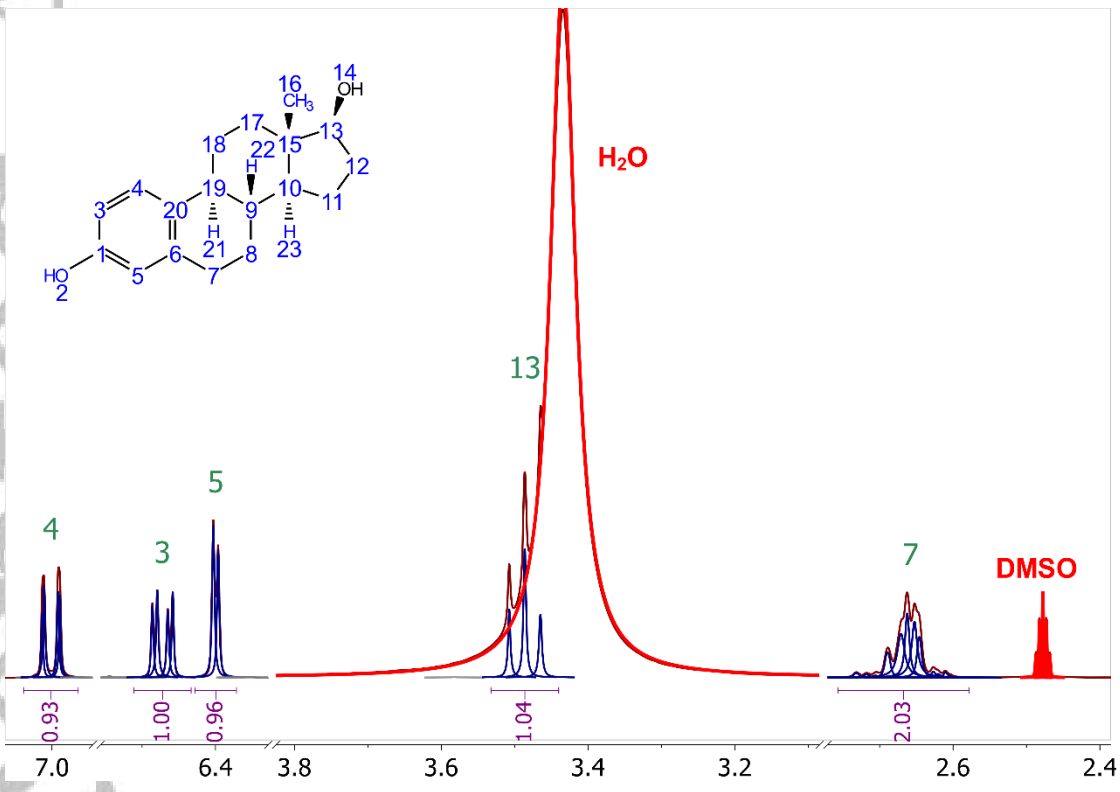

Figure 1. Expansions of the 1H NMR spectrum of Estradiol (some spectral regions have been cut, as shown by the marks in the horizontal scale). H-13 is heavily overlapped by the large residual water resonance. Attempts to quantify this multiplet with standard integration methods would fail. In this case, this triplet can be properly quantified by using GSD. Blue curves correspond to peaks classified as compounds whilst red curves are solvent peaks.

A standard ASV system does not address the question of whether other molecules might also match the spectral data, but some related procedures are devised to try to answer a different question: Given this spectrum and these *n* structures, which structure is the best match and is the difference significant? Strictly speaking this is not ASV but rather *structure discrimination* or *ranking*. Such process usually involves the comparison between the experimental spectrum with the predicted one

at different levels. For example, structure elucidation software packages typically produce many different molecules that are compatible with the NMR constraints. These structures are then ranked according to the *distance* or *similarity* between the experimental chemical shifts and the predictions. This ranking is usually performed by using simple statistics such as the standard deviation between experimental and predicted values or other metrics such as the mean absolute error (MAE) or correlation coefficient, R^2. Because the number of resulting structures can be quite large, these predictions usually consist of fast methods (e.g. ML-based methods, *vide supra*) and DFT calculations are only used in special cases, particularly when stereochemistry discrimination is important. In this case, more sophisticated statistical methods have been developed, such as the DP4[47] and DP4+[48] (see [49] for an excellent review on those metrics). Very recently, new metrics, the J-DP4 methods,[50] that can accommodate scalar couplings, in addition to chemical shifts, have also been proposed resulting in a 2.5-fold performance improvement.

All these methods assume that the correct molecule is present in the list of proposed structures. This can be useful in situations where the constitution of the molecule is known and only the relative configuration needs to be confirmed. There exist fully automatic software solutions, usually intended for the first problem, that is, confirmation of the molecular constitution,[51,52,53] although some of them can also be complemented with more specialized tools aimed at resolving the relative configuration.[54] However, as none of them can be considered ML methods, they will not be discussed here.

One of the first ML-based approaches for ASV applications in which only one putative molecular structure needs to be provided was proposed by Sarotti in 2013.[55] In this work, the author used two-layer feed-forward networks and 13C-NMR predictions with DFT calculations (mPW1PW91 functional and the 6-31G(d) basis set). The overall procedure is as follows: as a training set, 100 small-to-medium sized compounds were selected with known assigned 13C-NMR. For each of those molecules, incorrect structures were created by manually modifying the original structures. Next, statistical features calculated from the correlation between the experimental and predicted NMR chemical shifts were used as the input vector for the NN model building. This method showed very good results mostly in the identification of connectivity errors, that is, in the classification of the proposed molecular constitution as being compatible or not with the experimental 13C spectrum. However, it did not show discriminating power in cases where the molecules differ in stereochemistry (e.g. different diastereoisomers).

In order to take this approach a step further so that it can detect subtle structural differences typically found in regio- or stereochemical problems, in 2015 the same group extended the application of ANN to include HSQC experiments.[56] Following a conceptually similar strategy, several parameters of correlation between experimental and calculated monodimensional 1H and 13C data (MAE, CMAE, etc.) were computed. In addition, 18 extra parameters were computed from the correlation between experimental and calculated HSQC data. It was found that the new ANNs afforded very high (>92%) percentage of correct classification even in the case of very challenging stereochemistry problems.

ML has also found its applications in the field of Natural Products dereplication, that is, the process of assessing the uniqueness of a new compound in relationship to all known ones. In 2017, Zhang et al[57] used a Deep Convolutional Neural Network with contrastive loss trained on a database containing over 2,054 HSQC spectra to map the spectra into a cluster space where similar compounds are near one another and dissimilar compounds are far apart. This tool, dubbed as SMART (Small Molecule Accurate Recognition Technology) was successfully applied to rapidly

associate newly isolated natural products with their known analogues for the automatic compound dereplication and assignment to molecular structure families.

Conclusions

There has been tremendous progress in the application of ML techniques in NMR spectroscopy and this review has emphasized their role in the field of small molecules. Successful applications include fast processing of NUS spectra (not strictly confined to small molecules), prediction of NMR spectra and automatic structure verification. However, it is important to temper excessive optimism as there are still some significant challenges ahead.

One of the biggest problems plaguing ML in general, and DL in particular is that training them requires massive datasets before they can give useful results. If a model is fed poorly, it will only give poor results. For instance, robust prediction of NMR parameters must cope with the huge molecular universe. If the training set is not diverse enough, predictions will be limited to the narrow chemical space available in the training set. In this context, it is very important to highlight the fact that in general, predictions should always be accompanied by an estimation of their uncertainties. This is particularly important in structure verification and elucidation problems.

In NMR spectroscopy, the lack of large data sets is a strong limiting factor. Whilst NMR spectra can be acquired in high throughput, manual labeling (e.g. spectral assignments), by experts make it really challenging to generate very large training benchmarks. A popular approach to circumvent this problem relies on the usage of synthetic training data and several examples have been shown in this work. This does not only help in producing large data sets, but also in a controlled and systematic way that is usually not possible with real data. However, it is also important to bear in mind that models trained on synthetic data do not necessarily generalize well to real data.

[1] Reilly, C. A., & Kowalski, B. R. (1971). Nuclear magnetic resonance spectral interpretation by pattern recognition. The Journal of Physical Chemistry, 75(10), 1402–1411. https://doi.org/10.1021/j100680a008

[2] Kowalski, B. R., & Bender, C. F. (1972). K-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. Analytical Chemistry, 44(8), 1405–1411. https://doi.org/10.1021/ac60316a008

[3] Brereton, R. G. (2015). Pattern recognition in chemometrics. Chemometrics and Intelligent Laboratory Systems, 149, 90–96. https://doi.org/10.1016/j.chemolab.2015.06.012
[4] Lindon, J. C., Holmes, E., & Nicholson, J. K. (2001). Pattern recognition methods and applications in biomedical magnetic resonance. Progress in Nuclear Magnetic Resonance Spectroscopy, 39(1), 1–40. https://doi.org/10.1016/s0079-6565(00)00036-4
[5] Mendez, K. M., Broadhurst, D. I., & Reinke, S. N. (2019). The application of artificial neural networks in metabolomics: a historical perspective. Metabolomics, 15(11). https://doi.org/10.1007/s11306-019-1608-0
[6] El-Deredy, W. (1997). Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: a review. NMR in Biomedicine, 10(3), 99–124. https://doi.org/10.1002/(sici)1099-1492(199705)10:3<99::aid-nbm461>3.0.co;2-#
[7] Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., & Erickson, B. J. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. Journal of Digital Imaging, 30(4), 449–459. https://doi.org/10.1007/s10278-017-9983-4
[8] Hoch, J. C. (2019). If machines can learn, who needs scientists? Journal of Magnetic Resonance, 306, 162–166. https://doi.org/10.1016/j.jmr.2019.07.044
[9] Qu, X., Huang, Y., Lu, H., Qiu, T., Guo, D., Agback, T., … Chen, Z. (2019). Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning. Angewandte Chemie International Edition. https://doi.org/10.1002/anie.201908162
[10] Hansen, D. F. (2019). Using Deep Neural Networks to Reconstruct Non-uniformly Sampled NMR Spectra. Journal of Biomolecular NMR, 73(10–11), 577–585. https://doi.org/10.1007/s10858-019-00265-1
[11] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[12] Deep Learning Applications in NMR Spectroscopy, Bruker 2019. https://www.bruker.com/fileadmin/user_upload/5-Events/2019/BBIO/EUROISMAR/Deep-Learning-Applications-in-NMR-low-res.pdf . Last accessed 8/12/2019
[13] Klukowski, P., Augoff, M., Zięba, M., Drwal, M., Gonczarek, A., & Walczak, M. J. (2018). NMRNet: a deep learning approach to automated peak picking of protein NMR spectra. Bioinformatics, 34(15), 2590–2597. https://doi.org/10.1093/bioinformatics/bty134
[14] Morita, N., & Konishi, O. (2004). A method of estimation of magnetic resonance spectroscopy using complex-valued neural networks. Systems and Computers in Japan, 35(10), 14–22. https://doi.org/10.1002/scj.10705
[15] El-Bakry, H. M. (2007). A Novel Hopfield Neural Network For Perfect Calculation Of Magnetic Resonance Spectroscopy. Zenodo. https://doi.org/10.5281/ZENODO.108179
[16] Peng, C., Yuan, S., Zheng, C., & Hui, Y. (1994). Efficient Application of 2D NMR Correlation Information in Computer-Assisted Structure Elucidation of Complex Natural Products. Journal of Chemical Information and Modeling, 34(4), 805–813. https://doi.org/10.1021/ci00020a013
[17] Cobas, C., Bernstein, M., & Sýkora, S. (2014). An Integrated Approach to Structure Verification Using Automated Procedures. In Structure Elucidation in Organic Chemistry (pp. 445–492). https://doi.org/10.1002/9783527664610.ch12
[18] Cobas, C., Seoane, F., Vaz, E., Bernstein, M. A., Dominguez, S., Pérez, M., & Sýkora, S. (2013). Automatic assignment of1H-NMR spectra of small molecules. Magnetic Resonance in Chemistry, 51(10), 649–654. https://doi.org/10.1002/mrc.3995
[19] Lodewyk, M. W., Siebert, M. R., & Tantillo, D. J. (2011). Computational Prediction of1H and13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. Chemical Reviews, 112(3), 1839–1862. https://doi.org/10.1021/cr200106v
[20] Davidge, R. Predicting Spectra Using Rule Induction and Neural Nets. AISBQ Postgrad. AI Workshop 1990, 16-21
[21] Doucet, J. P., Panaye, A., Feuilleaubois, E., & Ladd, P. (1993). Neural networks and carbon-13 NMR shift prediction. Journal of Chemical Information and Modeling, 33(3), 320–324. https://doi.org/10.1021/ci00013a007
[22] Ivanciuc, O., Rabine, J.-P., Cabrol-Bass, D., Panaye, A., & Doucet, J. P. (1996). 13C NMR Chemical Shift Prediction of sp2Carbon Atoms in Acyclic Alkenes Using Neural Networks. Journal of Chemical Information and Computer Sciences, 36(4), 644–653. https://doi.org/10.1021/ci950131x
[23] Sklenak, S.; Kvasnicka, V.; Pospichal, J. Prediction of 13C NMR Chemical Shifts by Neural Networks in a Series of Monosubstituted Benzenes. Chem. Papers 1994, 48, 135-140
[24] West, G. M. J. (1993). Predicting phosphorus NMR shifts using neural networks. Journal of Chemical Information and Modeling, 33(4), 577–589. https://doi.org/10.1021/ci00014a009

[25] Meiler, J., Meusinger, R., & Will, M. (2000). Fast Determination of13C NMR Chemical Shifts Using Artificial Neural Networks. Journal of Chemical Information and Computer Sciences, 40(5), 1169–1176. https://doi.org/10.1021/ci000021c

[26] Aires-de-Sousa, J., Hemmer, M. C., & Gasteiger, J. (2002). Prediction of1H NMR Chemical Shifts Using Neural Networks. Analytical Chemistry, 74(1), 80–90. https://doi.org/10.1021/ac010737m

[27] Binev, Y., & Aires-de-Sousa, J. (2004). Structure-Based Predictions of1H NMR Chemical Shifts Using Feed-Forward Neural Networks. Journal of Chemical Information and Computer Sciences, 44(3), 940–945. https://doi.org/10.1021/ci034228s

[28] (i) Modgraph Consultants (www.modgraph.co.uk – Last accessed 2/1/2020) includes a NN prediction method that complements their HOSE-based prediction. This prediction system is commercially available from Mnova NMRPredict (https://mestrelab.com/software/mnova/nmr-predict/ - Last accessed 2/1/2020). (ii) ACD/Labs also includes a NN prediction method, both for 1H and 13C (https://www.acdlabs.com/products/adh/nmr/nmr_pred/ - Last accessed 2/1/2020).

[29] Kuhn, S., Egert, B., Neumann, S., & Steinbeck, C. (2008). Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. BMC Bioinformatics, 9(1), 400. https://doi.org/10.1186/1471-2105-9-400

[30] Laatikainen, R., Hassinen, T., Lehtivarjo, J., Tiainen, M., Jungman, J., Tynkkynen, T., … Tuppurainen, K. (2014). Comprehensive Strategy for Proton Chemical Shift Prediction: Linear Prediction with Nonlinear Corrections. Journal of Chemical Information and Modeling, 54(2), 419–430. https://doi.org/10.1021/ci400648s

[31] Blinov, K. A., Smurnyy, Y. D., Churanova, T. S., Elyashberg, M. E., & Williams, A. J. (2009). Development of a fast and accurate method of 13C NMR chemical shift prediction. Chemometrics and Intelligent Laboratory Systems, 97(1), 91–97. https://doi.org/10.1016/j.chemolab.2009.01.010

[32] Bremser, W. (1978). Hose — a novel substructure code. Analytica Chimica Acta, 103(4), 355–365. https://doi.org/10.1016/s0003-2670(01)83100-7

[33] Subham S. Sahoo, Christoph H. Lampert, Georg Martius, "Learning Equations for Extrapolation and Control", 2017, https://arxiv.org/abs/1806.07259v1

[34] Ruddigkeit, L., van Deursen, R., Blum, L. C., & Reymond, J.-L. (2012). Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. Journal of Chemical Information and Modeling, 52(11), 2864–2875. https://doi.org/10.1021/ci300415d

[35] Kuhn, S., & Schlörer, N. E. (2015). Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 - a free in-house NMR database with integrated LIMS for academic service laboratories. Magnetic Resonance in Chemistry, 53(8), 582–589. https://doi.org/10.1002/mrc.4263

[36] Carlos Cobas, "Ensemble NMR Prediction", https://resources.mestrelab.com/ensemble-nmr-prediction/ Last Accessed, 27/10/2019

[37] Carlos Cobas, "1H NMR Prediction: unity creates strength": http://nmr-analysis.blogspot.com/2019/05/1h-nmr-prediction-unity-creates-strength.html - Last Accessed: 2/2/2020

[38] Jonas, E., & Kuhn, S. (2019). Rapid prediction of NMR spectral properties with quantified uncertainty. Journal of Cheminformatics, 11(1). https://doi.org/10.1186/s13321-019-0374-3

[39] Schütz, V., Purtuc, V., Felsinger, S., & Robien, W. (1997). CSEARCH-STEREO: A new generation of NMR database systems allowing three-dimensional spectrum prediction. Fresenius' Journal of Analytical Chemistry, 359(1), 33–41. https://doi.org/10.1007/s002160050531

[40] Kuhn, S., & Johnson, S. R. (2019). Stereo-Aware Extension of HOSE Codes. ACS Omega, 4(4), 7323–7329. https://doi.org/10.1021/acsomega.9b00488

[41] Willoughby, P. H., Jansma, M. J., & Hoye, T. R. (2014). A guide to small-molecule structure assignment through computation of (1H and 13C) NMR chemical shifts. Nature Protocols, 9(3), 643–660. https://doi.org/10.1038/nprot.2014.042

[42] Gerrard, W., Bratholm, L. A., Packer, M. J., Mulholland, A. J., Glowacki, D. R., & Butts, C. P. (2020). IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. Chemical Science. https://doi.org/10.1039/c9sc03854j

[43] Faber, F. A., Christensen, A. S., Huang, B., & von Lilienfeld, O. A. (2018). Alchemical and structural distribution based representation for universal quantum machine learning. The Journal of Chemical Physics, 148(24), 241717. https://doi.org/10.1063/1.5020710

[44] Castillo, A. M., Bernal, A., Dieden, R., Patiny, L., & Wist, J. (2016). "Ask Ernö": a self-learning tool for assignment and prediction of nuclear magnetic resonance spectra. Journal of Cheminformatics, 8(1). https://doi.org/10.1186/s13321-016-0134-6

[45] Castillo, A. M., Bernal, A., Patiny, L., & Wist, J. (2015). Fully automatic assignment of small molecules' NMR spectra without relying on chemical shift predictions. Magnetic Resonance in Chemistry, 53(8), 603–611. https://doi.org/10.1002/mrc.4272
[46] C. Cobas, F. Seoane, S. Domínguez, S. Sykora and A. N. Davies, "A new approach to improving automated analysis of proton NMR spectra through Global Spectral Deconvolution (GSD)", *Spectroscopy Europe* vol 23 (1), 2010
[47] Smith, S. G., & Goodman, J. M. (2010). Assigning Stereochemistry to Single Diastereoisomers by GIAO NMR Calculation: The DP4 Probability. Journal of the American Chemical Society, 132(37), 12946–12959. https://doi.org/10.1021/ja105035r
[48] Grimblat, N., Zanardi, M. M., & Sarotti, A. M. (2015). Beyond DP4: an Improved Probability for the Stereochemical Assignment of Isomeric Compounds using Quantum Chemical Calculations of NMR Shifts. The Journal of Organic Chemistry, 80(24), 12526–12534. https://doi.org/10.1021/acs.joc.5b02396
[49] Grimblat, N., & Sarotti, A. M. (2016). Computational Chemistry to the Rescue: Modern Toolboxes for the Assignment of Complex Molecules by GIAO NMR Calculations. Chemistry - A European Journal, 22(35), 12246–12261. https://doi.org/10.1002/chem.201601150
[50] Grimblat, N., Gavín, J. A., Hernández Daranas, A., & Sarotti, A. M. (2019). Combining the Power of J Coupling and DP4 Analysis on Stereochemical Assignments: The J-DP4 Methods. Organic Letters, 21(11), 4003–4007. https://doi.org/10.1021/acs.orglett.9b01193
[51] Mnova Verify, version 14.1.1, Mestrelab Research S.L., Spain, www.mestrelab.com, 2019
[52] ACD/Structure Verification, version 2018.1, Advanced Chemistry Development, Inc., Toronto, ON, Canada, www.acdlabs.com, 2019
[53] CMC-assist, version 2.10, Bruker, www.bruker.com, 2019
[54] Mnova Stereofitter, Mestrelab Research S.L., Spain, www.mestrelab.com, 2019
[55] Sarotti, A. M. (2013). Successful combination of computationally inexpensive GIAO 13C NMR calculations and artificial neural network pattern recognition: a new strategy for simple and rapid detection of structural misassignments. Organic & Biomolecular Chemistry, 11(29), 4847. https://doi.org/10.1039/c3ob40843d
[56] Zanardi, M. M., & Sarotti, A. M. (2015). 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, 80(19), 9371–9378. https://doi.org/10.1021/acs.joc.5b01663
[57] Zhang, C., Idelbayev, Y., Roberts, N., Tao, Y., Nannapaneni, Y., Duggan, B. M., … Gerwick, W. H. (2017). Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. Scientific Reports, 7(1). https://doi.org/10.1038/s41598-017-13923-x