Perspectives in Magnetic Resonance

If machines can learn, who needs scientists?

Jeffrey C. Hoch

UConn Health, Farmington, CT 06030-3305, USA

ARTICLE INFO

ABSTRACT

Machine learning has been used in NMR in for decades, but recent developments signal explosive growth is on the horizon. An obstacle to the application of machine learning in NMR is the relative paucity of available training data, despite the existence of numerous public NMR data repositories. Other challenges include the problem of interpreting the results of a machine learning algorithm, and incorporating machine learning into hypothesis-driven research. This perspective imagines the potential of machine learning in NMR and speculates on possible approaches to the hurdles.

© 2019 Published by Elsevier Inc.

1. Introduction

Machine Learning (ML) is here, and it has captured the public's attention. News worthy examples are algorithms that beat the best human experts at ancient games like chess and Go, but also modern computer games like Starcraft (Fig. 1). More utilitarian and industrial applications abound, notably for autonomous vehicles but also financial engineering (e.g. dynamic pricing), drug design, and of course image classification. Large Internet companies (Google, Apple, Microsoft, Facebook) rely heavily on ML for their digital assistants (Siri, Google Assistant, Alexa), and have made important developments in the infrastructure, both software and hardware, to support large ML applications. The TensorFlow software library from Google reduced the computational cost of back-propagation for training neural nets, and Google's Machine Learning Crash Course has proven to be a useful entrée into ML for thousands of developers.

Closer to home, the Google Deep Mind team made a recent splash [1,2] in the Critical Assessment of Structure Prediction (CASP) competition [3]. This "shared task" competition challenges participants to predict protein structures given only the amino acid sequence. CASP has been run annually for 25 years, and has played a significant role in advancing the field of protein structure prediction. The Google team entered the competition for the first time in 2018, and immediately established themselves as the leading performers. Their success no doubt benefitted from Google's deep pockets and their computational resources, nevertheless the results are undeniable. For 43 challenge targets consisting only of sequence information, Google's DeepMind team using their Alpha-

Fold software outscored the other 97 competing teams 25 times. The nearest competitor achieved the top score on only 3 of these challenge targets. For a broader set of targets that includes individual domains of target proteins, AlphaFold achieved first or second ranking 40 times for 124 targets. The closest competitor attained first or second rank for 24 targets. A striking measure of the success of AlphaFold is not only that it outranked the competitors, but it also did so by substantial margins (Fig. 2). AlphaFold was responsible for every instance in which the best score exceeded the second-best score by more than 10.

Even before the recent success of AlphaFold, some commentators were questioning whether the advent of ML heralds changes in the way science is conducted, whether it will usurp the role of theory [4] or upend hypothesis-driven research altogether. Here I explore some of the limitations of ML to argue that scientists will remain indispensable, and that ML will not supplant established approaches to science but become a powerful new tool in the conduct of hypothesis-driven research. In fact, this role for ML is already firmly established, and we need look no further than NMR and the pages of this Journal for evidence.

2. ML in NMR

A definitive accounting of ML appearing in JMR is complicated by evolving nomenclature, but a casual search using the terms "machine learning", "neural network", and "artificial intelligence" yields more than 50 citations dating from 1988. Broadened to include other terms that rightfully represent forms of ML, for example Principle Component Analysis (PCA), the count of ML applications in JMR reaches into the hundreds.
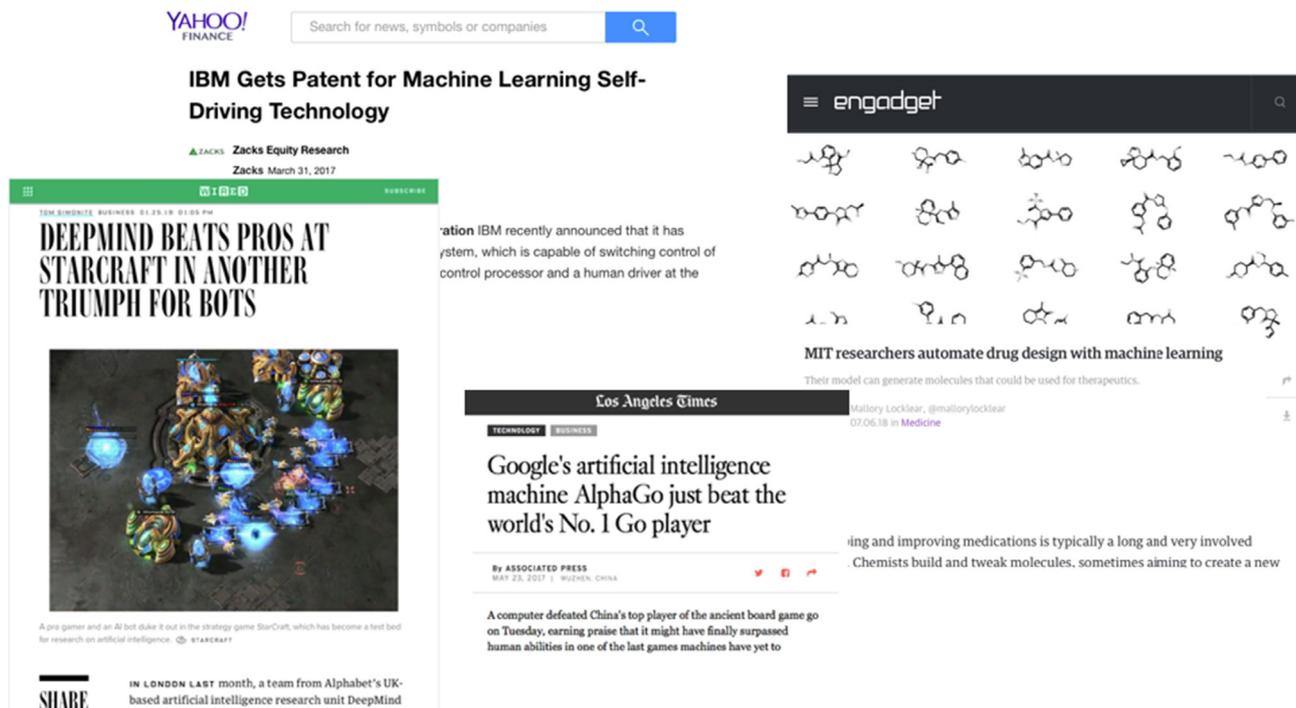
E-mail address: hoch@uchc.edu
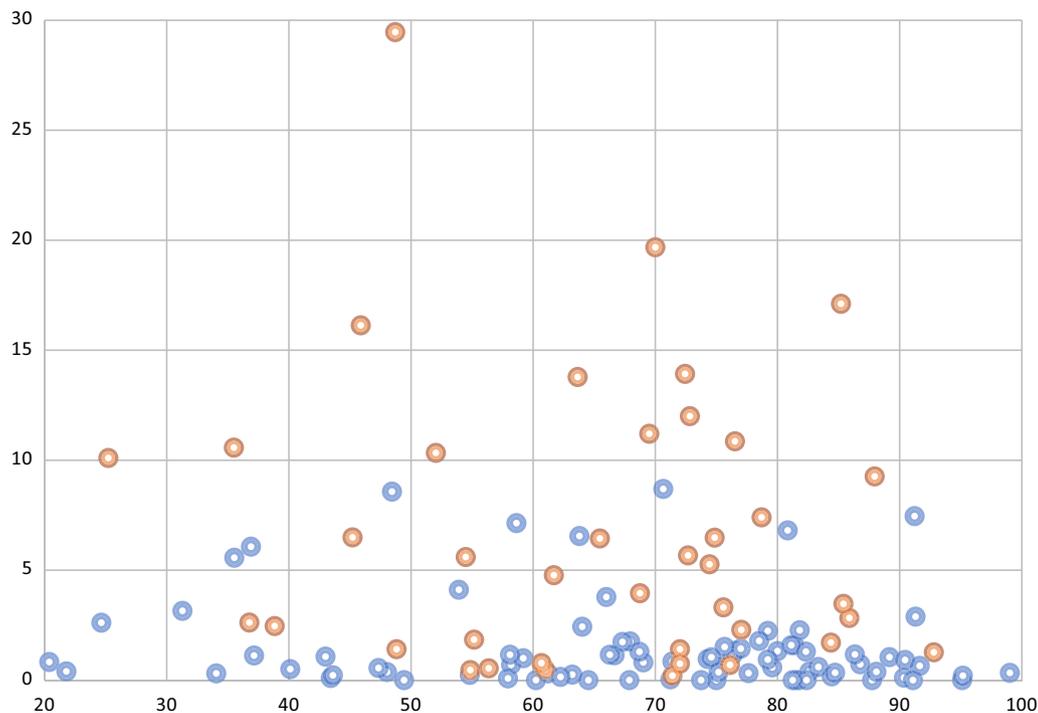
Fig. 1. Machine Learning in the news.

Fig. 2. Difference between the best and second-best CASP 13 "GDT_TS" scores (vertical) vs. the best score (horizontal). Results for Google DeepMind are depicted in orange.

Though ML is often described as an approach to "artificial intelligence", the applications of ML in NMR are frequently more prosaic, tending toward utilitarian. Widely-used examples in protein NMR are SHIFTX2 [5], used to predict ^{1}H, ^{13}C, and ^{15}N chemical shifts from protein structures, and TALOS-N [6], which predicts protein backbone conformation from chemical shifts. ML has made major inroads in the analysis of NMR metabolomics data [7,8]. More recently ML has been used to recapitulate expensive Density Functional Theory calculations [9], achieving remarkable accuracy

while dramatically reducing the computational cost. These examples demonstrate that far from replacing scientists, an important role for ML is to provide powerful new tools to scientists.

3. Interpretability, inverse vs. forward modeling, and hypothesis testing

As powerful as ML is proving to be, a weakness is the so-called "interpretability" problem. Employing an ANN, for example, results
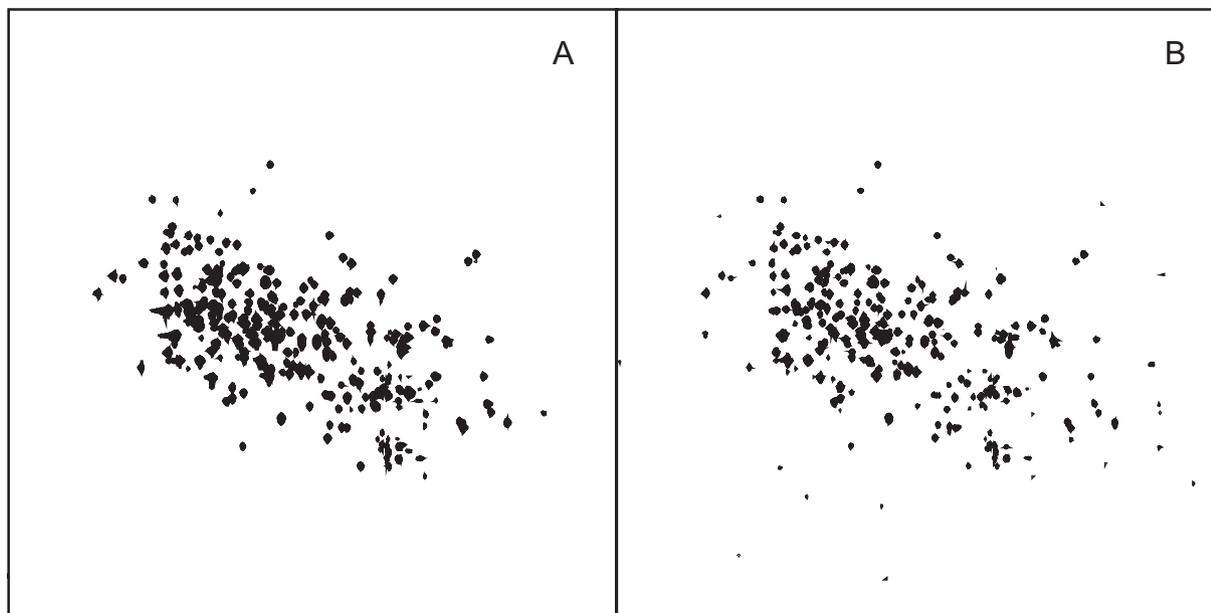
Fig. 3. MaxEnt spectral reconstructions using NUS data comprising 11% of the uniform Nyquist grid. A. ML-optimized schedule. B. Random schedule.
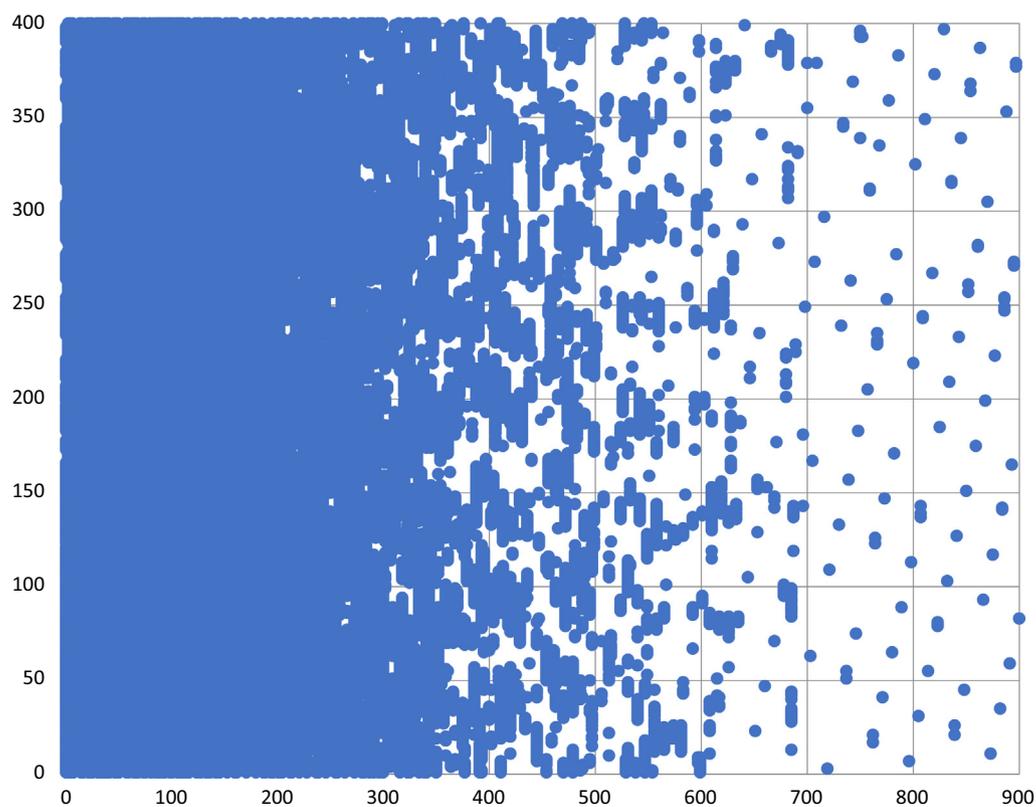

Fig. 4. ML-optimized sampling schedule comprising 11% of the Nyquist grid (used in Fig. 3A).

in numerical weights for the nodes and connections, and while these values enable the use of ANNs as "black boxes", they are otherwise opaque and don't provide useful insight into how an ANN "learned".

ML applications in NMR have demonstrated their use for both inverse modeling (e.g. deriving protein conformation from chemical shifts) and forward modeling (predicting chemical shifts from structure). Combining the two opens the possibility of adversarial training for ML, but also points at a possible approach to hypothesis testing as a means to overcome the interpretability problem. Where we have useful theory, predicting expected NMR parameters from a model can be used to generate "mock data" that can then be used to challenge a ML algorithm, testing both the hypothesis (the model) and the ML algorithm. When a ML algorithm that
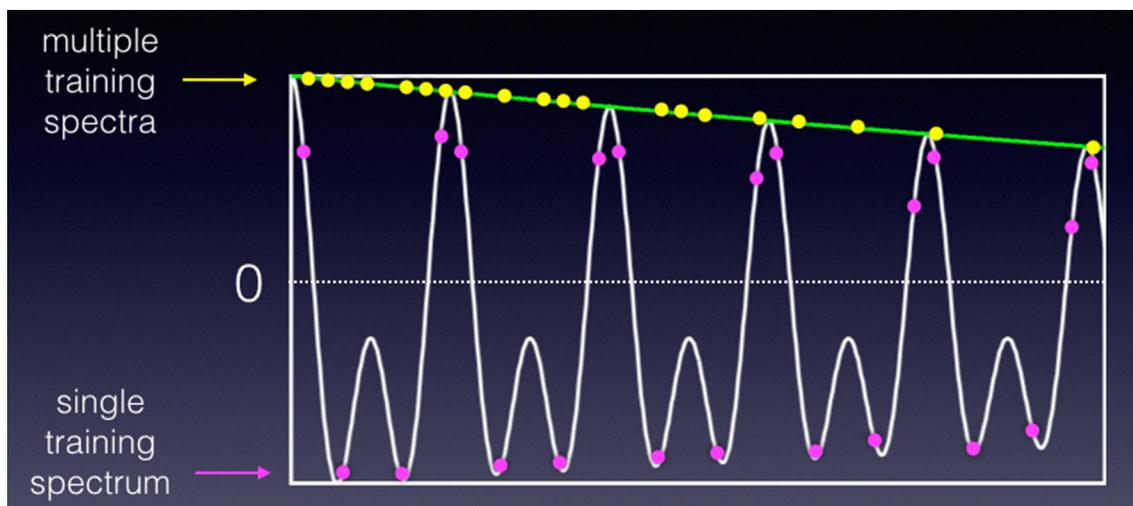
Fig. 5. ML-optimized NUS schedules optimized for a single training data set tend to concentrate at the peaks of the time-domain data. Schedules optimized for multiple data sets tend to mirror the decay of the signal envelope.

is able to accurately detect/distinguish "ground truth" (e.g. empirical data for known, standard samples) is unable to distinguish mock data from empirical data, the model/hypothesis is validated.

4. How much data?

Years ago our group turned to ML for a solution to a nonuniform sampling (NUS) problem (K.-B. Li and J.C. Hoch, unpublished). The problem we posed was this: given a reference multidimensional spectrum computed by conventional discrete Fourier transformation (DFT) from a uniformly-sampled data set, what subset of the samples in the indirect dimensions (for a given fraction of the original number of samples), when fed to a maximum entropy algorithm to compute the spectrum for the NUS data, yields a spectrum closest to the spectrum obtained using uniform sampling and DFT? When trained against a single spectrum, the results are excellent, and much better than randomly choosing a subset. Results for ML-optimized and random sampling employing 11% of the uniformly-sampled data are shown in Fig. 3(A and B), together with a representative ML-optimized sampling schedule (Fig. 4). Though the ML-optimized sampling schedule works quite well, better than random sampling for recovering this spectrum, it does worse than random sampling when used to recover a similar spectrum with some of the peaks located at different frequencies. Close inspection of the sampling schedule reveals a level of regularity at long evolution times, regularity that will result in sampling artifacts. Apparently what ML learned, in this instance, is to place the artifacts where the peaks are located. In contrast, if we train the ML algorithm against *multiple* spectra, we get sampling schedules that exhibit less regularity, with a sampling density that mirrors the decay of the signal envelope (so-called envelope-matched sampling) instead of closely hewing to the detailed amplitude changes for a specific data set (beat-matched sampling) (Fig. 5). Since exponentially-biased sampling schedules are easier and cheaper to compute than ML-optimized schedules that look very similar, we didn't pursue ML as an approach to finding optimal sampling schemes.

This ML exercise did provide a valuable lesson, however, on the importance of training data. The poverty of the ML-optimized schedule determined from a single training data set points to the need for "coverage" in the training data – the training data should contain examples of the full range of possibilities, both to identify features common to data sets as well as to ensure detection of

specific instances. The results from training against multiple data sets reveal that the predominant feature common to possible data sets is the shape of the signal envelope. Whether ML algorithms are capable of identifying data types not represented in the training data remains somewhat controversial – to do so is considered "innovation". The ability of AlphaGo to discover a novel move [10] in the game of Go is cited as an example of innovation by ML.

Attention to coverage in the training data has informed a number of applications of ML in NMR structural biology. For example, Talos-N [6], a neural net for predicting dihedral angles and secondary structure from chemical shifts, was trained against a database composed of peptide fragments in the Protein Data Bank (PDB) for which chemical shifts are available in the Biological Magnetic Resonance Data Bank (BMRB). Though it's unlikely that all possible protein folds are represented in the PDB, it's very likely that all feasible (thermally accessible) conformations of mainchain dihedral angles are represented.

The notion that it is difficult to discern truths from sets of examples that don't contain pertinent examples is not new. The philosopher of science Karl Popper used this as the basis of his critique [11] of the social sciences when historical records are used to try to predict the future. Popper's argument remains relevant today.

5. More data!

Access to curated training data represents a significant hurdle to wider application of ML in NMR. Commercial NMR databases, mainly for small molecules, are quite extensive, but require sometimes expensive subscriptions. Publically accessible databases, notably BMRB [12] and The Human Metabolome Database [13] (HMDB), are valuable resources but the number of entries pales in comparison to the amount of data available in collections such as ImageNet (http://www.image-net.org, >10^7 images), used to train image recognition ML algorithms.

Simply put, more data is needed. A tremendous amount of NMR data is collected that never becomes publicly accessible; precise numbers are elusive, but conservatively it seems safe to say that the majority of data collected is unavailable for use by the broader community. Presently BMRB hosts around 10^3 time domain data sets for biological macromolecules. For comparison, this amounts to just a few days output from all the NMR spectrometers in the world dedicated to biomolecular NMR. Where there are relevant existing databases, additional efforts are needed to lower the

barriers to deposition. Additional incentives in the form of requirements from publishers for deposition of primary, derived, and supporting data in public data archives could have a large beneficial impact, not only for applications of ML, but also for the reproducibility of published studies. Automatic upload of empirical NMR data in real time to a shared online repository, once unimaginable, now does not face any technical limitations.

In some instances, ML could be deployed using forward modeling, for example prediction of biomolecular NMR spectra from molecular structures. This approach is already informing methods for using predicted chemical shifts to infer structure, notably from unassigned carbon-carbon correlation spectra [14]. The use of ML to replace non-Fourier methods of spectrum analysis could proceed without large empirical data sets, using simulated time domain data[1]. After all, the discrete Fourier Transform can be described as a least-squares fit of a basis set (the Fourier basis) to time domain data; ML is quite successful at regression.

There is also need for new public data repositories for areas not covered by existing archives. A new initiative called the "Local Spectroscopy Database Infrastructure" (LSDI), housed within The Materials Project (materialsproject.org/) will be launched in summer 2019. This resource will provide DFT-computed ^{29}Si chemical shielding sensors for crystalline materials, with the expectation that the effort will be expanded to encompass additional nuclei. The success of LSDI depends on access to abundant empirical data.

6. Concluding remarks

This perspective gives only an anecdotal view of ML applied to NMR. Yet it is clear that ML is well entrenched in NMR, and recent advances suggest many exciting applications lie ahead. The lesson from NMR is that the primary significance of ML will be as a source of new tools that scientists will use to accelerate their discovery of knowledge, rather than as a replacement for scientists. Furthermore, scientists remain essential as curators of the data used to train ML algorithms. Machines can "learn", but not without scientists to craft the lessons.

Acknowledgements

References

[1] R.F. Service, Google's DeepMind aces protein folding, Science (2018). https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding.

[2] I. Sample, Google's DeepMind predicts 3D shapes of proteins, The Guardian, <https://www.theguardian.com/science/2018/dec/02/google-deepminds-ai-program-alphafold-predicts-3d-shapes-of-proteins>.

[3] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)–round x, Proteins 82 (Suppl 2) (2014) 1–6, https://doi.org/10.1002/prot.24452.

[4] F. Mazzocchi, Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science, EMBO Rep. 16 (2015) 1250–1255, https://doi.org/10.15252/embr.201541001.

[5] B. Han, Y. Liu, S.W. Ginzinger, D.S. Wishart, SHIFTX2: significantly improved protein chemical shift prediction, J. Biomol. NMR 50 (2011) 43–57, https://doi.org/10.1007/s10858-011-9478-4.

[6] Y. Shen, A. Bax, Protein structural information derived from NMR chemical shift with the neural network program TALOS-N, Methods Mol. Biol. 1260 (2015) 17–32, https://doi.org/10.1007/978-1-4939-2239-0_2.

[7] J. Kikuchi, K. Ito, Y. Date, Environmental metabolomics with data science for investigating ecosystem homeostasis, Prog. Nucl. Magn. Reson. Spectrosc. 104 (2018) 56–88, https://doi.org/10.1016/j.pnmrs.2017.11.003.

[8] J. Xia, N. Psychogios, N. Young, D.S. Wishart, MetaboAnalyst: a web server for metabolomic data analysis and interpretation, Nucleic Acids Res. 37 (2009) W652–W660, https://doi.org/10.1093/nar/gkp356.

[9] F.M. Paruzzo et al., Chemical shifts in molecular solids by machine learning, Nat. Commun. 9 (2018) 4501, https://doi.org/10.1038/s41467-018-06972-x.

[10] M. McFarland, in: Washington Post. 2016.

[11] K. Popper, The Poverty of Historicism, Routledge and Kegan Paul, 1957.

[12] E.L. Ulrich et al., BioMagResBank, Nucleic Acids Res. 36 (2008) D402–D408, https://doi.org/10.1093/nar/gkm957.

[13] D.S. Wishart et al., HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res. 46 (2018) D608–D617, https://doi.org/10.1093/nar/gkx1089.

[14] J.M. Courtney et al., Experimental protein structure verification by scoring with a single, unassigned NMR spectrum, Structure 23 (2015) 1958–1966, https://doi.org/10.1016/j.str.2015.07.019.

---

[1] Note added in proof: Hansen recently published just such a study, Hansen, D.F., Using Deep Neural Networks to Reconstruct Non-uniformly Sampled NMR Spectra. J Biomol NMR (2019). https://doi.org/10.1007/s10858-019-00265-1.