HHS Public Access

Author manuscript

J Magn Reson. Author manuscript; available in PMC 2020 September 01.

Published in final edited form as:

J Magn Reson. 2019 September ; 306: 162–166. doi:10.1016/j.jmr.2019.07.044.

If Machines Can Learn, Who Needs Scientists?

Jeffrey C. Hoch

UConn Health, Farmington, CT 06030-3305 USA

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.

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 (Fig. 2). 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 AlphaFold software outscored the other 97 competing teams 25 times. The nearest competitor achieved the top score on only 3 of these challenge targets.

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.

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.

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.

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 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 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.

How much Data?

Years ago we turned to ML for a solution 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, 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 samping) instead of closley 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 schudules 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 main-chain 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.

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, we need more data. 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. 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.

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.

Concluding Remarks

Machine Learning 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.

Acknowledgements

Kuo-Bin Li conducted early experiments using ML for the NUS problem in my lab at the Rowland Institute for Science in the mid-1980’s. I thank Sophia Hayes for useful discussions on NMR databases and the Local Spectroscopy Data Infrastructure project, David Donoho, Hatef Monajemi, and Hamid Eghbalnia for advice on ML, and Guy Montelione for insights on the CASP 2018 results. I’m grateful to the ANZMAG community for their hospitality in Australia and for comments on the ideas presented here. Support from the US National Institutes of Health, National Institute of General Medical Sciences is gratefully acknowledged (grants P41GM111135, R01GM123249, R01GM109046).

References

1. Service RF Google’s DeepMind aces protein folding. Science (online) (2018). <https://www.sciencemag.org/news/2018/12/google-s-deepmind-aces-protein-folding>.
2. Sample I 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. Moult J, Fidelis K, Kryshtafovych A, Schwede T & Tramontano A Critical assessment of methods of protein structure prediction (CASP)--round x. Proteins 82 Suppl 2, 1–6, doi:10.1002/prot.24452 (2014).

4. Mazzocchi F Could Big Data be the end of theory in science? A few remarks on the epistemology of data-driven science. EMBO Rep 16, 1250–1255, doi:10.15252/embr.201541001 (2015). [PubMed: 26358953]
5. Han B, Liu Y, Ginzinger SW & Wishart DS SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50, 43–57, doi:10.1007/s10858-011-9478-4 (2011). [PubMed: 21448735]
6. Shen Y & Bax A Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. Methods Mol Biol 1260, 17–32, doi:10.1007/978-1-4939-2239-0_2 (2015). [PubMed: 25502373]
7. Kikuchi J, Ito K & Date Y Environmental metabolomics with data science for investigating ecosystem homeostasis. Prog Nucl Magn Reson Spectrosc 104, 56–88, doi:10.1016/j.pnmrs.2017.11.003 (2018). [PubMed: 29405981]
8. Xia J, Psychogios N, Young N & Wishart DS MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37, W652–660, doi:10.1093/nar/gkp356 (2009). [PubMed: 19429898]
9. Paruzzo FM et al. Chemical shifts in molecular solids by machine learning. Nat Commun 9, 4501, doi:10.1038/s41467-018-06972-x (2018). [PubMed: 30374021]
10. McFarland M in Washington Post (2016).
11. Popper K The Poverty of Historicism. (Routledge and Kegan Paul, 1957).
12. Ulrich EL et al. BioMagResBank. Nucleic Acids Res 36, D402–408, doi:10.1093/nar/gkm957 (2008). [PubMed: 17984079]
13. Wishart DS et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res 46, D608–D617, doi:10.1093/nar/gkx1089 (2018). [PubMed: 29140435]

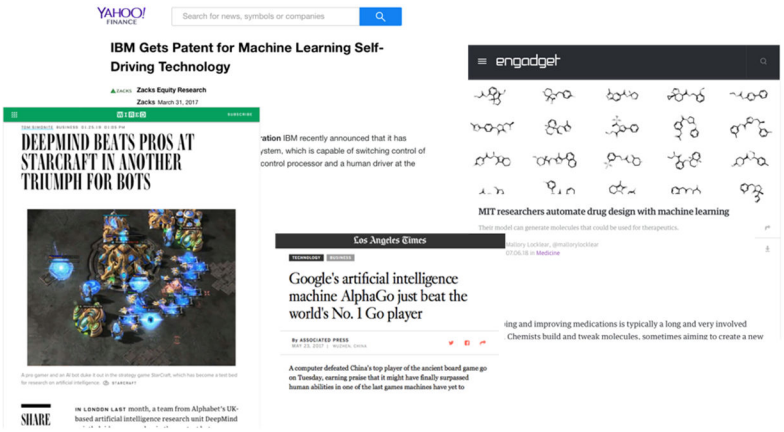

Figure 1.
Machine Learning in the news.

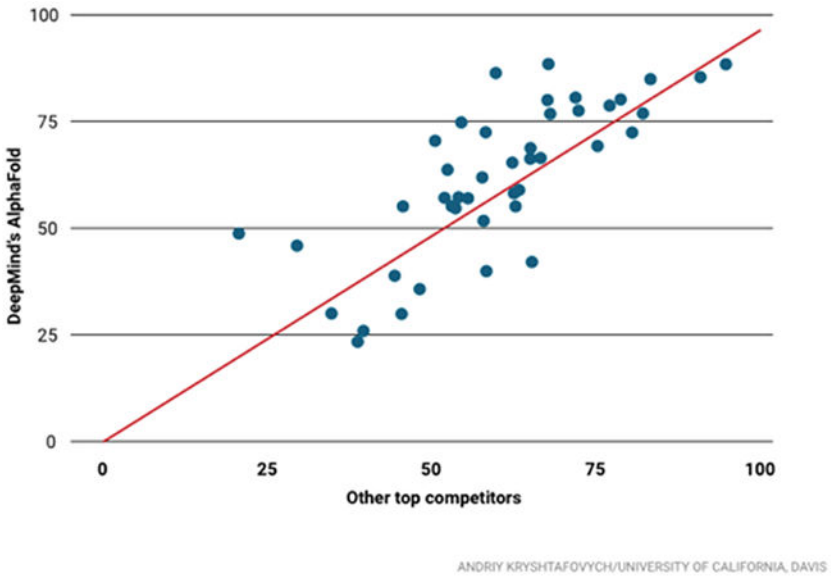

Figure 2.
CASP scores for Google DeepMind (vertical) vs. the best score for the other competitors (horizontal). Adapted from *Science* online[1].

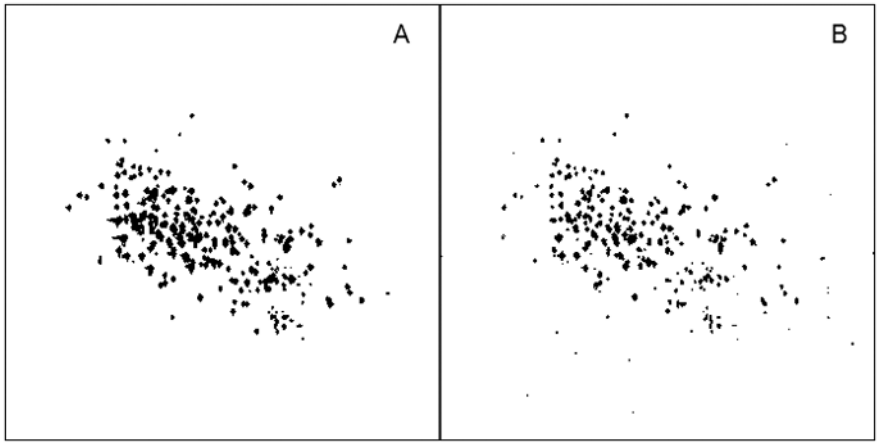

Figure 3.
MaxEnt spectral reconstructions using NUS data comprising 11% of the uniform Nyquist grid. A. ML-optimized schedule. B. Random schedule.

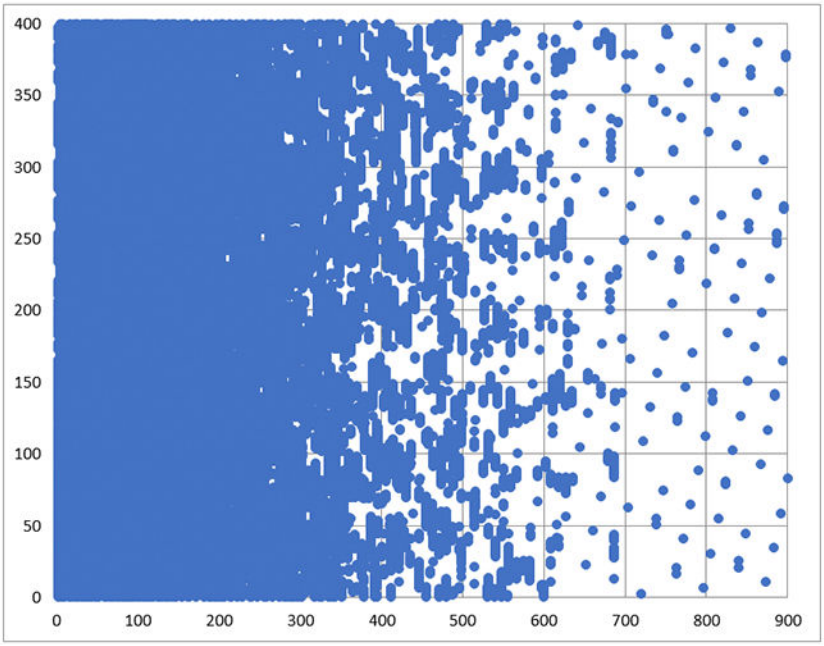

Figure 4.
ML-optimized sampling schedule comprising 11% of the Nyquist grid (used in Fig. 3A).

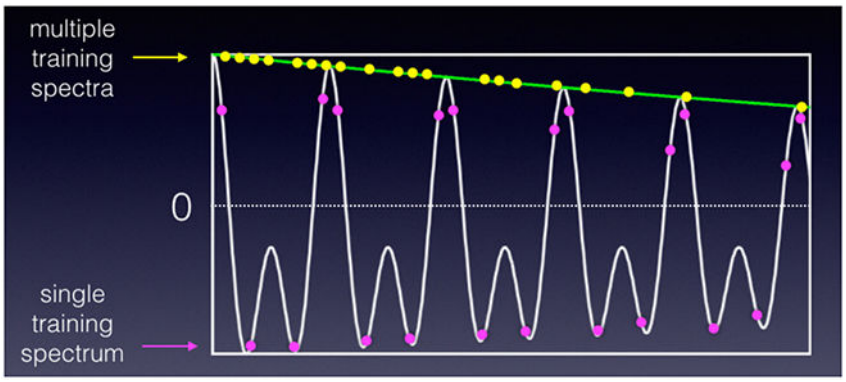

Figure 5.

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.