Message Passing Neural Networks for Predicting ^{1}H and ^{19}F Chemical Shifts

Adam P. Jones[a], Santi Ponte[b], Isaac Iglesias[b], Nicola Tonge[b], David Williamson[b], Vera Martos[b], Till Orth[c], Torsten Schoenberger[c], Carlos Cobas[b], Katharina T. Huber[a], E. Kate Kemsley[a,b*]

[a] University of East Anglia, Faculty of Science, Norwich, UK
[b] Mestrelab Research S.L., Santiago de Compostela, Spain
[c] Bundeskriminalamt, Wiesbaden, Germany
* Author for correspondence: kate.kemsley@mestrelab.com

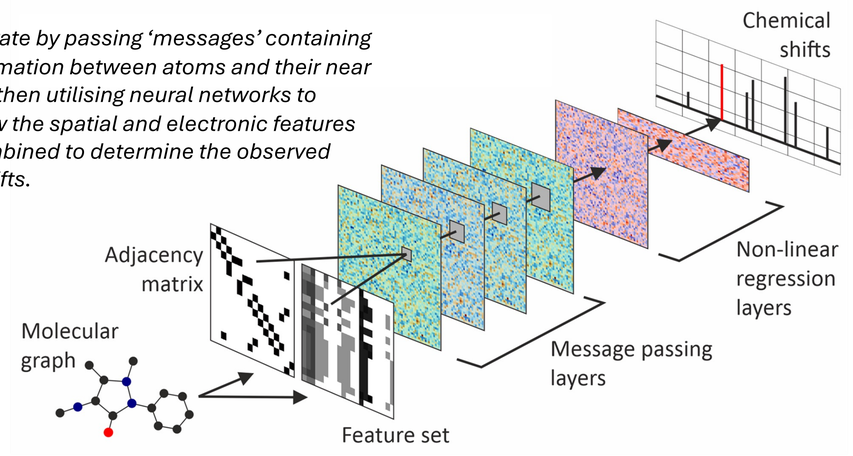
Mestrelab Research

University of East Anglia

Deep learning for chemical shift prediction

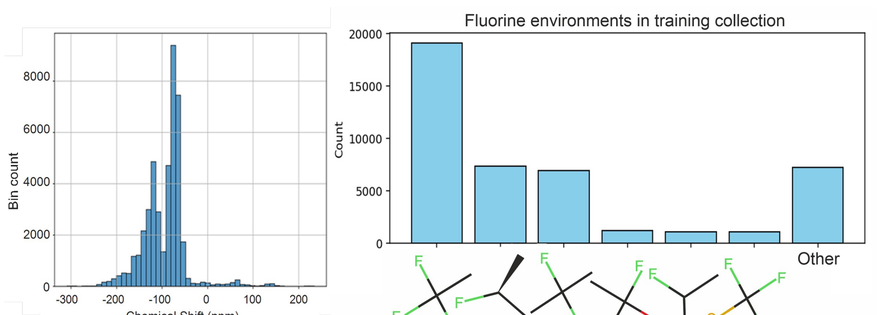
By representing molecules as molecular graphs, Message Passing Neural Networks (MPNNs) have proven to be powerful tools for predicting NMR chemical shifts. By encoding atoms as nodes, their bonds as edges and atomic information (e.g. valency, hybridisation) as node features, these deep learning models can capture complex relationships and make predictions that traditional methods such as HOSE (Hierarchically Ordered Spherical Environment) Code may struggle to carry out effectively.

MPNNs operate by passing 'messages' containing feature information between atoms and their near neighbours, then utilising neural networks to optimise how the spatial and electronic features are best combined to determine the observed chemical shifts.

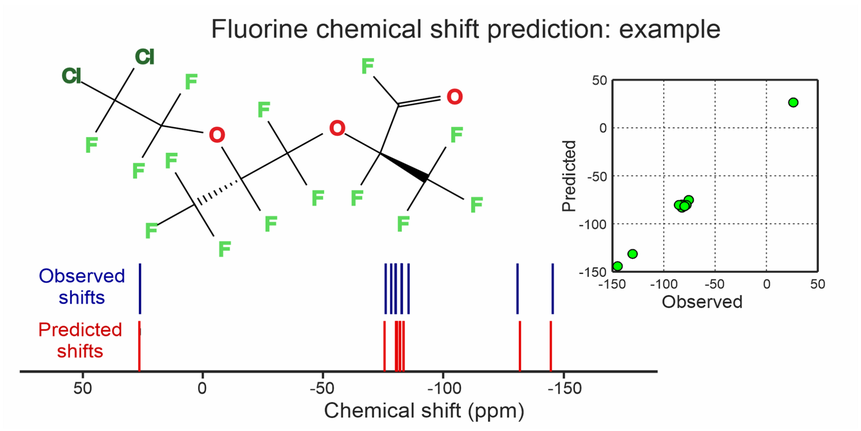


Predicting fluorine (^{19}F) chemical shifts

An MPNN was trained using >14,000 labelled organo-fluorine compounds to predict ^{19}F chemical shifts. The distribution of these is strongly centred between -50 and -150 ppm, although with a significant tail outside this range. This suggests a diverse range of environments are possible for fluorine, even if these are sparsely represented in the training collection. This emphasises the requirement for any predictive model to generalise well.

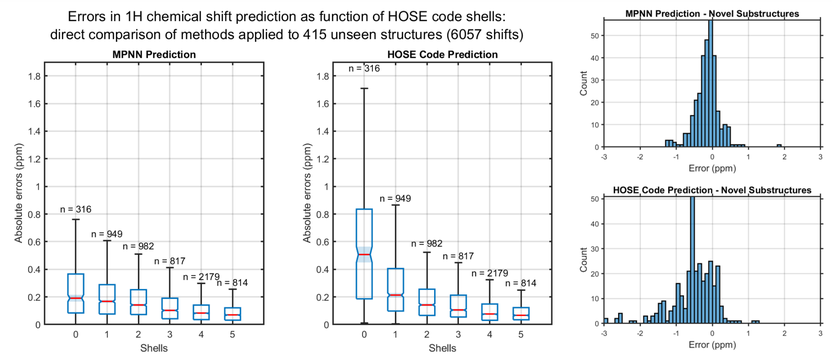


The constitution of fluorine environments in the dataset, with the majority comprising only of carbon, other fluorine atoms and hydrogen, presents a challenge, making prediction of less common neighbourhoods ('Other') more difficult. Nevertheless, an ensemble of MPNNs yielded an effective model with a median absolute error in prediction of ~2.2 ppm. Relative to the chemical shift range, this is comparable performance to MPNN models for other active nuclei (^{1}H, and ^{13}C reported in [1]).



Predicting ^{1}H chemical shifts: MPNNs cf. HOSE Code

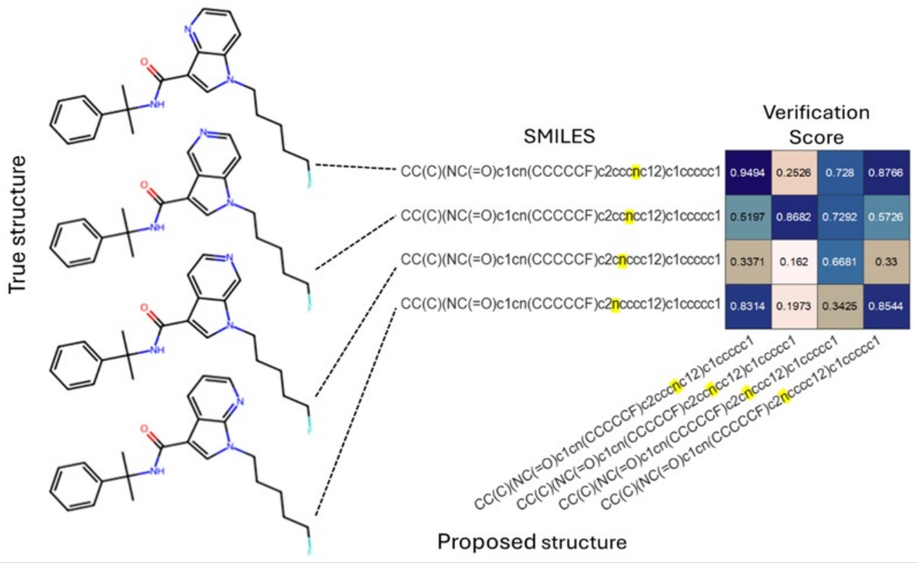
An ensemble of MPNNs was trained for ^{1}H chemical shifts from >30,000 molecular structures. The median absolute prediction error obtained from a test collection of ~400 structures of novel psychoactive substances was 0.10 ppm. This accuracy was compared with that by HOSE Code, in which an atom's environment is encoded as concentric 'shells' of neighbourhood information and predictions obtained by matching these, fully or partially, to a reference database which also contains chemical shift annotations.



The boxplots summarise the absolute prediction errors, grouped by the number of HOSE Code shells used, allowing for direct comparison of the two methods. The most notable difference is seen for uncommon or novel structures (groups with low numbers of matched shells) for which it is clear that MPNNs offer a substantial performance advantage.

Accurate prediction leads to effective verification

A primary motivation for accurate chemical shift prediction is their downstream use in structural verification. We previously reported the use of ^{13}C shifts predicted by MPNNs to make automated assignments and verification [1]. Here we show that this approach is effective for ^{1}H also, in an example taken from the novel psychoactives test set. These four structures are positional isomers of Cumyl-5F-P7AICA, a synthetic cannabinoid of the class P4/5/6/7AICA. They differ only in the location of a nitrogen on an aromatic ring.



In cross-verification, a probabilistic assignment score is calculated for each possible pairwise match of the predicted to observed chemical shifts from all structures. The heatmap shows the outcome: in all cases, the highest score in each row is found on the diagonal, indicating that the closest match is from the corresponding proposed structure.

Reference

[1] D. Williamson et. al. Chemical shift prediction in 13C NMR spectroscopy using ensembles of message passing neural networks (MPNNs). J Magn Res. Volume 368, 2024, 107795

-Mestrelab Confidential-