Chemical Shift Prediction using Message Passing Neural Networks

Carlos Cobas[2], Isaac Iglesias[2], E. Kate Kemsley[1,2,*], Marcel Lachenmann[2], Santi Ponte[2], Nicola Tonge[2], David Williamson[2]

[1]University of East Anglia, Norwich Research Park, NR7 6TJ, United Kingdom
[2]Mestrelab Research SL, C/Feliciano Barrera, 9B-Bajo, 15706, Santiago de Compostela, Spain
*Author for correspondence

Mestrelab Research

University of East Anglia

Why do we need chemical shift prediction?

Accurate chemical shift predictions are the foundation of spectral assignment and automated structural verification in NMR spectroscopy. We are using the modern AI approaches of artificial neural networks, deep learning, and ensembles to improve the prediction of chemical shifts from molecular properties. This is known as the 'forward' problem:

Properties of atoms in a molecule
e.g. valency, hybridization,...

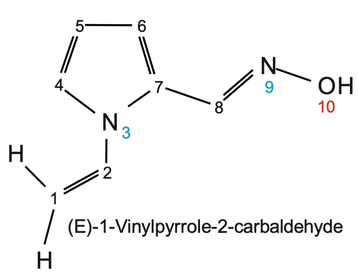
PREDICTION
- Spectral assignment
- Spectral verification

Chemical shifts of active nuclei
(^{13}C, ^{1}H)

"Predictors", "Regressors", "Explanatory", "Independent variables" — *the "Forward" problem* → *"Target", "Outcome", "Regressand", "Response", "Dependent variable"*

Message passing neural networks (MPNNs)

MPNNs are a form of graph convolutional neural network and are a promising approach for tackling this problem. They provide a natural framework for handling molecular structures as graphs, with atoms represented as nodes, bonds as edges, and the graph connectivity described by an adjacency matrix. Associated with each node are the various physical properties of each atom, which can be concatenated to give a feature matrix for the structure as a whole:



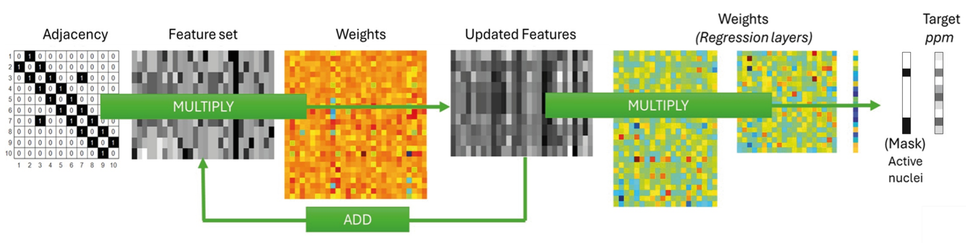
Ajacency matrix (Atom connectivity)

	1	2	3	4	5	6	7	8	9	10
1		■								
2	■		■							
3		■		■			■			
4			■		■					
5				■		■				
6					■		■			
7			■			■		■		
8							■		■	
9								■		■
10									■	

Feature set (some typical variables)

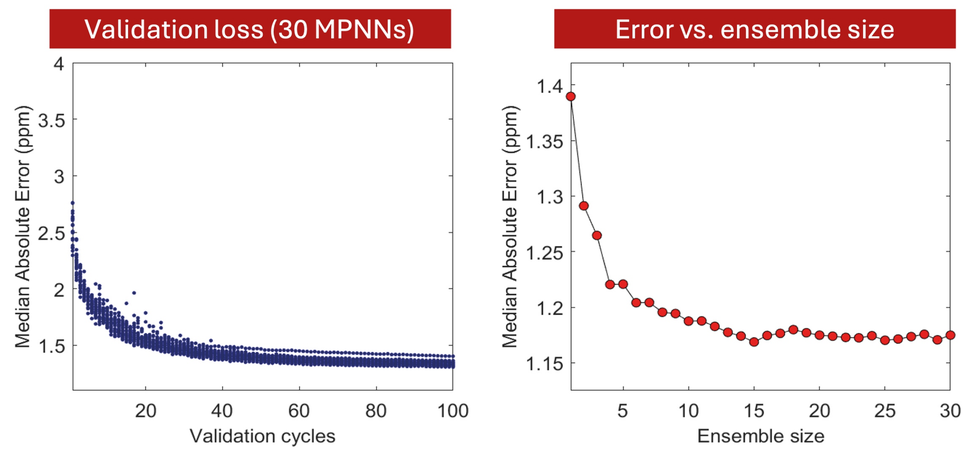
Atomic Num	Degree	Valency	isAromatic	TotalNumHs
6	1	4	0	2
6	2	4	0	1
7	3	3	1	0
6	2	4	1	1
6	2	4	1	1
6	2	4	1	1
6	3	4	1	0
6	2	4	0	1
7	2	3	0	0
8	1	2	0	1

A characteristic of MPNNs is their simultaneous utilization of node feature and connectivity information via multiple, iterative 'message passing' layers. This is illustrated in the schematic of the network architecture below, in which the grayscale heatmaps represent the predictor and outcome variables, and the RGB heatmaps the learnable model parameters.



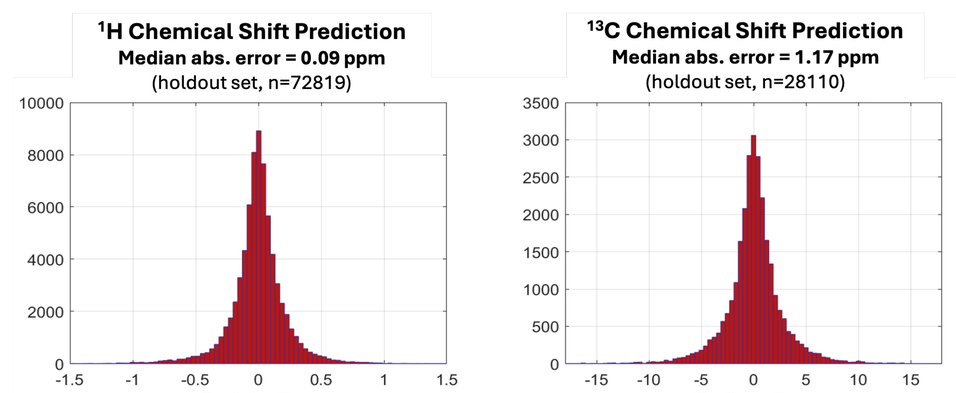
An ensemble approach

Ensembles of MPNNs were trained using large, well-curated collections of diverse molecular structures that were comprehensively annotated with experimentally observed proton (^{1}H) and carbon (^{13}C) chemical shifts. An ensemble approach was adopted to achieve error reduction via the 'wisdom of crowds', since main sources of stochastic training variance are the partitioning and weight initialisations. The MPNNs were trained using the 'Adam' optimizer, widely-used for deep learning architectures, and the model loss was evaluated from an internal validation partition at regular intervals. The validation loss as a function of training cycles is illustrated below for an ensemble of 30 MPNNs. The prediction error from the aggregated outcomes reduces with ensemble size, reaching a stable minimum of ~1.2 ppm median absolute error for ^{13}C chemical shift prediction.

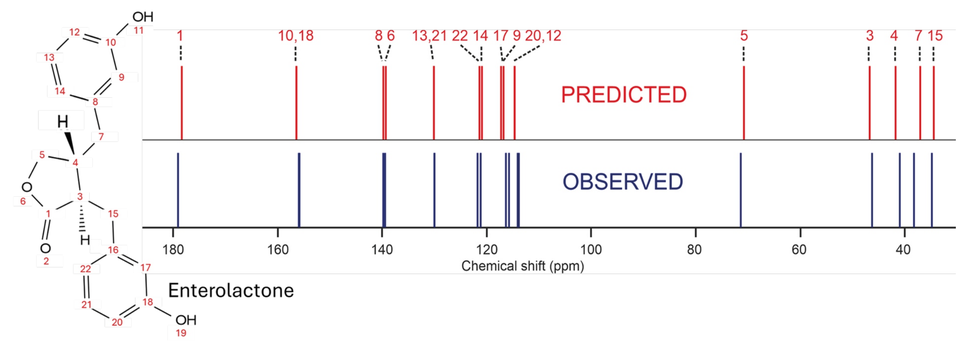


^{13}C and ^{1}H chemical shift prediction

An analogous ensemble was trained for 1H chemical shifts. The relative prediction performance is remarkably similar for both nuclei as seen by the error histograms. These are symmetric around 0 ppm and fat-tailed in comparison with a normal distribution. The collection of prediction errors for an individual structure can be treated as a sample from these distributions.



The distributions of the mean absolute error at the structure level can also be characterised. For a sample size n = 18, the 5th percentile for ^{13}C predictions is 0.6 ppm. This is exemplified below using an item from the holdout set; we can expect 1 in 20 molecules to be predicted with a similar accuracy.



Acknowledgements: Carlos C., Isaac I. and Nicola T. thank the SmartGlobalLab Project: A global pharmaceutical data marketplace (No. 2021/C005/00150505) under the 2021 Call for grants for research and development projects in artificial intelligence and other digital technologies and their integration into value chains (C005/21-ED), promoted by Red.es and carried out under the Recovery, Transformation and Resilience Plan - Funded by the European Union - NextGenerationEU.

Financiado por la Unión Europea NextGenerationEU

Gobierno de España · Ministerio para la Transformación Digital y de la Función Pública · Secretaría de Estado de Digitalización e Inteligencia Artificial · red.es · Plan de Recuperación, Transformación y Resiliencia