



estrelab Research

From instruments and analytical data
to knowledge and decisions

RESEARCH HIGHLIGHTS

Graph Convolutional Neural Networks (GCNNs) for accurate NMR ^{19}F -NMR chemical shift prediction.

Kate Kemsley

Overview

Accurate prediction of NMR chemical shifts is the foundation for spectral assignment, structural elucidation and automated structural verification. In the work reported here, we are developing Graph Convolutional Neural Network (GCNNs) to represent molecular structures and learn their relationships with ^{19}F -NMR chemical shifts. In this framework, molecules are treated as graphs in which atoms are nodes and bonds are edges. Node features such as valency, electronegativity, and hybridisation capture each atom's chemical environment, allowing the network to infer relationships between structure and NMR shift directly from data.

The GCNN exploits 'message passing', where each node exchanges information with neighbouring atoms to learn its extended chemical context, followed by prediction, where aggregated node features are processed through a neural network trained on experimental chemical shifts. Such networks are referred to as Message Passing neural Networks (MPNNs). The approach captures subtle dependencies between molecular structure and NMR properties that conventional empirical methods can miss.



Results

A model for ^{19}F shift prediction was trained on structural and chemical shift information from more than 14,000 labelled organo-fluorine compounds. The distribution of the target chemical shifts is densest between -50 and -150 ppm, but includes long tails extending either side of this region. Further, whereas the majority of fluorine environments involve only carbon, hydrogen and other fluorine atoms, rarer environments are much more sparsely represented giving a chemically diverse, heterogeneously distributed dataset overall.

An ensemble of MPNNs achieved a median absolute error in prediction of approximately 2.2 ppm. This performance is comparable to that of ^1H and ^{13}C models we have previously reported in the literature (Williamson et al, 2024). The prediction errors are symmetrically distributed and well-behaved; however, accuracy is notably less for under-represented environments. To address this, future work will investigate a hybrid strategy in which chemical shifts calculated by DFT (density functional theory) from under-represented environments are used to supplement experimental data. The goal is to improve representation without requiring extensive new measurements or additional annotated data.

Work has also been conducted on refining the MPNN architecture. Attention mechanisms have been explored with the aim of improving the model's ability to focus on chemically relevant node interactions; however, to date, no significant improvement in prediction ability has been obtained, suggesting the feature set as it currently exists is both sufficient and necessary. Additional refinement through alternative encodings (e.g. categorical rather than one-hot) will be explored in an attempt to reduce the dimensionality of the feature set. At minimum, this should lead to improvements in training speed, but it may also mitigate the tendency to overfitting by reducing the number of parameters to optimise. Finally, a tandem model for addressing prediction of non-equivalent CF_2 shifts will be trained and integrated into the framework. Given the relatively small number of such instances (~400), this is a logical target for the rare environment augmentation approach.

Summary

The current MPNN framework achieves high accuracy for ^{19}F chemical shift prediction, commensurate with the performance obtained for other active nuclei studied. Continued development will focus on extending chemical coverage, addressing diastereotopic fluorines, and integrating the predictive tool into automated workflows for NMR-based prediction and structural verification.

Reference

D. Williamson, S. Ponte, I. Iglesias, N. Tonge, C. Cobas, E.K. Kemsley, "Chemical shift prediction in ^{13}C NMR spectroscopy using ensembles of message passing neural networks (MPNNs)" *Journal of Magnetic Resonance*, Volume 368, 2024, 107795, ISSN 1090-7807

Acknowledgements

Project title "Verification of molecular structures with AI for the safe purchase of chemical compounds" (Ref. IG408M-2025-000-000050).

Under the Resolution of the call IG408M-IA360. Development of technology for the improvement of economic and industrial activity and for the commercialization of new products and services based on Artificial Intelligence.

This operation is co-financed 25% by the Autonomous Community of Galicia's own funds and 75% by funds from the Recovery and Resilience Mechanism within the framework of the Recovery, Transformation and Resilience Plan, financed by the European Union-NextGenerationEU, within the framework of component 16, National Artificial Intelligence Strategy.