Determining NMR Spectrum Peaks from Chemical Structure

By

Aidan McCrillis and Daniel Roche

Introduction:

The goal of this project is to accurately predict the structure of a chemical from its ^{1}H NMR spectrum. This capability is essential in organic chemistry, where identifying the exact molecules produced by reactions is crucial. This identification process can be challenging, especially with large molecules that may contain numerous functional groups, making it a promising area for the application of machine learning and neural networks.

In recent years, significant progress has been made in this area. Models such as CASCADE[2] predict NMR peaks based on a given structure, while others, like ANN-PRA[4] and the DP4/DP5[3] models, assess whether a proposed structure aligns with its NMR data. Our project's objective is to predict the NMR peaks that a specific chemical structure would produce under ^{1}H NMR analysis.

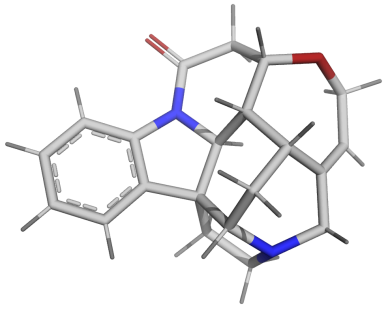
Problem:

Our dataset comprises 50,000 chemical structures and the atomic coordinates for each structure. The goal is to take a given chemical structure and predict its corresponding ^{1}H NMR spectral peaks. This capability would support the confirmation of structural predictions by comparing predicted peaks with experimental spectra.

Data:

Our dataset includes NMR spectra labeled with known compound classifications and their

corresponding chemical structures. Pre-processing steps involve filtering out noise and removing irrelevant data points to ensure model accuracy. The data is divided into training, validation, and test sets to allow the model to generalize across various compound types. An example chemical structure from the dataset, strychnine, is shown below. The output of the model will be a list of predicted NMR peaks for each molecule.



Methods:

This project employs a graph neural network (GNN) trained on chemical structure data. The data undergoes pre-processing, and the training set is used to train the GNN to generate NMR peak predictions for each chemical structure.

Next Steps:

For this project the next step will be to work on the data and train a basic graph neural network on the data to see how it performs and where the model can be improved and whether a more complicated model needs to be tested for this problem. Until this preliminary model is trained and tested, it's impossible to see what will need to be taken next.

Citations:

1. Cortés, I., Cuadrado, C., Daranas, A. H., & Sarotti, A. M. (2023). Machine learning in computational NMR-aided structural elucidation. *Frontiers in Natural Products*, *2*. https://doi.org/10.3389/fntpr.2023.1122426
2. Guan, Y., Sowndarya, S. V. S., Gallegos, L. C., St John, P. C., & Paton, R. S. (2021). Real-time prediction of 1H and 13C chemical shifts with DFT accuracy using a 3D graph neural network. Chemical Science, 12(36), 12012–12026. https://doi.org/10.1039/d1sc03343c
3. Howarth, A., & Goodman, J. M. (2022). The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. Chemical Science, 13(12), 3507–3518. https://doi.org/10.1039/d1sc04406k
4. Wang, F., Sarotti, A. M., Jiang, G., Huguet-Tapia, J. C., Zheng, S., Wu, X., Li, C., Ding, Y., & Cao, S. (2020). Waikikiamides A–C: Complex Diketopiperazine Dimer and Diketopiperazine–Polyketide Hybrids from a Hawaiian Marine Fungal Strain Aspergillus sp. FM242. Organic Letters, 22(11), 4408–4412. https://doi.org/10.1021/acs.orglett.0c01411