Predicting NMR Parameters Through Machine Learning

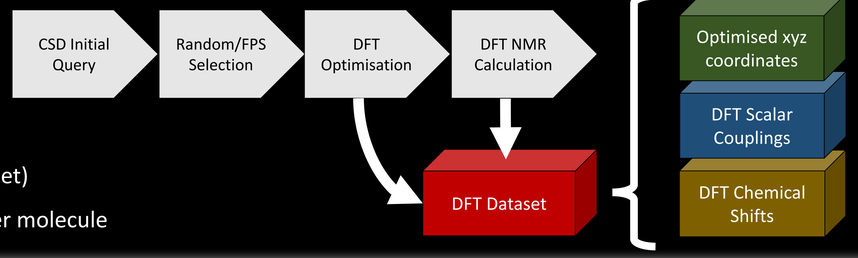
Will Gerrard, Lars Bratholm, Mathew Davy, Claire Dickson, Emily Johnston, Adrian Mullholland, Craig Butts

University of Bristol, School of Chemistry, Cantocks Close, Bristol, BS8 1TS. Email: wg12385@bristol.ac.uk

Current DFT methods can accurately predict NMR parameters for small molecules in 1-100 CPU hours. Machine learning (ML) models are capable of making the same predictions in 1-100 seconds. Existing methods use databases of experimental data to make predictions [1], however DFT could provide a more reliable way of producing training datasets. Early results indicate that ML models can predict DFT chemical shifts and coupling constants to within 5%.

Dataset Production

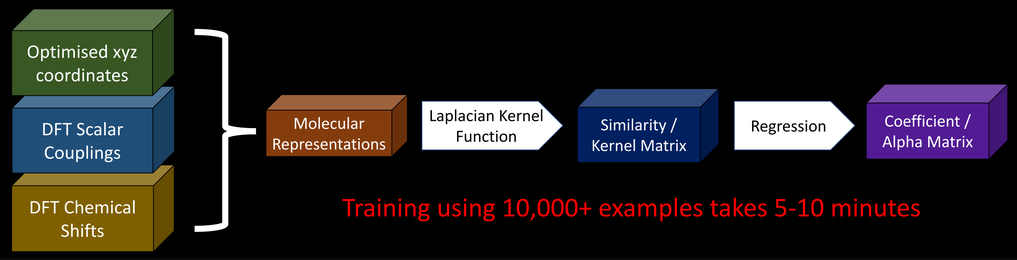
- Initial structures from the Cambridge structural database selected for:
  - No charge, no errors, H, C, N, O atoms only, R factor ≤ 5
- 500 Structures chosen at random (Test Dataset)
- 2000 Structures chosen by Furthest Point Sampling (FPS) (Training Dataset)
- DFT geometry optimisation + NMR calculation takes 10-50 CPU hours per molecule



The production of the current dataset took around 100,000 CPU hours

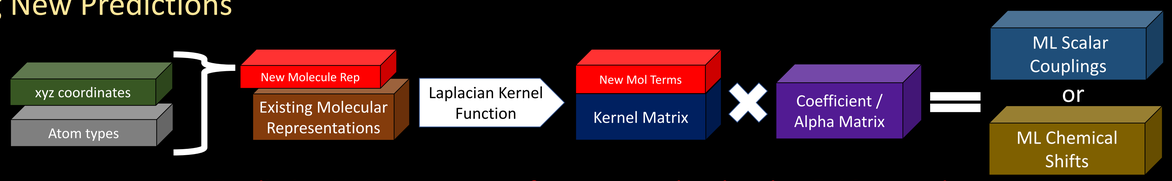
Training by Kernel Ridge Regression

- The raw DFT data is converted into molecular representations, which are the input for the ML algorithm.
- A measure of the similarity between each pair of representations is encoded in the kernel matrix determined via the kernel function.
- The 'training' consists of calculating the coefficient matrix. The coefficient matrix is the matrix that when multiplied by the kernel matrix, returns the vector of predicted values.



Training using 10,000+ examples takes 5-10 minutes

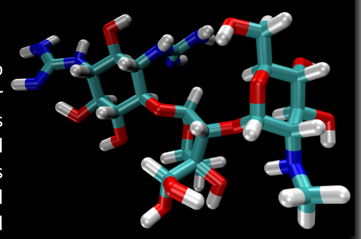
Making New Predictions

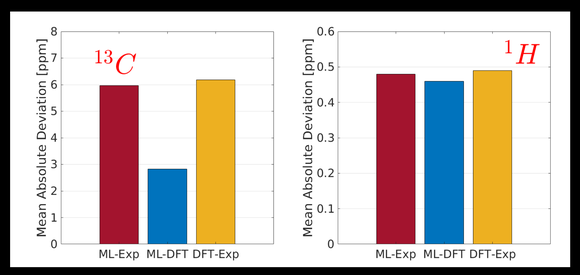


Predicting NMR parameters for a new molecule takes 5-10 seconds

Streptomycin

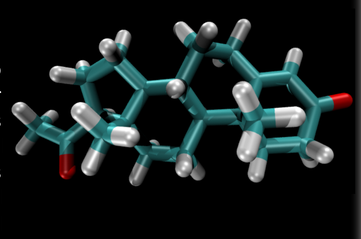
The ShML training set was used to predict the chemical shifts for Streptomycin, and these values were compared to experimental measurements. The ML predictions were accurate to 5.96 MAD and 5.79 SD for ^{13}C and 0.48 MAD and 0.78 SD for ^{1}H .

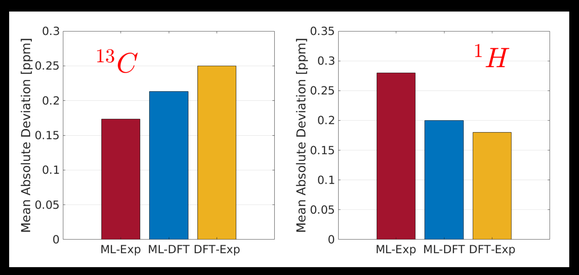




Progesterone

The ShML training set was used to predict the chemical shifts for Progesterone, and these values were compared to experimental measurements. The ML predictions were accurate to 3.47 MAD and 4.55 SD for ^{13}C and 0.28 MAD and 0.37 SD for ^{1}H .





Datasets, References, Acknowledgements

Extra data and info available on github

Current Dataset:
- HCNO only, molecules chosen from ShML [1], 2000 structures (current)

Future Datasets:
- HCNOF only, new FPS selection method, 2000 structures
- HCNOF+, FPS selected, multiple conformers per molecule

References
1. ACD/Labs NMR prediction software - www.acdlabs.com/products/adh/nmr/nmr_pred/
2. *Paruzzo, Federico M., et al.* "Chemical Shifts in Molecular Solids by Machine Learning." arXiv preprint arXiv:1805.11541 (2018).
3. *Huang, B., and O. A. von Lilienfeld*. "The 'DNA' of chemistry: Scalable quantum machine learning with 'amons'. arXiv preprint arXiv:1707.04146

Acknowledgements
- This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol – www.bristol.ac.uk/acrc/
- EPSRC NPIF project reference 1954848

ShML Test Set

Using the current dataset, NMR parameters for 400 test structures were predicted using 1600 training structures. This was done using the SLATM [3] representation via Kernel Ridge Regression. The accuracy of the ML predictions are relative to the DFT calculated values.

NMR Parameter	No. Atoms	Mean Absolute Deviation	Standard Deviation	Typical DFT MAD
^{13}C Chemical Shift	7522	**3.22 ppm**	**4.10 ppm**	**1-2 ppm**
^{1}H Chemical Shift	8474	**0.38 ppm**	**0.38 ppm**	**0.1-0.2 ppm**
$^{1}J_{CH}$ Coupling	7786	**1.78 Hz**	**1.90 Hz**	**2 Hz**