PREDICTING NMR SPECTRA USING LOW AMOUNTS OF DATA

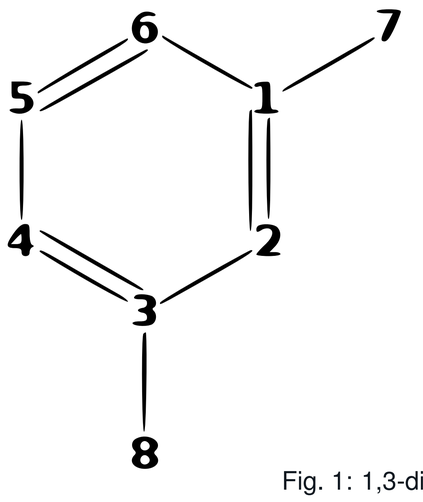
Herman Rull, Computer Science Master's

Supervisor: Stefan Hermann Kuhn, PhD

UNIVERSITY OF TARTU
1632

Introduction

Nuclear magnetic resonance (NMR) spectra are the fingerprints of the molecule that can be used to identify the molecule's structure, (Fig 1). These spectra are often visualized using graphs, as seen in Fig 2. Reading the spectrum requires specialized knowledge, and even then the decision often comes down to some potential candidates. Databases exist that collect NMR spectra so that the specialist can, through process of elimination, remove non-matching candidates. However, since the chemical space is vast and the machines required for measuring shifts are relatively expensive, then there is a need for accurate models that can predict spectra.



Atom	Shift(ppm)
1	137.56
2	130.00
3	137.65
4	126.25
5	128.25
6	126.25
7	21.25
8	21.25

Fig. 1: 1,3-dimethylbenzene and its NMR spectrum

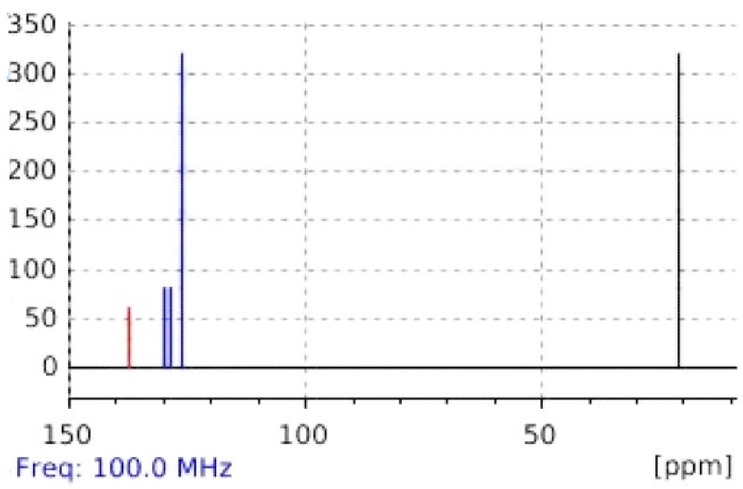


Fig. 2: 1,3-dimethylbenzene spectrum as a graph. Higher peaks indicate that there are multiple atoms with same shift.

There are two different approaches to predicting NMR shifts: data-driven and *ab initio*. Methods based on the latter predict the spectra based on detailed quantum computational calculations. In contrast to data-driven methods, this approach does not rely on existing data, however each prediction requires long and complex calculations. Therefore there are ongoing efforts to get better results with low amounts of data.

Methodology

- **Obtaining Data** - All spectra were from open-sourced web database nmrshiftdb2[4],
- **Preparing Data**
  - Descriptor/feature calculation and selection - rdkit[6], mendeleev[7]
  - Formatting descriptors as a graph - pytorch[5]
- **Training models**
  - HOSE-code based model. This is a relatively simple model originating from 1978[1], that generates a map of neighboring atoms of each atom in the training set and associates the neighborhood to the measured shift value. During the prediction phase, the model searches for the largest identical neighborhoods and, if found, then outputs its shift. Fig 3 illustrates neighborhoods of different widths for a carbon atom in the ibuprofen molecule.
  - GNN model architecture from an article[2] where it was used to predict ^{2}H NMR order parameters.
- **Comparing results**
  - The HOSE-code-based model, GNN model (2023) and the previous best GNN-based model (2019) from paper of Jonas and Kuhn were compared[3]
  - Models were tested on ^{19}F (957 molecules) and ^{13}C (44370 molecules) NMR spectra datasets.

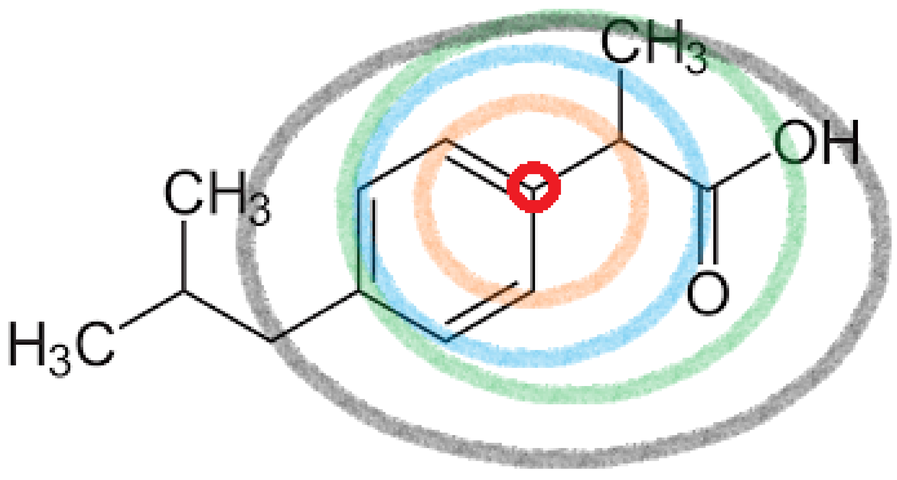


Fig. 3: Image shows neighborhoods used to construct HOSE codes for carbon atoms circled in red. The bigger the radius of the circle, the more layers of nearest atoms are considered to be in the neighborhood and therefore, if two atoms have identical HOSE codes up to a large radius (6+), then they also have almost equal NMR shift.

Results

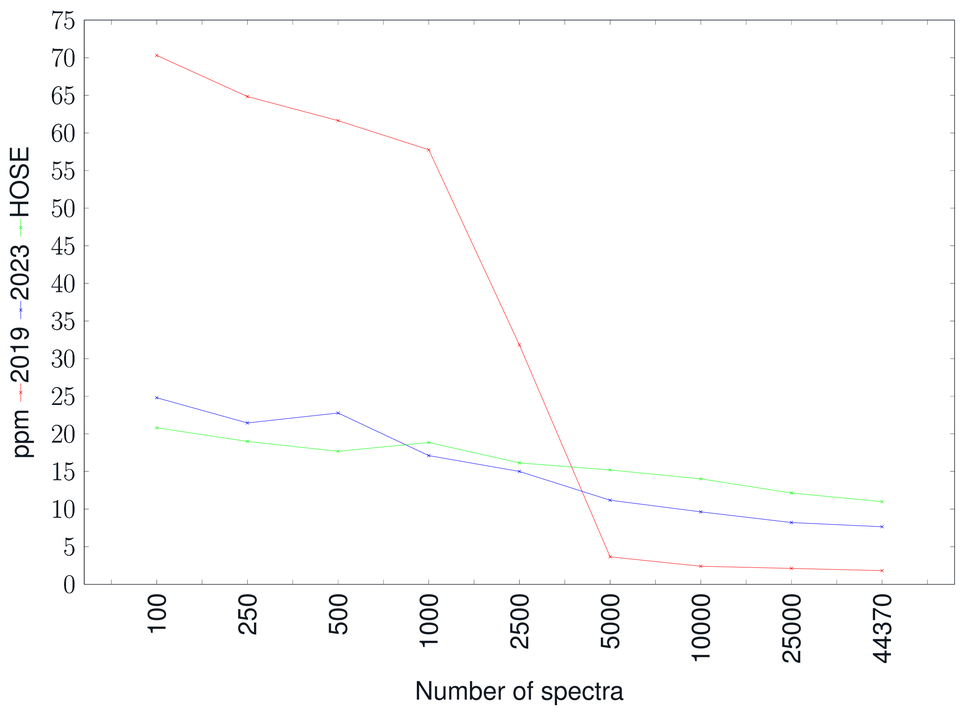


Fig. 4: Mean average error of ^{13}C NMR shift prediction in relation to number of available spectra

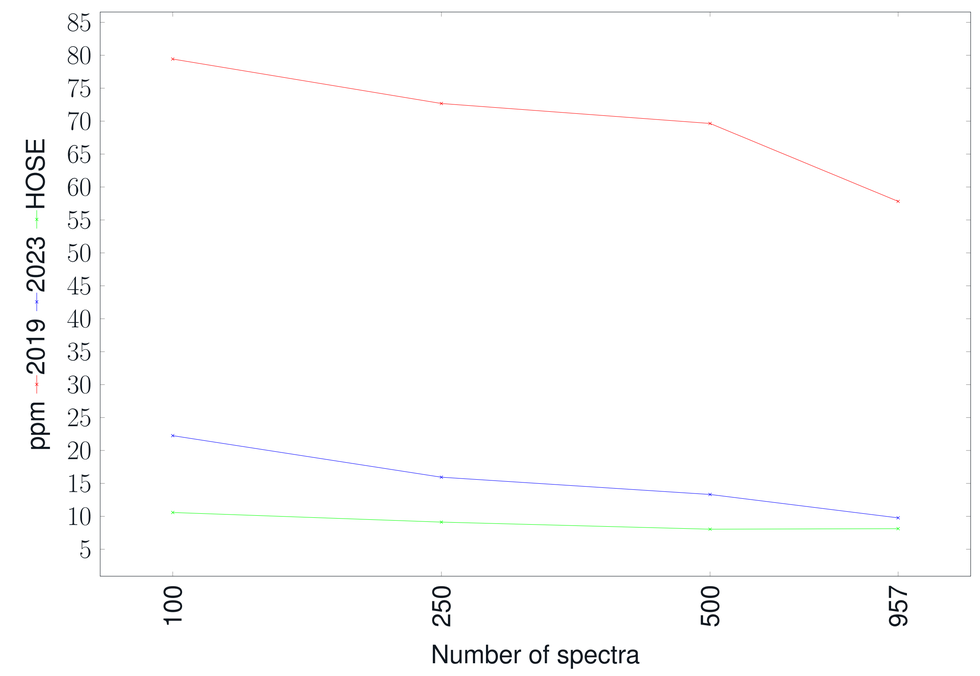


Fig. 5: Mean average error of ^{19}C NMR shift prediction in relation to number of available spectra

Project Colab page

Conclusions

- The model developed during this project has better MAE accuracy than NN-based methods when available spectra are low (<5000).
- When less than 1000 spectra are available, the HOSE-code-based model performs better than all NN-based methods.
- A sweet spot exists between 1000-5000 spectra, where the model developed in this project could perform better than all other data-driven methods.

References

[1] W. Bremser. "Hose — a novel substructure code". en. In: *Analytica Chimica Acta* 103.4 (Dec. 1978), pp. 355–365. ISSN: 0003-2670. DOI: 10.1016/S0003-2670(01)83100-7. URL: https://www.sciencedirect.com/science/article/pii/S0003267001831007 (visited on 05/25/2023).

[2] Markus Fischer et al. "Predicting 2H NMR acyl chain order parameters with graph neural networks". In: *Computational Biology and Chemistry* 100 (2022), p. 107750. ISSN: 1476-9271. DOI: https://doi.org/10.1016/j.compbiolchem.2022.107750. URL: https://www.sciencedirect.com/science/article/pii/S147692712200130X.

[3] Eric Jonas and Stefan Kuhn. "Rapid prediction of NMR spectral properties with quantified uncertainty". In: *Journal of Cheminformatics* 11.1 (Aug. 2019), p. 50. ISSN: 1758-2946. DOI: 10.1186/s13321-019-0374-3. URL: https://doi.org/10.1186/s13321-019-0374-3 (visited on 05/25/2023).

[4] *nmrshiftdb2 - open nmr database on the web*. URL: https://nmrshiftdb.nmr.uni-koeln.de/portal/js_pane/P-Help (visited on 05/25/2023).

[5] *PyTorch*. en. URL: https://www.pytorch.org (visited on 05/25/2023).

[6] *RDKit*. original-date: 2013-05-12T06:19:15Z. May 2023. URL: https://github.com/rdkit/rdkit (visited on 05/25/2023).

[7] *Welcome to mendeleev's documentation — mendeleev 0.13.1 documentation*. URL: https://mendeleev.readthedocs.io/en/stable/ (visited on 05/25/2023).