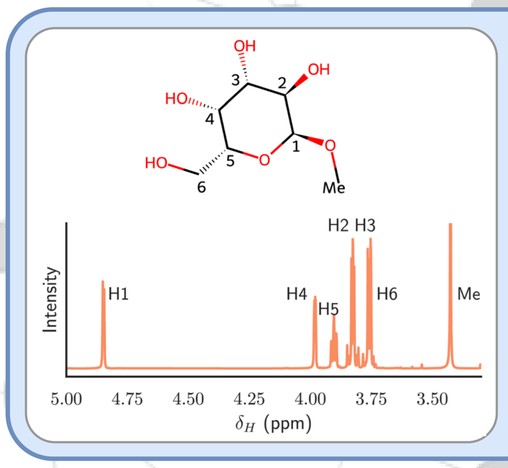
CarbPred

Using machine learning to predict nuclear magnetic resonance spectra for carbohydrates via graph neural networks



Carbohydrates are complex molecules, critical to drug and vaccine development. Their study often involves the analysis of nuclear magnetic resonance (NMR) spectra, which are challenging to interpret, even for experts.

We build upon prior work, GeqShift, employing E(3)-equivariant graph neural networks (GNNs) to automate this process. We explore a variety of methods and accomplish accuracy and computational efficiency improvements in multiple categories.

NextShift

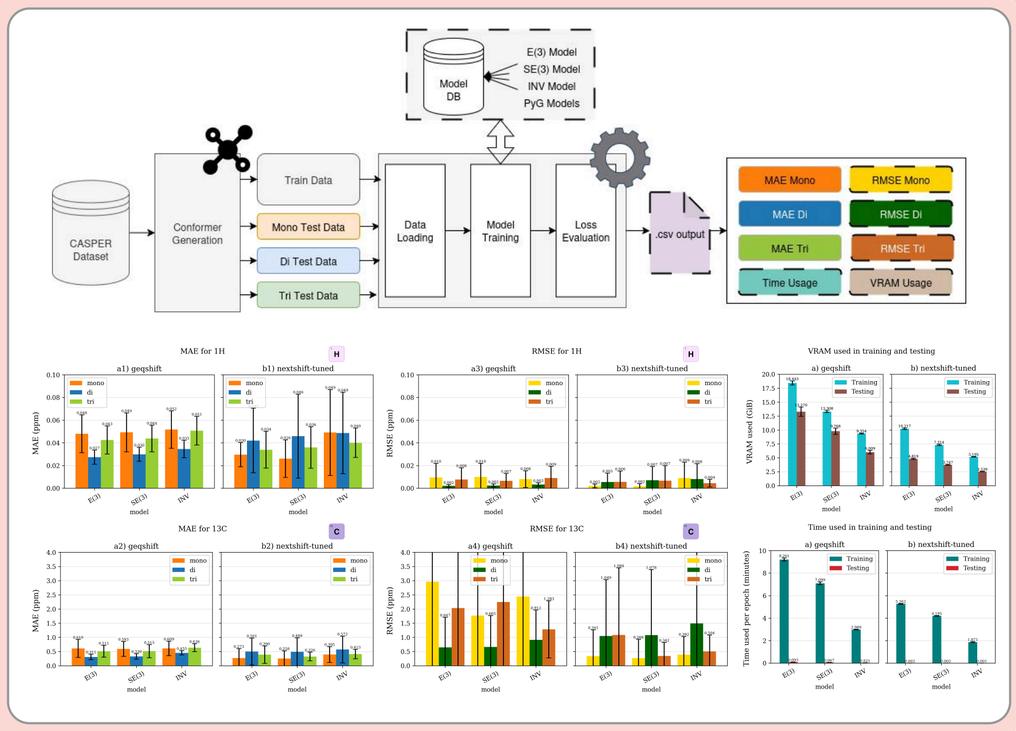
NextShift: next-generation, modular codebase, faster and more accurate

Aim Develop a next-generation, modular codebase (NextShift) featuring alternative models, equivariances and novel ML techniques. Evaluate NextShift against baseline GeqShift in prediction accuracy (MAE and RMSE) and computational efficiency (time and VRAM usage) for training and inference when using the original GeqShift dataset of 400 samples.

Method We refactor GeqShift by streamlining procedures and optimizing methods. We add automatic mixed precision, mini-batch training, random shuffling, regularization, alternative PyG models and SE(3) equivariance. We extensively hyper-parameter tune on new codebase.

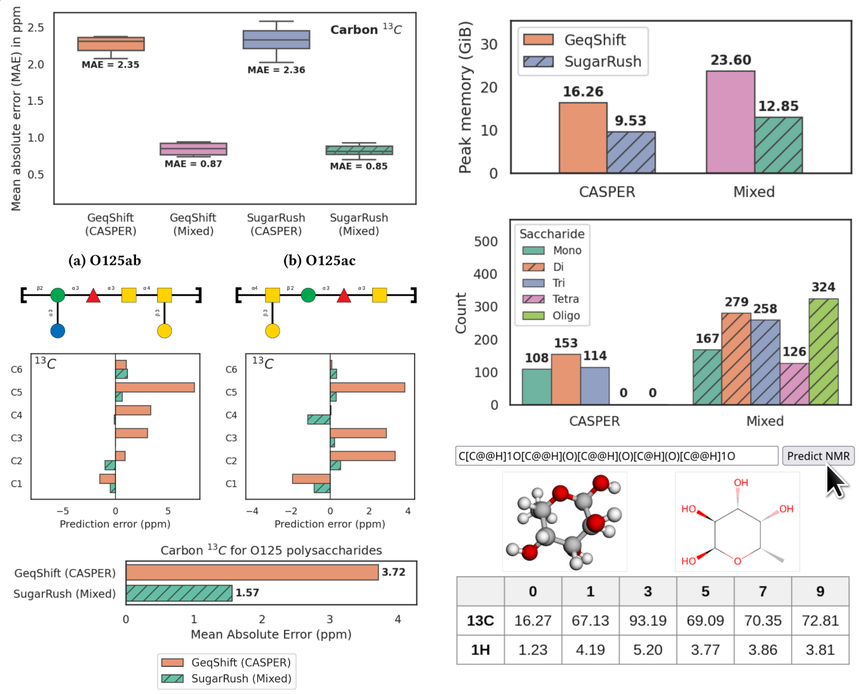
Results We accomplish 1.5x greater accuracy in 2 of 3 carbohydrate categories, 1.8x less time and memory usage for training and 46x less time and memory usage for testing. We demonstrate modularity with PyG models and provide an easy-to-use web interface.

Concl. NextShift is significantly more efficient than GeqShift and outperforms in accuracy for mono- and tri-saccharides but underperforms for di-saccharides. For equivariances, we find SE(3) more efficient than E(3) but with acceptable accuracy margins.



SugarRush

SugarRush: trained on more data to produce a more accurate and general model



C[C@@H]1O[C@@H](O)[C@@H](O)[C@H](O)[C@@H]1O Predict NMR

	0	1	3	5	7	9
13C	16.27	67.13	93.19	69.09	70.35	72.81
1H	1.23	4.19	5.20	3.77	3.86	3.81

SugarRush: trained on more data to produce a more accurate and general model

Aim Establish whether increasing the size of the dataset through adding additional data sources improves the accuracy of the GeqShift model measured in Mean Absolute Error (MAE).

Method We assemble a larger "Mixed" dataset using additional data from the CSDB and GlycoNMR, train the model and evaluate it against the original "CASPER" dataset. We implement automatic mixed precision (AMP) training to reduce memory consumption by using smaller datatypes where appropriate. We build an example web interface to demonstrate general usability.

Results An almost 3x improvement in accuracy while roughly halving memory consumption. Out-of-distribution testing on the *E. coli* O125 serogroup shows significant improvements in generalization ability. The web interface works to improve usability and responds in a reasonable amount of time.

Concl. A larger/more diverse dataset improves accuracy and generalization ability for GNN models like GeqShift. The memory optimizations and web interface make the model more accessible to researchers.

UniMol 2

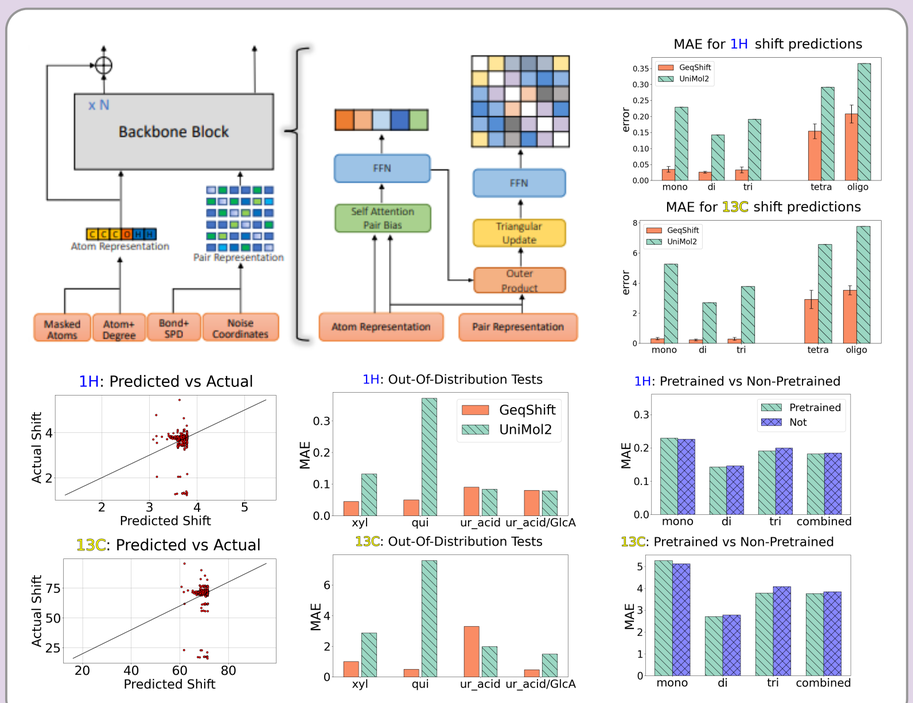
UniMol2: transformer-based, state-of-the-art chemical pre-trained model

Aim Fine-tune UniMol2 to predict NMR chemical shifts and evaluate its performance against GeqShift in terms of prediction accuracy (MAE and RMSE) using the original GeqShift dataset of 400 samples.

Method UniMol2 was initially designed to make a single prediction per molecule. To enable per-atom predictions, the input data were divided into distinct atom-specific subsets, each parsed into the model's backbone. A subgroup of backbone layers was frozen to retain general chemical knowledge acquired during pre-training.

Results UniMol2 achieved 5–10× lower accuracy than GeqShift across all carbohydrate categories, primarily due to its inability to handle the outlier shift values common in carbohydrate spectra. However, UniMol2 demonstrated better generalization to novel structures with narrower shift ranges, such as uronic acids. Additionally, retraining the model improved accuracy in two of the three carbohydrate categories compared to the baseline without pre-training.

Concl. Pre-training shows promise in improving model generalization to complex and novel carbohydrates. Future work could explore alternative pre-trained models or architectural modifications to enhance outlier handling and predictive performance.



By Unays Bhad (NextShift) [bhduna001@myuct.ac.za]
By Channing Bellamy (SugarRush) [bllcha013@myuct.ac.za]
By Matthew Dean (UniMol2) [dnxmat002@myuct.ac.za]

Supervised by Prof. Michelle Kuttel and Dr Jan Buys

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad
UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
SCHOOL OF IT

Unleash human potential for a fair and just society • UCT Vision 2030 •