

Project Proposal: Using machine learning (ML) to predict nuclear magnetic resonance (NMR) spectra for carbohydrates via graph neural networks (GNNs)

Unays Bhad [BHDUNA001]

bhduna001@myuct.ac.za

University of Cape Town

Cape Town, South Africa

Channing Bellamy

[BLLCHA013]

blcha013@myuct.ac.za

University of Cape Town

Cape Town, South Africa

Matthew Dean [DNXMAT002]

dnxmat002@myuct.ac.za

University of Cape Town

Cape Town, South Africa

Abstract

Carbohydrates are complex molecules that are important for researchers to study to advance medical research. Their study often involves the analysis of nuclear magnetic resonance (NMR) spectra, which can be costly and time-consuming to produce. Techniques to predict these NMR spectra using machine learning (ML) have been successful but operate on relatively small datasets and suffer weak generalisation ability. We propose methods to improve upon this by integrating advanced ML architectures and inductive biases, expanding the dataset size through data augmentation and the use of alternative data sources and utilising fine-tuned pre-trained ML models. We build on the work of GeqShift [2].

CCS Concepts

- Computing methodologies → Neural networks.

Keywords

Machine Learning, Graph Neural Networks, Equivariance, Molecular Property Prediction, Nuclear Magnetic Resonance (NMR), Carbohydrates, Data Augmentation

1 Introduction

Carbohydrate molecules are vital components of biological systems. They are present in many micro-organisms [18] and play a fundamental role in the development of medical drugs and vaccines [16].

These molecules have complex 3D structures that significantly impact their properties. For example, starch and cellulose are structurally identical apart from being mirror images of each other. Starch is digestible by humans, while cellulose is not.

As such, carbohydrates are typically identified using a specialised technique known as Nuclear Magnetic Resonance (NMR). It produces spectra which show the resonance peaks of the hydrogen and carbon atoms within the molecule. Researchers can infer valuable information about carbohydrates from the position of these peaks along the x-axis, referred to as the chemical "shift" of an atom. However, this process is costly and time-consuming, making an automated computerized approach attractive to researchers.

Various machine learning (ML) methods have been used for predicting molecular properties such as NMR [7]. Notably, graph neural networks (GNNs), a specialised subset of neural networks (NNs), have yielded successful results. This is attributed to their

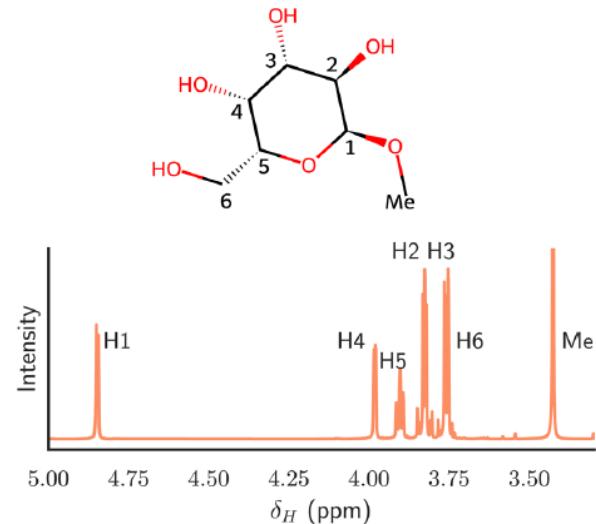


Figure 1: An NMR spectrum where each peak corresponds to a hydrogen atom in the displayed carbohydrate [2]

ability to effectively represent the geometric structure of molecules as graphs.

GeqShift is an E(3) equivariant GNN model dedicated to NMR prediction for carbohydrates [2]. It has outperformed the state-of-the-art (SOTA) in this task and has exceeded researcher expectations as prediction errors were demonstrated to be below error margins for practical measurements.

We seek to improve upon GeqShift in NMR prediction accuracy and generalisation ability by incorporating alternative techniques.

2 Background and Related Work

Our primary focus of related work is GeqShift [2]. It acts as our baseline for comparison. Furthermore, we inherit our methodology from it to ensure fair comparison.

2.1 GNN architectures and inductive biases

2.1.1 GNN theory. GNNs operate by iteratively performing two primary functions on a graph: message passing and aggregation. The definitions of these functions, define the GNN architecture.

We define a graph as $G = (V, E)$, where V and E are the set of nodes and edges, respectively. We assign each node a node feature, denoted as h_i for node i [7].

Message passing is whereby a node sends messages to its neighbouring nodes and is performed by all nodes simultaneously in parallel, as defined in Equ. 1. Neighbourhood aggregation is whereby received messages are aggregated and the aggregate is used to update the receiver node for the next iteration as defined in Equ. 2.

We define the message function and update function at a layer ℓ (i.e. an iteration):

$$m_{ij}^\ell = \psi_m(h_i^\ell, h_j^\ell, e_{ij}), \quad (1)$$

$$h_i^{\ell+1} = \psi_h(h_i^\ell, \{m_{ij}^\ell\}_{j \in N(i)}), \quad (2)$$

where $N(i)$ is the set of neighbours around node i (without self-loop by default), and ψ_m, ψ_h are parametric functions. [7]

2.1.2 GNN architectures. GeqShift's GNN architecture can be classified as *vanilla* GCN: the simplest variant of GNN in which messages are aggregated using a simple summation operation. It is defined:

$$h_i^{\ell+1} = \text{ReLU}\left(U^\ell \sum_{j \in N_i} h_j^\ell\right), \quad (3)$$

where $U^\ell \in \mathbb{R}^{d \times d}$. A bias is also included and node h_i^ℓ can be included via self-loops or residual connections.

Our motivation for alternative GNN architectures stems from a popular benchmarking survey [5]. It presents the performance of 25 architectures in molecular property prediction for two large datasets across various metrics. The results demonstrate the architectures Gated-GCN as the most accurate and GCN as the second most accurate.

The GCN (Graph ConvNet) architecture [10] replaces summation with symmetric normalization. It is defined:

$$h_i^{\ell+1} = \text{ReLU}\left(U^\ell \frac{1}{\sqrt{\deg_i} \sqrt{\deg_j}} \sum_{j \in N_i} h_j^\ell\right), \quad (4)$$

where \deg_i is the in-degree of node i and similarly for j .

The GatedGCN architecture [3] builds on GCN by utilising anisotropy, residual connections, batch normalization, edge gates [12] and explicit maintenance of edge features \hat{e}_{ij} at each layer. It is defined:

$$h_i^{\ell+1} = h_i^\ell + \text{ReLU}\left(\text{BN}\left(U^\ell h_i^\ell + \sum_{j \in N_i} \hat{e}_{ij}^\ell \odot V^\ell h_j^\ell\right)\right), \quad (5)$$

where $U^\ell, V^\ell \in \mathbb{R}^{d \times d}$, \odot is the Hadamard product and the edge gates \hat{e}_{ij}^ℓ are defined in a sophisticated manner.

Further in-depth definition of models and aforementioned terminologies can be found in Appendix A below.

2.1.3 Molecules as graphs. A molecule is abstracted as a graph where nodes represent atoms and edges represent bonds, as shown in Figure 2. The feature vector for an atom can be any collection of scalar properties such as proton number, neutron number, weight etc. Edge feature vectors are not strictly necessary as the graph inherently represents bonds through edges but can be defined to encode additional information [7].

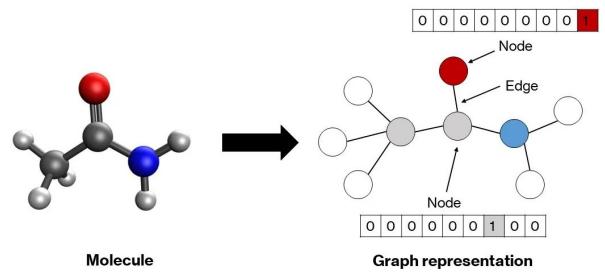


Figure 2: Molecules can be abstracted as graphs [4].

2.1.4 Inductive biases. Equivariance is the property whereby if a transformation is applied to the input of a function, the output also changes via the same transformation. This property can be utilised as an inductive bias of a GNN. It can significantly improve data efficiency and generalisation ability of a model. We focus here on geometric equivariance. Intrinsic GNN equivariance is discussed in Appendix A below.

$E(3)$ equivariance refers to the Euclidean group $E(3)$ which maintains translation, rotation and reflection symmetries. For example, translations of molecules in 3D space are identified as equivalent molecules. However, $E(3)$ is not ideal for molecules as reflections of molecules are not equivalent. $SE(3)$ is the Special $E(3)$ group which maintains translation and rotation but disregards reflection.

2.2 Carbohydrate Datasets

A large dataset containing high-quality carbohydrate structure and corresponding NMR spectra is important for training related ML models. There exist several databases which vary in size and completeness.

2.2.1 CSDB (Carbohydrate Structure Database). The CSDB is a curated repository of bacterial, fungal, and plant glycans, featuring near-complete coverage up to 2020. As of March 2025, it contains 32,937 carbohydrate structures and 19,728 NMR spectra [14]. Both experimental and simulated spectra are included.

Data export to various formats and encoding schemes is available on request, and web access is free. Structure data is annotated with assigned NMR spectra and other information where available. CSDB makes use of the CSDB Linear notation [17].

We obtained an export of 7,073 experimental ^1H and ^{13}C NMR spectra from CSDB in April 2025 by special request.

2.2.2 GlycoNMR. GlycoNMR is a curated carbohydrate-specific NMR dataset, aiming to address the scarcity of carbohydrate data for ML [23]. It includes both experimental and simulated shifts for 2,609 carbohydrate structures. The addition of simulated shifts expands the dataset, although it may introduce inaccuracies as the simulated shifts are not real experiments. GlycoNMR uses data from CSDB, and uses the same CSDB Linear notation.

2.2.3 NMRShiftDB2. NMRShiftDB2 [11] is an open-source database for organic molecules containing both experimental and simulated NMR spectra. The data is peer reviewed by a board of reviewers.

Table 1: Comparison of datasets

Dataset	Carbohydrate specific?	Sample size
CSDB	Yes	7,073
GlycoNMR	Yes	2,609
NMRShiftDB2	No	68,467
GeqShift	Yes	375

NMRShiftDB2 contains 271,668 structures, 68,467 experimental NMR spectra, and 396,583 simulated NMR spectra as of March 2025. [11].

Although this database does not focus on carbohydrates specifically, the large amount of data may be beneficial for model generalisation ability.

2.2.4 GeqShift Dataset. The small dataset used for GeqShift is based on published data used by CASPER. It contains 375 carbohydrate structures with 5,356 ^1H and 4,713 ^{13}C shifts.

2.3 Data Augmentation

Molecules may have many conformations (variants). A single conformation used in training may not fully represent the factors that influence NMR shifts. Data augmentation using conformational sampling is one method to increase the size of the dataset and account for these variants.

2.3.1 RDKit. RDKit is a chemoinformatics toolkit. GeqShift used RDKit to augment the data by generating multiple conformations for each carbohydrate molecule. This improved the model's accuracy, with a notable decrease in MAE.

2.3.2 CHARMM. CHARMM is molecular simulation software which targets biological systems (including carbohydrates) and features analysis and modelling tools [13]. CHARMM may produce more realistic conformations than RDKit due to its focus on molecular simulation.

2.4 Pre-trained models

Chemically pre-trained models are trained on large molecular datasets to predict various properties - such as atom types or bond lengths - in order to develop an understanding of these features. During training, the models' internal weights are adjusted to improve prediction accuracy. They can then be fine-tuned for a specific task, such as predicting NMR spectra, by retraining on data relevant to that task; during this process, the weights are adjusted further. To preserve knowledge of the learned chemical properties, some weights are frozen and cannot be modified during retraining [9].

2.4.1 UniMol+. UniMol+ [15] is a pre-trained model with a Transformer architecture. Transformers decompose the input molecular information into individual elements and learn the relationships among them all at once, rather than in sequential steps. While learning these relationships, the model assigns different weights to each element depending on its context [19]. UniMol+ was specifically designed to predict conformation-dependent properties, like NMR spectra. A molecule's conformation is the spatial arrangement of

its atoms resulting from rotation about a single bond. Some conformations are more stable than others; the most stable is called the equilibrium conformation. Although datasets often include data for the equilibrium conformation, they typically do not provide its coordinates, making reconstruction difficult. UniMol+ is trained to predict these equilibrium-conformation coordinates. During training, it uses those coordinates - alongside other molecular features - to predict target properties. By retraining on task-specific data, UniMol+ can be fine-tuned to perform specialized predictions.

3 Problem Statement and Research Questions

3.1 Problem Statement

GeqShift outperformed the SOTA for NMR prediction [2]. However, there are potential avenues for improvement with regards to prediction accuracy and generalisation ability to unseen samples.

GeqShift utilised the simplest "*vanilla*" architecture of GNNs, as described above. Benchmarking surveys show that alternatives GCN [10] and Gated-GCN [3] outperform other models for molecular property prediction [5].

Furthermore, GeqShift leveraged the $E(3)$ symmetry group to enable equivariance bias, but $E(3)$ is general and not necessarily the ideal symmetry group for molecular geometry. We hypothesize that the alternative $SE(3)$ equivariance better represents the geometric symmetries of molecules by treating chiral symmetries as distinct molecules [7].

The small dataset (375 samples) used to train GeqShift may have limited its ability to generalize to new carbohydrate molecules. RDKit was used to generate multiple conformations for each molecule but CHARMM [20] may be better suited to generating conformations as it relies on simulation instead of approximation. This data augmentation may increase dataset size for training.

For certain tasks, pre-trained models have outperformed manually trained NNs. They alleviate the data and computational limits in traditional training. We explore UniMol+ as a complete alternative that is not hindered by the architectural limits of GNNs.

Our work aims to improve upon GeqShift by experimenting with alternative GNN architectures and inductive biases, utilising a larger dataset and applying data augmentation techniques, and exploring chemical pre-trained models.

3.2 Research Questions

3.2.1 GNN architectures and inductive biases. How do GNN models using the architectures GCN and Gated-GCN and the inductive bias $SE(3)$ equivariance compare to the GeqShift with respect to prediction accuracy (measured using mean absolute error (MAE) and its standard deviation) and generalisation ability (measured via ablation studies) when trained and tested on the original dataset of 400 samples?

3.2.2 Dataset and data augmentation. How does a different, larger dataset, obtained from alternative data source(s) such as CSDB, improve the accuracy of the model (measured using MAE)?

Does augmenting the dataset, using molecular simulation software such as CHARMM to generate multiple conformations for each carbohydrate structure, improve the accuracy of the model (measured using MAE)?

3.2.3 Pre-trained models. When fine-tuned on NMR data, will the pre-trained model UniMol+ produce more accurate shift predictions than GeqShift, measured by a decrease in MAE for the predictions?

4 Procedures and Methods

4.1 Common methodology

We intend to follow the methodology of GeqShift to ensure comparable results. We will use its source code as a foundation of ours.

We will begin by following the methodology used in GeqShift to reproduce its results. We will use these results as our primary baseline of comparison.

We intend to use Python v3.9.13, PyTorch v2.0.0 and Cuda v11.7 as our ML resources. They are common in the field. We will use e3nn v0.5.1, an open-source equivariance framework, for the symmetry group implementation as it has gained extensive use. [6]. Training will be performed on NVIDIA A100 GPUs as done in GeqShift. These are accessed via UCT HPC resources.

We will use the same dataset employed by GeqShift which consists of 375 experimental samples of ^1H and ^{13}C NMR chemical shifts of mono- to trisaccharides. It is sourced from <http://www.casper.org.su.se/casper/liter.php>.

The dataset will be split for tenfold validation. This is where 10% of the data is used for testing and the experiment is repeated on 10 occasions using different test sets. From the 90% training set for each instance, 5% will be used for the validation set. Splitting the dataset allows us to verify if the model is generalizing well instead of recalling training samples. This is a common ML approach and was utilised by GeqShift.

Experiments will be performed in a grid format, where all combinations of architectures and symmetries will be represented. This also includes GeqShift as a baseline for measurement.

The aforementioned methodology will be employed for all components of our work. Component specific methodologies are provided below.

4.2 GNN architectures and inductive biases

We intend to experiment with GCN and Gated-GCN by replacing the message parsing procedure in the GeqShift source code to follow the mathematical principles of the respective architectures, as defined above. For GCN, this involves simple max-pooling. For Gated-GCN, this is much more complex and involves edge gates, edge encoding, residual connections and batch normalisation [10]. We illustrate them in Figure 3.

We intend to experiment with SE(3) equivariance by tweaking the parameters of the e3nn framework through its API.

4.3 Dataset and data augmentation

We intend to increase the size of the training dataset. We will replace the GeqShift dataset with a larger dataset and use alternative molecular simulation software to augment the dataset by generating multiple conformations for each carbohydrate.

4.3.1 Dataset. A new dataset will be constructed from one or more data sources, such as CSDB. The data, consisting of carbohydrate structures and NMR shifts for ^1H and ^{13}C atoms, will be converted into the correct format for training the ML models using similar

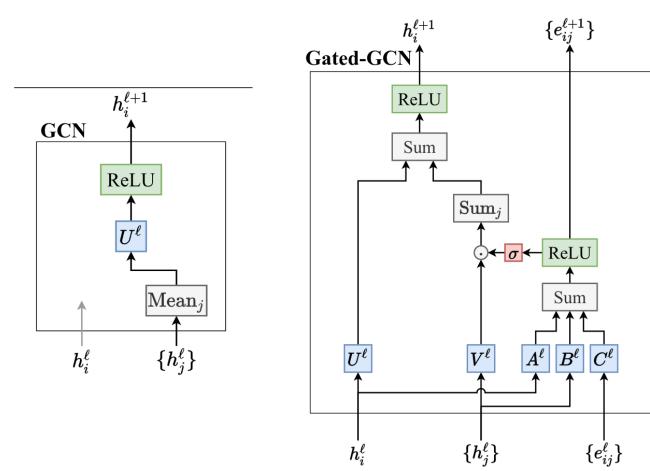


Figure 3: Architectures of GCN and Gated-GCN [5].

methods as used in GeqShift. Invalid data will be filtered out. The data, if obtained from multiple sources, will be combined into a single dataset.

4.3.2 Data Augmentation. Both the original dataset as used in GeqShift and the new dataset constructed will be augmented using molecular simulation software, such as CHARMM [20]. The data will be pre-processed for the molecular simulation software.

Multiple conformations, at least 100, for each molecule will be generated, the results converted into the appropriate format, and the additional conformers added to the datasets to expand their sizes. Conformations will be filtered depending on stability and energy levels.

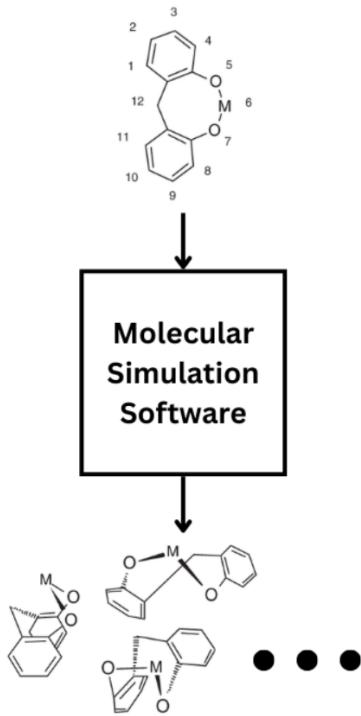


Figure 4: Molecular simulation software can be used to generate many conformations for a molecule

We will thus obtain several datasets:

- GeqShift Dataset
- GeqShift Dataset + GeqShift RDKit augmentation
- GeqShift Dataset + New augmentation
- New Dataset
- New Dataset + New augmentation

All datasets will be compared to each other using the original GeqShift model. Their impact to prediction performance will be evaluated.

4.4 Pre-trained models

UniMol+ takes atom and edge features as well as the atoms' 3D coordinates to make a prediction. The atom features are stored in a 2D matrix where each row corresponds to an atom while each column corresponds to a feature, such as its element type. The edge features are stored in a 3D matrix where each entry stores the value of a feature relating to each pair of atoms in the molecule. Such features include the distance between them and, if applicable, the type of bond that exists between them.

4.4.1 Pre-training. The model will be pre-trained on a large, publicly available database of general molecular information. During this stage, it will learn to predict basic properties - such as atom types and bond lengths - by masking features in each molecule and training itself to recover them. The model will also be trained to predict a molecule's equilibrium conformation, specifically by minimizing the MAE between its predicted coordinates and the reference coordinates in the training data.

4.4.2 Fine-tuning. The pre-trained model will then be fine-tuned on NMR data. If the dataset lacks 3D coordinates, we will generate the equilibrium-conformation coordinates synthetically by first sampling a random conformation and refining it with UniMol+. From that refined equilibrium conformation, we extract a list of 3D coordinates for the next step. If coordinates are already provided, we use them directly.

Using the dataset and 3D coordinates, we construct atom and edge features. From these features, we build two representations that the model updates at each layer. The first is the atom representation, initialized with the atom features; the second is the pair representation, constructed from the 3D coordinates and edge features. At each layer, both representations are updated and passed forward. Each element of the atom representation is updated based on its own value and those of all other atoms, with influence weighted by trainable parameters. The pair representation provides a bias term that is added to the element's updated value. The pair representation, itself, is updated via a series of learned update functions.

The model outputs a list of predicted chemical shifts, one for each 1H or ^{13}C atom in the molecule. We enforce the same atom ordering as in the experimental data so we can compute pairwise comparisons. The mean absolute error (MAE) between predicted and experimental shifts is calculated and minimized by adjusting the model's weights. To preserve knowledge gained during pre-training, we can freeze weights in some layers while allowing others to adapt—or alternatively freeze all pre-trained weights and add new, trainable layers. In this project, we will explore which strategy is most effective.

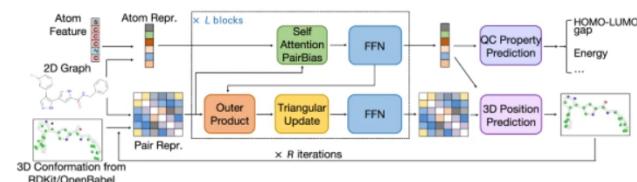


Figure 5: Model for how UniMol+ makes property and equilibrium conformation predictions [15]

Finally, UniMol+ will also be fine-tuned on synthetic data generated via CHARMM. Because this synthetic data typically includes atomic coordinates, it bypasses the need for synthetic equilibrium conformation generation. This allows us to compare which approach is more effective.

4.5 Final Model

We will construct a final model, using each of the approaches that achieved the greatest performance in our evaluation, in an effort to maximize prediction accuracy. This model will be evaluated and compared against the original GeqShift model as well as all models trained within our work.

4.6 Evaluation

Our primary metric for evaluation is mean absolute error (MAE) for prediction accuracy and its standard deviation for which lower indicates better. In NMR spectra predictions, Mean Absolute Error

(MAE) quantifies how close the predicted chemical shifts are to the actual values. A lower MAE indicates a more accurate prediction. Since the GNN approaches were trained to minimize MAE, this metric serves as a reliable measure for comparing their accuracy.

We intend to compare the MAE of our models against each other and against GeqShift. We will start by comparing the accuracy for simple mono-, di- and trisaccharides. We will also extend to comparing the accuracy for more complex polysaccharides.

We intend to perform ablation studies to measure generalisation ability. Specific samples will be removed from the training set and the model will be evaluated on its performance on those excluded samples. This will be compared to GeqShift. A minimum of six ablation studies will be performed where the excluded samples are based on those used in GeqShift [2].

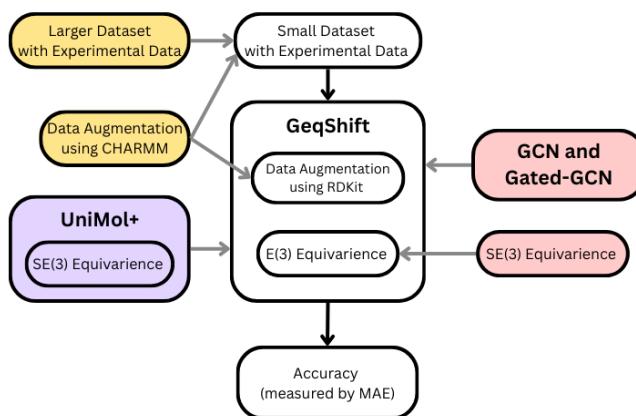


Figure 6: The original GeqShift experiment (in white) and what we plan to implement to improve the accuracy

5 Anticipated Outcomes

We will measure the success of our work based on a decrease in MAE against GeqShift and the model's pass rate in ablation studies.

We accomplish success in prediction accuracy should any of our models achieve lower MAE than GeqShift in the majority of test cases (75%). This is our primary measure of success.

Additionally, we accomplish success in generalisation ability should any of our models pass more ablation tests than GeqShift or pass the tests with more correct predictions.

Success of our work may result in the automation and reduction of cost in NMR prediction. Our work may provide a larger carbohydrate NMR spectra dataset for training future ML models in NMR prediction and related tasks. If successful, we improve on the accuracy of GeqShift, thus creating a more accurate and useful tool to researchers. This positively impacts the field of glycomics, and contributes to existing research in ML, as well as improve established databases such as the CSDB by providing more accurate NMR prediction tools. Finally, our work in solving this problem contributes to assisting the inverse problem (carbohydrate prediction from NMR spectra) which is a complex problem but has far-reaching impact in fields such as medicine and drug discovery.

5.1 GNN architectures and inductive biases

We anticipate a successful result as benchmark surveys show that alternative models have demonstrated previous success. We expect that the inherent properties of the architectures and the inductive biases that better represent the dataset contribute to better model performance.

5.2 Dataset and data augmentation

We anticipate a successful result as it is generally accepted that more data and more accurate data contributes to better model performance.

5.3 Pre-trained models

We anticipate a successful result because the model gains knowledge of general chemical properties through pre-training. These properties influence the NMR spectra of molecules, so we anticipate that a model aware of these influences will make more accurate predictions.

A significant portion of experimentally obtained chemical shift values are either incorrectly assigned to atoms or left unassigned [1]. ML models cannot be reliably trained on such data without compromising their accuracy. The anticipated results would demonstrate that even a limited amount of correctly labelled NMR data can be sufficient for training ML models—provided it is used in combination with large amounts of general, unlabelled molecular data.

6 Ethical, Professional and Legal Issues

Ethics clearance is not required for our work. Our experiments are computational and do not involve sensitive data. We make use of freely available data as well as open-source software.

One ethical concern is caused by high power consumption due to compute-intensive ML model training, which may contribute to global warming and climate change. Another concern involves the use of closed-source firmware and driver software for ML accelerators (such as NVIDIA GPUs). The continued general use of closed-source proprietary software may contribute to its dominance, hampering adoption of open-source alternatives.

The source code for UniMol+ is licensed under the terms of the MIT license. This means we are allowed to copy and modify the code. Any modifications to the code will be available under the MIT license as well. The other GNN architectures do not have such restrictions and their source code will be available under an open-source license.

7 Project Plan

Here we list the key milestones in our project. We also explain our timeline and provide the allocation of work. We emphasise dependencies and prioritise tasks. We concisely summarise this in a provided Gantt Chart.

7.1 Risks

7.1.1 GNN architectures and inductive biases.

- GeqShift source code may be deprecated

- Probability: Low. Code is fairly recent, access to original publishers.
- Impact: High. Significant more development if we have to start from scratch.
- Mitigation: Test GeqShift code early by reproducing their work.
- Management: Fix GeqShift code, reach out to original publishers or use an alternative PyTorch framework.
- Monitoring: Stay up-to-date on Python libraries.
- Difficulty in implementing Gated-GCN due to its inherent complexity
 - Probability: Medium. Many different properties and lack of domain specific knowledge.
 - Impact: Low. Gated-GCN is not critical as GCN will also be explored.
 - Mitigation: Request assistance from supervisor Dr Jan Buys.
 - Management: Ensure GCN is operational before implementing Gated-GCN, if insufficient time then do not explore Gated-GCN.
 - Monitoring: Mark progress on Gated-GCN implementation.
- Monitoring: Check whether UniMol+ becomes successfully fine-tuned.
- It may not be possible to pre-train UniMol locally due to insufficient computing power.
 - Probability: Medium.
 - Impact: Low.
 - Mitigation: Request access to UCT HPC cluster.
 - Management: Pre-train on a smaller dataset. Consider using a different model.
 - Monitoring: Check status of UCT HPC access request.

7.2 Timeline

The Gantt chart in Figure 7 (Appendix B) provides a timeline for when the high level tasks will be completed.

7.3 Resources Required

7.3.1 Equipment. Access to sufficiently powerful graphics processing units (GPUs) is required to train the ML model in reasonable time and may also be required for data augmentation with molecular simulation software. The GPUs must have sufficient memory to train the ML model. We intend to use the UCT HPC cluster to access GPUs for our project.

7.3.2 Software. Molecular simulation may require CHARMM software for data augmentation. CHARMM can be obtained after registration on the CHARMM website [13]. The appropriate ML software stack (example: ROCm on AMD or CUDA on NVIDIA) and UNIX environment capable of running ML pipelines using PyTorch is required.

7.3.3 Data. Carbohydrate NMR spectra data from a data source such as CSDB is required for the new dataset. The original GeqShift CASPER dataset is required to train the various GNN architectures and to fine-tune the UniMol+ model. This dataset will be provided by Göran Widmalm.

7.4 Milestones and Deliverables

- GNN architectures and inductive biases
 - GCN model implemented by June 20.
 - SE(3) and Gated-GCN implemented by August 15.
 - Models are trained upon completion and conclude by August 15 for evaluation.
- Dataset and data augmentation
 - Data source obtained by May 17
 - This data is used to construct a new dataset by May 30.
 - This new dataset is used to retrain GeqShift by June 20.
 - Datasets are augmented with simulation software by August 10 so they can be used to train all models.
- Pre-trained models
 - UniMol+ pre-trained by May 30 so it is ready for fine-tuning.
 - NMR Data is reformatted so UniMol+ can make predictions by June 15, indicating that fine-tuning can begin.
- General

7.1.2 Dataset and data augmentation.

- Alternative data may not be usable due to incompleteness.
 - Probability: Low. Multiple curated data sources (such as CSDB) exist.
 - Impact: High.
 - Mitigation: Obtain data from multiple data sources in advance.
 - Management: Filter or convert data where possible or use another data source instead.
 - Monitoring: Keep track of how much valid data has been obtained.
- Lack of knowledge on molecular simulation software may hinder data augmentation.
 - Probability: Medium. We are not familiar with molecular simulation.
 - Impact: Medium. Not as critical as the alternative dataset.
 - Mitigation: Request assistance from supervisor Prof. Michelle Kuttel.
 - Management: Use alternative augmentation methods such as the original RDKit method used in GeqShift.
 - Monitoring: Mark progress on molecular simulation process.

7.1.3 Pre-trained models.

- UniMol+ cannot be fine-tuned to predict NMR spectra.
 - Probability: Low.
 - Impact: High.
 - Mitigation: UniMol+ has already been fine-tuned on other datasets. We can copy the implementation of these.
 - Management: There are other pre-trained models, like GeoGNN [21], MolFormer [8] and UniMol[22], that could be considered.

- Original GeqShift experiment conducted by Widmalm et al [2] reproduced by June 20.
- June 20 marks the end of block 2 and a checkpoint for some tasks to be completed in preparation for an intermediate demo.
- Most of the individual tasks completed by August 15 so that evaluation of models can begin.
- Final report compiled by September 5 for submission.

7.5 Work Allocation

- Unays Bhad.
 - GCN and Gated-GCN architectures
 - SE(3) equivariance
 - Modification of GeqShift source code
 - Benchmarking and evaluation of GNNs' performance
- Channing Bellamy
 - New dataset
 - Data augmentation (molecular simulation)
 - Scriptwriting for parsing and converting data formats
 - Benchmarking and evaluation of dataset impact on performance
- Matthew Dean
 - Pre-train UniMol+ on ZINC
 - Script to construct embeddings for UniMol+ from CASPER and our new dataset
 - Fine-tune UniMol+ on CASPER and our new dataset

Acknowledgments

We would like to acknowledge Prof. Michelle Kuttel and Dr. Jan Buys from the University of Cape Town (UCT) for their supervision of our project.

References

- [1] Nathan Argaman and Guy Makov. 2000. Density functional theory: An introduction. *Am. J. Phys.* 68 (Jan. 2000), 69–79. doi:10.1119/1.19375
- [2] Maria Bánkestad, Keven M Dorst, Göran Widmalm, and Jerk Rönnols. 2023. Carbohydrate NMR chemical shift predictions using E (3) equivariant graph neural networks. *arXiv preprint arXiv:2311.12657* (2023).
- [3] Xavier Bresson and Thomas Laurent. 2017. Residual gated graph convnets. *arXiv preprint arXiv:1711.07553* (2017).
- [4] Gaurav Deshmukh. 2024. Building a graph convolutional network for molecular property prediction. <https://medium.com/data-science/building-a-graph-convolutional-network-for-molecular-property-prediction-978b0ae10ec4>
- [5] Vijay Prakash Dwivedi, Chaitanya K Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking Graph Neural Networks. *arXiv preprint arXiv:2003.00982* (2020).
- [6] Mario Geiger and Tess Smidt. 2022. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453* (2022).
- [7] Jiaqi Han, Yu Rong, Tingyang Xu, and Wenbing Huang. 2022. Geometrically equivariant graph neural networks: A survey. *arXiv preprint arXiv:2202.07230* (2022).
- [8] Vijil Chenthamarakshan Inkit Padhi Youssef Mroueh Jerret Ross, Brian Belgodere and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* 4 (Dec. 2022), 1256–1264. doi:10.1038/s42256-022-00580-7
- [9] Yuanqi Du Stan Z Li Jun Xia, Yanqiao Zhu. 2022. A Systematic Survey of Chemical Pre-trained Models. (Oct. 2022). doi:10.48550/arXiv.2210.16484
- [10] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
- [11] Stefan Kuhn. 2024. NMRShiftDB2 – Open NMR Database on the Web. <https://nmrshiftdb.nmr.uni-koeln.de>
- [12] Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826* (2017).
- [13] Harvard Molecular Mechanics. 2025. CHARMM. <https://www.academiccharmm.org>
- [14] Yuri A. Knirel et al Philip V. Toukach, Ksenia S. Egorova. 2023. Carbohydrate Structure Database (CSDB). <http://csdb.glycoscience.ru>
- [15] Di He Linfeng Zhang Shuqi Lu, Zifeng Gao and Guolin Ke. 2024. Data-driven quantum chemical property prediction leveraging 3D conformations with Uni-Mol+. *Nat Commun* 15 (Aug. 2024). doi:10.1038/s41467-024-51321-w
- [16] Therese Buskas Thomas J. Boltje and Geert-Jan Boons. 2009. Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nature Chemistry* 1, 8 (Jan. 2009), 611–622. doi:10.1038/nchem.399
- [17] Philip V. Toukach and Ksenia S. Egorova. 2020. New Features of Carbohydrate Structure Database Notation (CSDB Linear), As Compared to Other Carbohydrate Notations. *J. Chem. Inf. Model.* 60, 60 (2020), 1276–1289.
- [18] Ajit Varki. 2016. Biological roles of glycans. *Glycobiology* 27, 1 (Dec. 2016), 3–49. doi:10.1093/glycob/cww086
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbdd053c1c4a845aa-Paper.pdf
- [20] III Bernard R. Brooks Wonmuk Hwang, Charles L. Brooks. 2024. CHARMM at 45: Enhancements in Accessibility, Functionality, and Speed. *J. Phys. Chem. B*, 128 (2024), 9976–10042.
- [21] Jieqiong Lei Donglong He Shanzhuo Zhang Jingbo Zhou Fan Wang Hua Wu Xiaomin Fang, Lihang Liu and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* 4 (Feb. 2022), 127–134. doi:10.1038/s42256-021-00438-4
- [22] Gengmo Zhou, Zifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=6K2RM6wVqKu>
- [23] Lachele Foley Robert Woods Pengyu Hong Zizhang Chen, Ryan Paul Badman. 2024. GlycoNMR: DATASET AND BENCHMARKS FOR NMR CHEMICAL SHIFT PREDICTION OF CARBOHYDRATES WITH GRAPH NEURAL NETWORKS. *cs.LG* 2311, 17134v2 (2024).

A GNN architectures

We add upon the definitions aformentioned in the paper and provide further details.

We defined a graph of a GNN as $G = (V, E)$, where V and E are the set of nodes and edges, respectively. We assign each node a node feature, denoted as h_i for node i . We add that we may also optionally assign an edge feature e_{ij} for the edge connecting node i and j [7].

Accounting for weighted or directed graphs is a trivial modification. Weighted graphs can be implemented by accounting for edges e_{ij} . Directed graphs can be implemented by replacing the set $\{m_{ij}\}_{j \in N(i)}$ with $\{m_{ij}^\ell : j \rightarrow i\}$ which is the set of neighbouring nodes j pointed to node i . [5]

With regards to architectures, technically, *vanilla* GNN can be considered a variant of GCN because summation is a valid convolution operation. For consistency within literature, we refer to the normalising architecture as "GCN" and the simple summation-based architecture as "*vanilla* GCN".

For completeness, we describe the afovementioned terminologies:

ReLU is a common activation function for NNs which introduces non-linearity to improve a model's generalisation ability.

The in-degree of a node i refers to the number of incoming incoming edges to node i .

U^ℓ refers to the weight matrix applied to processed input. For brevity, we define that the bias term is incorporated within the weight matrix.

Self-loops and residual connections are techniques used to retain a node i feature vector for updating.

Anisotropy is a property exhibited by neural networks utilising attention mechanisms. It is the property whereby hidden representations align in specific directions. It is calculated using edge weights which account for the relationship between a node i and neighbouring node j .

Batch normalization (BN) is a technique to normalize the output activations of intermediate layers in the NN in mini-batches. Normalisation refers to ensuring the zero mean and unit variance within the mini-batch. It improves training stability (by preventing exploding or vanishing gradients) and increases training speed.

We observe that the update function (and by extension the model) exhibits the property of permutation equivariance if ψ_h is permutation equivariant. [7] For example, suppose ψ_h is the summation function. The order in which elements are summed is irrelevant hence the summation is permutation equivariant (more specifically, permutation invariant). When applied to graphs, this implies that the order of node aggregation is irrelevant. In other words, swapping any two nodes does not change the output of the GNN.

Permutation equivariance has been demonstrated to be highly beneficial in the modelling of real-world geometric systems. It reduces the combinatorial complexity of the function input ordering. This increases learning rate and reduces computational cost. [7]

We emphasise that whilst GNNs are typically permutation equivariant, they are not always geometrically equivariant.

B Gantt Chart

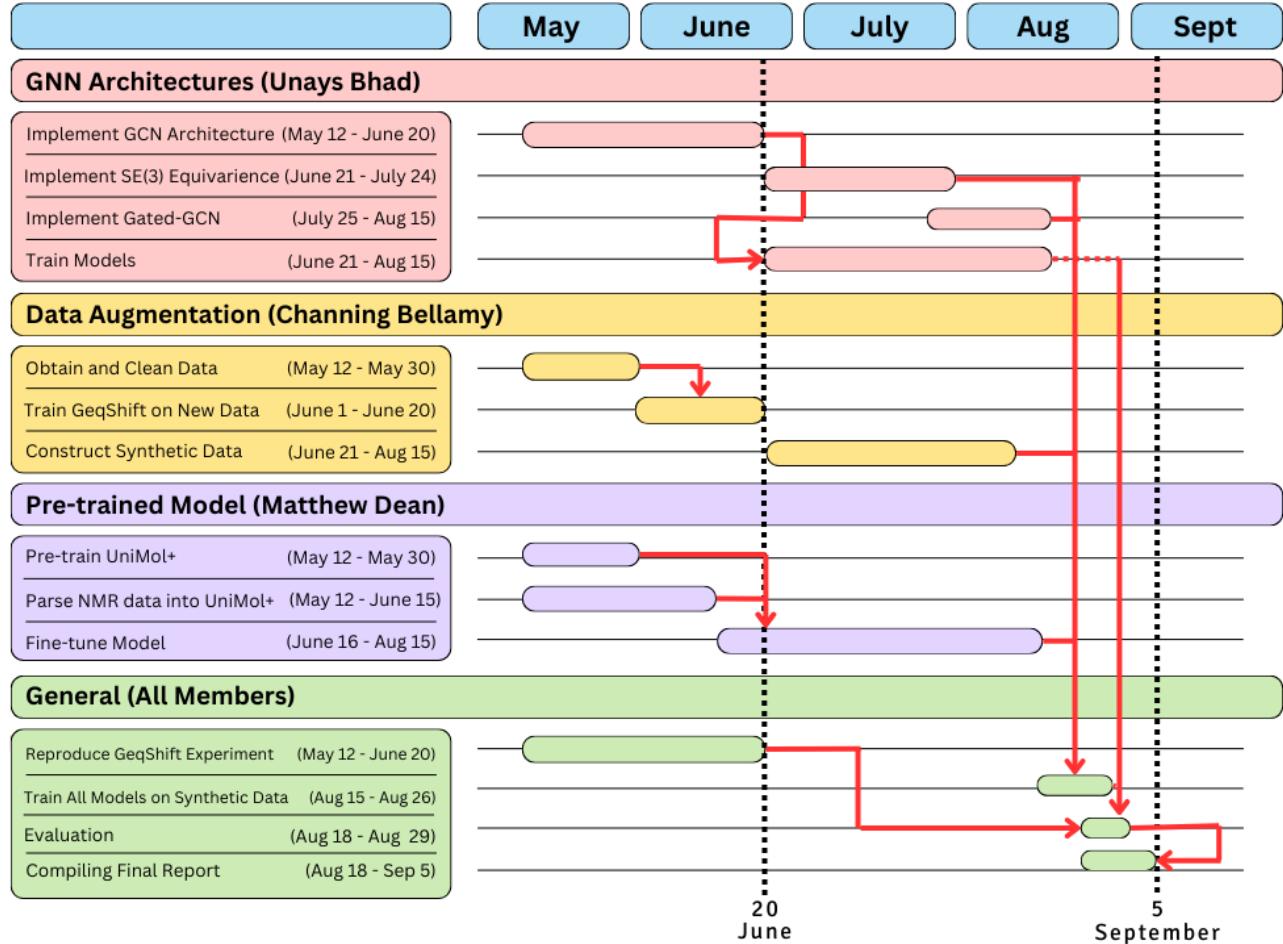


Figure 7: A Gantt chart providing a timeline for the project