SugarRush: Improving generalization and usability of machine learning (ML) models for prediction of carbohydrate nuclear magnetic resonance (NMR) spectra through an expanded dataset and web API

Channing Bellamy [BLLCHA013]
bllcha013@myuct.ac.za
University of Cape Town
Cape Town, South Africa

Abstract

Carbohydrates are complex molecules. Prediction of their nuclear magnetic resonance (NMR) spectra using machine learning (ML) has been successful but models are trained on small datasets and thus struggle to generalize to previously unseen carbohydrates. We build on the work of GeqShift [6] by expanding the dataset with additional data sources and optimizing the ML model to produce our model - SugarRush. In addition, we present a simple web API for integration with existing services. Our model shows increased prediction accuracy, up to fourfold for oligosaccharides, compared to previous models. Significantly reduced memory consumption and the development of a web API enables the model to be trained on less powerful hardware and be more easily used by researchers.

Keywords

Machine Learning, Graph Neural Networks, Molecular Property Prediction, Nuclear Magnetic Resonance (NMR), Carbohydrates, Datasets, CASPER, CSDB, GlycoNMR, Web API

1 Introduction

Carbohydrates are vital components of biological systems. They are present in many micro-organisms [40] and play a fundamental role in developing medical drugs and vaccines [35].

These molecules have complex 3D structures that significantly impact their properties. For example, starch and cellulose appear structurally similar as both are made of glucose monomers linked together, but are distinct due to differing linkages and hydroxyl group orientations. Starch is digestible by humans, while cellulose is not.

Carbohydrates are typically identified using a specialized technique known as Nuclear Magnetic Resonance (NMR). Molecules are exposed to electromagnetic radiation while being placed in a magnetic field, causing them to emit an electromagnetic signal at a certain frequency. Spectra are produced from these signals which show the resonance peaks of the hydrogen and carbon atoms within the molecule. Researchers can infer valuable information such as the primary structure, ring sizes, and linkages of carbohydrates from the positions of these peaks along the x-axis, referred to as the chemical "shift" of an atom. However, this process is challenging, making an automated computerized approach wherein structures are generated from NMR spectra by software attractive to researchers.

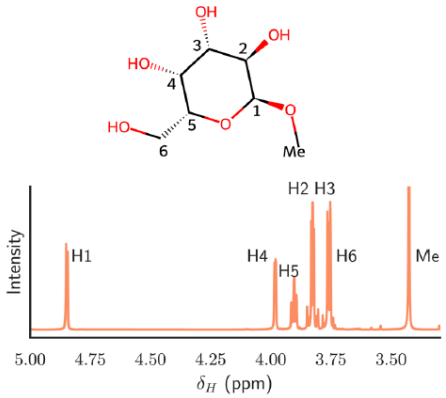


Figure 1: An NMR spectrum where each peak corresponds to a hydrogen atom in the displayed carbohydrate [6].

Various machine learning (ML) methods have been used for predicting molecular properties such as NMR [19, 22, 23, 27, 45]. Notably, graph neural networks (GNNs), a specialized subset of neural networks (NNs), have yielded successful results by predicting shifts with accuracies comparable to the best performing DFT functionals while using less processor time [20]. This is attributed to graphs providing effective representations for the geometric structure of molecules.

GeqShift is an E(3) equivariant GNN model dedicated to NMR prediction for carbohydrates [6]. It exceeded the researchers' expectations by outperforming state-of-the-art results from the SG-IMP-IR GNN model, delivering prediction errors that approach levels that qualify as error margins for experimental NMR measurements, but there is potential for improved prediction accuracy and generalization to unseen samples, especially for larger polysaccharides.

The dataset used in GeqShift is small, containing 375 samples [6]. The largest class of saccharides present in the dataset is trisaccharides, carbohydrates consisting of three simple carbohydrates known as monosaccharides linked together, while larger polysaccharides consisting of more than three monosaccharides such as tetra- (four linked monosaccharides) and oligosaccharides (five to ten linked monosaccharides) are not present. Training on these examples may improve the accuracy for larger and more complex carbohydrates.

We aim here to establish whether increasing the size of the dataset through adding additional data sources improves the accuracy of the model (measured using mean absolute error (MAE)

and root mean squared error (RMSE)). To establish a baseline for comparison and validate the results achieved by GeqShift, we first reproduce the results using the original dataset. We then expand the dataset using additional data sources, which we hypothesize to improve the model's generalization to unseen carbohydrates by learning from a larger number of more diverse examples. To allow the model to be more easily trained on larger datasets by a wider audience of researchers, we then optimize the model to reduce memory consumption. Finally, we present an example web API to allow the practical use of the model by researchers on the web.

2 Background and Related Work

In this section, we first provide an overview of carbohydrates and their structural aspects relevant to NMR prediction. We then introduce data augmentation as a strategy for expanding training datasets in machine learning. Next, we review GeqShift, outlining its dataset and model pipeline. Finally, we present alternative datasets that can be used to extend and improve upon the original GeqShift dataset.

2.1 Carbohydrates

Carbohydrates are structurally complex molecules consisting of carbon (C), hydrogen (H), and oxygen (O) atoms , containing a chain of carbons, hydroxyl (-OH) groups which consist of hydrogen and oxygen atoms, as well as an aldehyde or a ketone.

Carbohydrates are molecules made up of carbon (C), hydrogen (H), and oxygen (O) atoms [32]. Their structure is commonly based on a chain of carbon atoms, with oxygen- and hydrogen-containing groups known as hydroxyl (-OH) groups attached. In addition, one of the carbons usually carries a reactive group that classifies the molecule as either an aldehyde or a ketone.

"Saccharide" is a synonym for carbohydrate [3]. The fundamental building blocks of larger carbohydrates are monosaccharides, the simplest form of carbohydrate, which can range from three to nine carbon atoms [32] and form rings. Saccharides are divided into three main groups: monosaccharides, disaccharides (two linked monosaccharides), and polysaccharides (more than two linked monosaccharides). A polysaccharide can contain several thousand or more monosaccharides or residues, forming complex carbohydrates. In an effort to provide more granularity in this paper, we will further divide polysaccharides into tri-, tetra-, and oligosaccharides which consist of three, four, and five to ten residues respectively.

2.1.1 Configurations. The configuration of a carbohydrate refers to its topology of atoms and bonds as well as spatial arrangement. Multiple distinct carbohydrates can emerge from the same chemical formula, differing in their spatial molecular arrangement, and are known as isomers. A chiral carbon is one which is bonded to four different atoms or groups. Monosaccharides have two different isomers, D- or L-, depending on the orientation of the hydroxyl group on the chiral carbon farthest from the aldehyde or ketone which is known as the chiral center or stereocenter. These isomers can be thought of as mirror images of one another. [33].

In Figure 2, D-glucose and L-glucose are mirror images of one another. Even though the basic chemical formula for both glucose

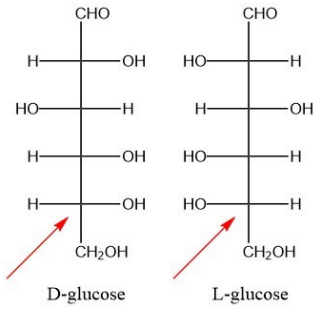


Figure 2: D-glucose and L-glucose isomers with the chiral carbon indicated by the red arrow. These isomers are mirror images of one another. [33]

forms is the same ($C_6H_{12}O_6$), the two resulting isomers differ in their spatial molecular arrangements.

Carbohydrates have a closed-ring five-membered (furanose) or six-membered (pyranose) form. The anomeric carbon forms a new stereocenter when the ring closes. The cyclic form has two versions: α- (alpha) and β- (beta), where the hydroxyl group on the anomeric carbon is pointing down and up respectively. The two forms, known as anomers, interconvert as the ring opens and closes [33]. Combined with the D- and L- forms, a monosaccharide can exist in four stereoisomeric forms (e.g., α-D-glucose, β-D-glucose, α-L-glucose, β-L-glucose).

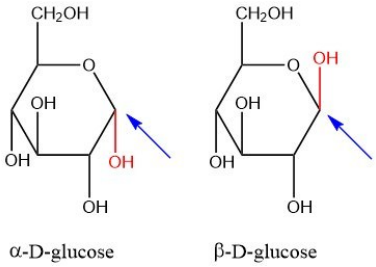


Figure 3: The alpha and beta forms of D-glucose, with the anomeric carbon indicated by the arrow, and the hydroxyl group in red [33].

Beyond configuration, carbohydrates also adopt multiple conformations, which are changes in spatial arrangement without altering bonds. Conformations interconvert dynamically depending on steric and electronic effects, and these differences can influence both molecular recognition and NMR chemical shifts [32].

2.1.2 Glycosidic Linkages. Polyaccharides consist of multiple monosaccharides bonded together through glycosidic linkages, which are covalent bonds between two monosaccharides [9].

A common disaccharide is the milk sugar – lactose. It consists of galactose and glucose, both monosaccharides, bonded together through a glycosidic linkage. The bond forms between the anomeric carbon of galactose and the fourth carbon of glucose. In Figure 4, galactose is in a β- anomeric configuration as the hydroxyl group on the anomeric carbon is pointing up.

If the α- anomeric configuration of galactose was used instead (hydroxyl group on the anomeric carbon point down), it would

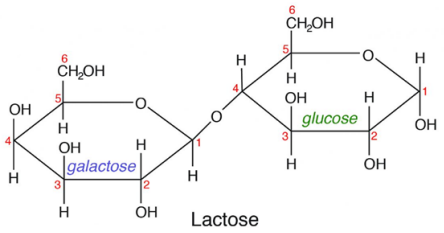


Figure 4: Lactose formed by galactose and glucose linked via a $\beta(1\rightarrow4)$ glycosidic bond. [32]

result in a different sugar, with linkage $\alpha(1\rightarrow4)$. This is true even though the chemical formula remains the same.

2.2 Nuclear Mangnetic Resonance

In nuclear magnetic resonance (NMR) spectroscopy, molecules emit electromagnetic signals at certain frequencies when they are exposed to electromagnetic radiation. For carbohydrates, ^{1}H NMR detects signals from hydrogen atoms, while ^{13}C NMR provides complementary information from carbon atoms. These signals are used to produce NMR spectra which show the resonance peaks of the atoms in the molecule. The positions of these peaks along the x-axis, referred to as the chemical "shift" of an atom, are affected by the local electronic environment surrounding atomic nuclei. Subtle differences in atomic environments - such as ring conformation, stereochemistry, or hydrogen bonding - produce distinct shifts, which can be studied by researchers to infer information such as the structure, ring size, and linkages in the molecule. Structural elucidation of carbohydrates from their NMR spectra is a challenging process [18]. While GeqShift focuses on predicting the NMR spectra for a given molecule, the inverse task of generating a molecular structure from NMR spectra is attractive to researchers and research may be assisted by first solving the forward problem of NMR spectra prediction.

2.3 Data Augmentation

Data augmentation is a technique that increases the size and/or diversity of a dataset by artificially creating new data based on existing data. Datasets used for machine learning consist of training samples with labels describing the training samples. Augmentation is achieved by creating new data samples without altering the corresponding labels, by way of a transformation operation applied to the original samples [29].

Machine learning (ML) models usually require large amounts of high-quality labeled data. Data augmentation serves to increase the size of a training dataset, which can improve the accuracy of ML models, improve generalization, and reduce overfitting [31, 43] - a phenomenon where the model adapts to its training data too closely which reduces its ability to generalize to previously unseen samples. Data augmentation is especially important when existing datasets are small, improving model robustness to unseen data [30].

2.4 GeqShift

GeqShift is an E(3) equivariant graph neural network (GNN) which outperformed the state-of-the-art model for NMR prediction. GeqShift achieved accuracies approaching levels considered as measurement error for practical NMR experiments [6]. However, GeqShift was trained on limited data. GNNs struggle when dealing with carbohydrates due to their complex stereochemical arrangements. A unique aspect of GeqShift is its use of E(3) equivariance, the group of transformations in the Euclidean space, encompassing rotation, translation, and mirroring. [6].

GeqShift comprises a model, a data augmentation method, and a dataset of carbohydrates and their associated chemical shifts. The GeqShift code and dataset is available on GitHub [5].

The pipeline from dataset input to result output consists of two main steps: dataset generation, and model training and testing.

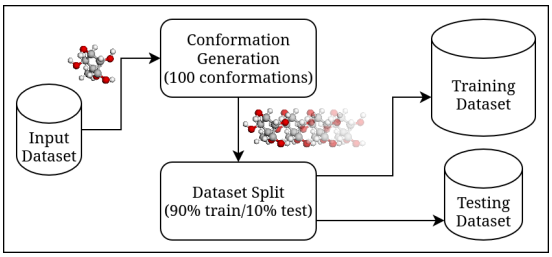


Figure 5: Overview of the GeqShift data pipeline showing how molecules from the input dataset have 100 conformations generated before being split into train and test datasets.

2.4.1 GeqShift Dataset. The dataset, which we will refer to as the CASPER dataset, is based on published data used by CASPER [34], a web tool designed to aid the classification and representation of carbohydrates, which also uses its own notation. It consists of 108 monosaccharides, 153 disaccharides, and 114 trisaccharides giving a total of 375 carbohydrate molecules and 4955 ^{13}C and 4856 ^{1}H chemical shifts. The dataset is small, containing just 375 carbohydrates, and is lacking larger polysaccharides.

The data is split into separate Structural Data Format (SDF) files for each class of saccharide, which contain sections consisting of an MDL MOL description of the atoms and their coordinates, their connections, and property tags containing a list of NMR chemical shifts for each atom type, ^{13}C and ^{1}H, matched by atom index.

RDKit [17], a free and open-source chemoinformatics software toolkit, is used to augment the dataset by generating 200 conformations per molecule using ETKDGv3, a stochastic method that utilizes distance geometry [42] to calculate atomic coordinates, with MMFF94 [21], a general force-field which describes the forces between atoms in a molecule for energy calculation during conformation generation. Only the 100 conformations with the lowest energy are kept - the rest are discarded. As molecules can have many conformations, a single conformation may not fully model factors that affect NMR chemical shifts. In GeqShift, training on 100 conformations per molecule reduced the mean absolute error (MAE) from 0.55 to 0.34 for ^{13}C chemical shifts [6]. Each conformation contains the same atom and connection table, but with differing 3D coordinates determined by the simulation. The dataset lacks residue mapping information to match each residue of a polysaccharide to their associated atoms. This poses challenges in generating conformations with alternative force-fields such as CHARMM [28] or

AMBER [39], which could model carbohydrates more accurately as they contain carbohydrate-specific definitions.

The data is split into separate training and test datasets for ^{13}C and ^{1}H. The testing datasets are further split into separate datasets for each saccharide class, where the CASPER data is split between mono-, di- and trisaccharides. These datasets are serialized into byte streams using the Python pickle module [13] and written as files, which are later used to train and test the model.

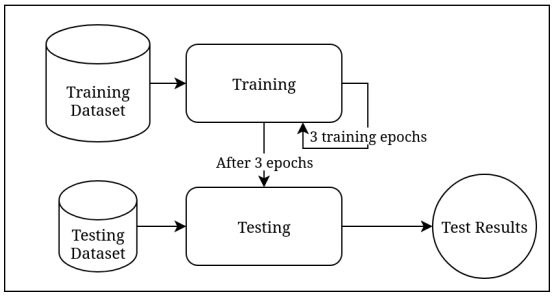


Figure 6: Overview of the GeqShift model pipeline showing how the training dataset is used during the training loop over three epochs, and the testing dataset being used for for testing before returning results.

2.4.2 GeqShift Model. The GeqShift model is built using PyTorch, PyTorch Geometric, and the e3nn Python libraries. The model runs in two main stages: training and testing. The training stage uses the training dataset provided by the previous dataset generation step, while the testing phase uses the split testing datasets for each class of saccharide. The model can only be trained to predict one of either ^{13}C or ^{1}H shifts, making it necessary to train and test two separate models on their respective shift type.

By default, the model trains during three epochs, where an epoch refers to a pass through the entire training dataset [41]. The batch size, which determines how many training examples the model is exposed to simultaneously before updating its internal learning parameters [41], is set to 32. The model automatically initiates the testing phase once training is complete. The mean and standard deviation of the training dataset is calculated before training begins, and is used to Z-score normalize and de-normalize the shift values before and after the model output respectively. Z-score normalization converts values from a normal distribution into Z-scores from a standardized normal distribution, improving learning performance and decreasing the number of epochs required for convergence [1].

The results are calculated as the mean absolute error (MAE) between expected and predicted shift values using the PyTorch `L1Loss()` criterion [15] for each test dataset saccharide class. The code to calculate the root mean squared error (RMSE), which was reported in the paper, is absent.

The model saves checkpoint files during the training phase which can be loaded for later predictions and testing without needing to retrain the model from scratch.

2.5 Alternative Carbohydrate Datasets

A large dataset containing high-quality carbohydrate structure and corresponding NMR spectra data is important for training machine learning models on carbohydrate NMR prediction.

Increasing the size of the dataset by making use of additional data sources is our primary interest. We consider two additional data sources: CSDB and GlycoNMR.

2.5.1 CSDB (Carbohydrate Structure Database). CSDB is a curated repository of bacterial, fungal, and plant glycans, featuring above 80% coverage of prokaryotes, fungi, and protista up to 2023. As of March 2025, it contains 32,937 carbohydrate structures and 19,728 NMR spectra [38].

Data export to various formats and encoding schemes is available on request, and web access is free. CSDB source data files can additionally be obtained and imported into a different database system, such as SQL [37]. Structure data is annotated with assigned NMR spectra and other information such as the temperature and solvent used in the experiment, where available. Structures are encoding using CSDB Linear notation, an unambiguous, human- and machine-readable linear string representation for carbohydrates [36]. An example of a data record from CSDB is shown in Figure 7.

```
 CSDB_ID: 313
CNMR SPECTRUM: #3_bDGlcp  105.4 74.9 77.1 71.0 77.4 ? //
         #_bDGalp  97.8  72.5 84.2 70.0 76.4 62.7 //
 TEMPERATURE: 298
 SOLVENT: D2O
HNMR SPECTRUM: #3_bDGlcp  4.65 3.35 3.47 3.40 3.42 ? //
          #_bDGalp  4.60 3.62 3.77 4.16 3.69 3.70 //
 STRUCTURE: bDGlcp(1-3)bDGalp
```

Figure 7: CSDB data record for bDGlcp(1-3)bDGalp, containing the CSDB record ID, carbon and hydrogen NMR spectra in CSDB NMR format, NMR experiment temperature in Kelvin, and NMR experiment solvent.

2.5.2 GlycoNMR. GlycoNMR is the first large-scale curated carbohydrate-specific NMR dataset, aiming to address the scarcity of appropriate carbohydrate data for machine learning [8]. It includes GlycoNMR.Exp (containing experimental ^{13}C and ^{1}H NMR shifts) and GlycoNMR.Sim (containing simulated shifts using GODESS, a web service for the prediction of carbohydrate NMR spectra [24–26]). In total, it contains 2,609 carbohydrate structures and 211,543 annotated NMR chemical shifts. A lack of raw spectra is a limiting factor, as experimentalists typically only report the peak positions of chemical shifts [8]. We only consider the experimental version which contains 299 carbohydrate molecules, including larger polysaccharides. We do not include simulated shifts in our training dataset as we believe that they may introduce additional error which can reduce the accuracy of the model.

Each data record in the GlycoNMR dataset consists of a Protein Data Bank (PDB) file, similar to the MDL MOL format, describing the carbohydrate molecular structure as well a CSV file containing the NMR chemical shifts.

3 Methods

In this section, we describe the methodology used to reproduce the results for GeqShift [6], then describe how we expanded the dataset and optimized the model to reduce memory consumption to counteract the increased memory demands of the expanded dataset. Finally, we describe our production of the example web API.

3.1 GeqShift Reproduction

We first attempt to reproduce the results shown in the GeqShift paper. We used the code located in the GeqShift GitHub repository [5]. The code contained numerous discrepancies and errors, requiring changes to the code in order to get it to work. A list of discrepancies, errors, and the associated changes are detailed in Appendix A.

We generate the dataset for each of the ten folds used in ten-fold cross-validation by running `create_data.py` with the appropriate parameters to load the provided input dataset while varying the `--fold` parameter from 0 through 9. Existing dataset directories are renamed between folds to ensure that they are not overwritten.

We then run `main.py`, specifying the appropriate parameters to load the dataset. The model is trained and then automatically tested by the software with the error results for each test dataset being output. This process is repeated for each fold used in ten-fold cross-validation for both ^{13}C and ^{1}H NMR shift types.

Due to difficulties encountered in attempting to reproduce the original results, GeqShift was tested on multiple machines with varying processors and operating systems, as well as with many varying parameters. We compare against the results exhibiting the lowest mean absolute error we were able to produce for our showing of GeqShift while using the model and dataset generation parameters from the original paper.

3.2 Expanded Dataset

We form our combined mixed-distribution dataset, hereafter referred to as Mixed, by combining the original CASPER data with data from two additional data sources: CSDB and GlycoNMR.

3.2.1 CSDB. We obtained a dump of the database for all molecules with experimentally recorded NMR ^{13}C and ^{1}H chemical shifts through special request. The dump file lists 7,073 carbohydrate molecules by CSDB ID in CSDB Linear format, along with their associated shifts and NMR experiment temperature and solvent.

As GeqShift expects data in SDF format, the CSDB Linear encoding must first be translated. CSDB provides a web-based API to facilitate translation to many formats including MDL MOL using their REStLESS software [7]. We used this translation service via a Python script to retrieve 13,686 MOL files. Some of the files contained errors, often due to being too large or having too many ambiguities. We filtered out these invalid files.

We processed the data using RDKit and Python scripting to form a sub-dataset consisting of 33 monosaccharides, 105 disaccharides, 128 trisaccharides, 115 tetrasaccharides, and 306 oligosaccharides giving a total of 687 carbohydrate molecules with 14,380 ^{13}C shifts and 13,646 ^{1}H shifts. We filtered out molecules for which less than 80% of ^{13}C or ^{1}H NMR shifts are known and those for which the ^{1}H shift variance exceeded 0.1. NMR experiments performed in a solution other than D_2O, a common aqueous solution used in NMR spectroscopy, were filtered out. We additionally filtered out those with atom types that are not recognized by the GeqShift data processing software. NMR chemical shifts are mapped to atom index values using mappings provided in the obtained MOL files. Some molecules contained a placeholder atom signifying a linkage point, used for forming repeating chains of the same molecules [10]. To allow for conformation generation, we removed these placeholder atoms as part of processing and treated these structures as non-repeating units.

3.2.2 GlycoNMR Dataset. We obtained the GlycoNMR.Exp data for 299 carbohydrate molecules from the download link provided on the GitHub page [7].

We processed the data using RDKit with Python scripting to form a sub-dataset of 26 monosaccharides, 21 disaccharides, 16 trisaccharides, 11 tetrasaccharides, and 18 oligosaccharides giving 92 carbohydrate molecules with 1,466 ^{13}C shifts and 1,232 ^{1}H shifts. We applied the same filtering rules as in Section 3.2.1. We mapped the NMR chemical shifts using the mappings and residue namings provided.

3.2.3 Combined Dataset. The combined dataset merged the three datasets into a single SDF file - the format used by GeqShift, requiring minimal changes to the existing code. RDKit was used to tag each molecule with the name of the source dataset and saccharide class. De-duplication, a technique that removes duplicate data, was performed by generating a canonical Simplified Molecular Input Line Entry System (SMILES) [44] string for each molecule using RDKit and comparing these strings across molecules. A SMILES string encodes the molecular structure in a simple string format, where two identical molecules will have the same SMILES string. GlycoNMR data was de-duplicated against CASPER data, while CSDB data was de-duplicated against both the CASPER and GlycoNMR data. The dataset is randomly shuffled before being written as an SDF file.

GeqShift was then modified to accept this single combined SDF file, filter by SDF source tag, and generate separate datasets for each test case. Modifications include those to additionally handle test data for tetra- and oligosaccharides, which are saved alongside the other classes of saccharide. To avoid data leakage, where testing data is included in the training data, each sub-dataset is split for training and testing before the datasets are combined. This ensures an equal distribution of testing data across datasets within the same fold. The datasets generated are the CASPER dataset as used by the original GeqShift model and the Mixed dataset containing the combined CASPER, GlycoNMR, and CSDB data.

Due to the conformation generation time requirements and failure rate becoming unacceptably high, we disable the RDKit parameter `useSmallRingTorsions`. This parameter controls whether RDKit will impose small ring torsion angle preferences on the molecule [16], which could potentially increase conformation generation accuracy by modeling the conformations of small rings more accurately. Disabling this parameter, as is the RDKit default, resolves these issues, which we believe are exacerbated by the presence of larger polysaccharides in the Mixed dataset. To further reduce the time required for conformation generation, we change the RDKit parameter `numThreads` from 10 to 0, allowing RDKit to use the

maximum number of processor threads supported by the system for conformation generation [16], providing a roughly three-fold speedup on a system with 40 logical processors. We generate all datasets we use with these settings to ensure a fair comparison.

Dataset files are written for both CASPER and Mixed datasets for each fold. The distribution of saccharides is visualized in Figure 8.

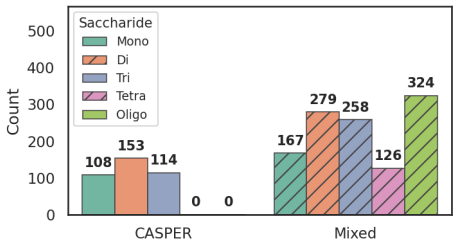


Figure 8: Datasets: a comparison of the molecule count per saccharide class for the datasets used in training and testing, where CASPER is the original dataset used in GeqShift and Mixed is the new dataset constructed from a combination of CASPER, GlycoNMR, and CSDB datasets.

3.3 Optimized Model

As the larger Mixed dataset results in a significant increase in training time and memory usage, we optimize the GeqShift model by implementing automatic mixed precision (AMP) training. AMP accelerates training and reduces memory consumption by performing certain operations in lower precision while keeping critical operations in higher precision to preserve accuracy [14]. Our aim with the optimized model, dubbed SugarRush, is to reduce memory consumption and maintain or improve accuracy while training in a similar time to the original GeqShift model.

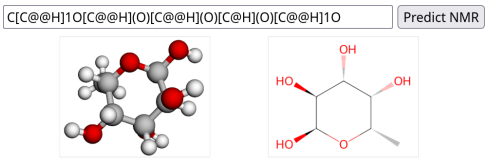
AMP requires the use of gradient scaling via `GradScaler` [14], as well as gradient clipping via `clip_grad_norm_()` [14] to prevent gradients from underflowing during training [14]. Gradient scaling temporarily multiplies loss values to keep gradients in a numerically stable range, while underflowing refers to gradients becoming too small to represent in floating-point precision. We implement AMP in the training phase using `torch.amp.autocast()` [14]. Gradient scaling is applied during the backward step (the phase where gradients are computed through backpropagation), with gradient clipping performed before the step operation using a maximum norm of 1.0. These optimizations are only applied during the training phase. We further optimize performance by removing statements that empty the GPU cache before every batch, instead, we empty the cache only once before each epoch to reduce memory allocation overhead.

The default number of epochs was adjusted from three to six to approximately match the running time of the original GeqShift model.

Finally, we modernized the versions of software used in the Python environment, enabling AMD GPUs to be used for training with AMD ROCm as well as easing development and deployment of the model on modern systems where setting up an older environment can be considerably frustrating. We have verified this to not significantly impact accuracy.

3.4 Web API

To enable integration with existing web services and make the model more accessible and generally useful to researchers, we build an example web API to use the model programmatically. Previously, the model did not provide an interface to easily obtain prediction results for a given carbohydrate. The web API makes the model usable for arbitrary predictions outside of controlled experiments. An example web front-end, shown in Figure 9, demonstrates how the web API could be integrated into existing websites.



	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

Figure 9: Example web front-end showing an input SMILES string and output predicted NMR shifts along with 3D and 2D representations of the molecule.

The web API is built using the Python Flask framework [12] with the Gunicorn web server [11]. It accepts an HTTP POST request at the `/predict` endpoint containing a JSON request with a single SMILES string in the `smiles` field. RDKit is used to convert the SMILES string into a molecule and generate conformations as described in Section 2.4.1. The model predicts ^{13}C and ^{1}H NMR shifts for the molecule which are then mapped to atom indices and returned in JSON format. The web API loads an existing model checkpoint file created from a previous training session. Figure 10 provides a high-level overview of the web API.

We attempted to export the model into an optimized format suitable for production inference deployment, however, challenges arose around the architecture of the model and redesign was considered out-of-scope for this work.

The model will automatically use any suitable GPU with CPU fallback. We believe the performance of the web API to be reasonable, returning results for various disaccharides within 3 seconds of request on a system with no available GPU. The software uses approximately 5 GiB of memory when using the SugarRush model.

3.5 Materials

3.5.1 Data and ML Models. We used the CASPER dataset of 375 carbohydrate molecules originally used in GeqShift [6]. In addition, we used data from the GlycoNMR dataset and data from CSDB as well as its web API for additional data retrieval.

We used the original GeqShift model as provided in the code hosted on GitHub.

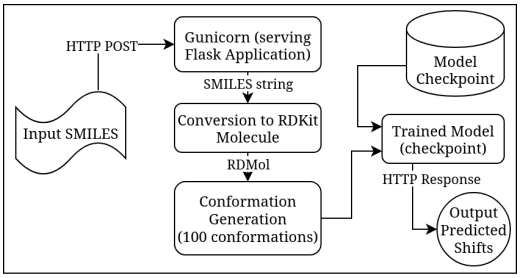


Figure 10: Overview of the web API pipeline. The Input SMILES string is sent through an HTTP POST method to the Gunicorn web server, it is converted to an RDKit molecule and 100 conformations are generated. The model, loaded from a checkpoint, is used to predict the NMR shifts which are returned via HTTP.

3.5.2 Software. Python 3.12.1, PyTorch 2.7.1 with NVIDIA CUDA 12.8, PyTorch Geometric 2.6.1, PyTorch Scatter 2.1.2, PyTorch Cluster 1.6.3, e3nn 0.5.6, RDKit 2025.3.5, Flask 3.1.1, and Gunicorn 23.0.0 were the software versions used for training and testing on the UCT HPC cluster running Rocky Linux 9.5 (kernel version 5.14.0) and NVIDIA kernel driver 575.51.03 with CUDA 12.9. The same software versions were used for local development, training, and testing with the exceptions of Python 3.13.7, PyTorch 2.7.1 with AMD ROCm 6.3, and Arch Linux (kernel version 6.16.1).

3.5.3 Hardware. UCT HPC ada and l40sfree cluster nodes were used for training and testing. The ada nodes were used for dataset generation and contained 48 Xeon 6442 CPU cores with 384 GB of memory. The l40sfree nodes were used for model training and testing and contained the same CPUs as the ada nodes, but with 512 GB of memory and four NVIDIA L40S 48 GB GPUs, of which one was used. A workstation containing 24 logical AMD Ryzen 9 5900X processors, 32 GB of memory, and an AMD Radeon RX 7900 GRE 16 GB GPU was used for local development, training, and testing.

3.6 Evaluation Metrics

We use Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and standard deviation for both forms of measurement as our primary evaluation metrics. We chose these metrics to match those used in the GeqShift paper. MAE calculates the average of absolute differences while RMSE calculates the square root of the average of the squared differences, giving larger errors an increased weight, thus making RMSE more sensitive to outlier values. MAE is obtained through the PyTorch `L1Loss()` criterion, while RMSE is obtained through the square root of the PyTorch `MSELoss()` criterion, both operating on the model outputs during the testing phase.

Ten-fold cross-validation is used to match GeqShift [6] for comparison. The dataset is split into ten equal parts, with nine parts used for training and one part used for testing, giving a train/test split of (90%/10%). The specific subset chosen for the test part is different for each of the ten folds. The model is trained and tested ten times, once for each dataset fold and the results averaged to produce the final results.

We consider a reduction in MAE and RMSE when testing on larger carbohydrate structures, specifically when testing on the mixed-distribution data, to be a successful outcome.

4 Results and Discussion

We compare the accuracy of the models across datasets, giving in-distribution prediction accuracies, where the testing data is within the same distribution as the training data. We then extend to comparing mixed-distribution prediction accuracies, where portions of the testing data are from a different dataset entirely to better evaluate the model's ability to generalize to new data. Finally, we include out-of-distribution testing on longer polysaccharides by evaluating and comparing the models on two complex oligosaccharides from the *Escherichia coli* O125 serogroup [2, 4].

4.1 GeqShift

Figure 11a shows box plots of mean absolute error (MAE) for ^{13}C and ^{1}H shift predictions on the CASPER test dataset. When trained on CASPER, the GeqShift model achieves $\delta_C = 0.38$ ppm and $\delta_H = 0.034$ ppm. These values are notably higher than the originally reported errors of $\delta_C = 0.28$ ppm and $\delta_H = 0.031$ ppm (shown by dotted lines). Our results are comparable for ^{1}H disaccharides, matching GeqShift with $\delta_H = 0.026$ ppm, but are considerably higher than the reported results in all other cases, exhibiting much larger RMSE and standard deviation (Table 4, Appendix B). In particular, when predicting ^{13}C shifts for trisaccharides, RMSE increases by more than two standard deviations from the reported results. We believe these higher than expected errors may be caused by discrepancies between the parameters described in the paper and the code provided on GitHub as well as our modifications made to make the software functional (detailed in Appendix A).

When GeqShift is trained on the Mixed dataset, errors drop from $\delta_C = 0.38$ ppm and $\delta_H = 0.034$ ppm to $\delta_C = 0.36$ ppm and $\delta_H = 0.032$ ppm when testing on the CASPER dataset (Figure 11a), showing a modest improvement in accuracy. Despite an overall improvement, accuracy deteriorates slightly when testing on ^{13}C monosaccharides, where it increases from $\delta_C = 0.41$ ppm to $\delta_C = 0.44$ ppm (Table 4, Appendix B). Accuracy is improved over GeqShift trained on CASPER data in all other cases, particularly for ^{13}C trisaccharides where the standard deviation for RMSE reduces by over five-fold. Although MAE is less sensitive to outlier values than RMSE, the effects of lower variance can still be seen in the box plot (Figure 11a), where GeqShift (Mixed) features a shorter box and whiskers.

As we move to testing on the Mixed dataset, which introduces tetra- and oligosaccharides, prediction accuracy worsens for all model and dataset combinations (Figure 11b), making the impact of dataset diversity more clear. The Mixed dataset is considerably larger, over three-fold, as compared the original CASPER dataset. In terms of shift count, it contains more than eight times the number of shifts. GeqShift trained only on CASPER data produces errors of $\delta_C = 2.35$ ppm and $\delta_H = 0.134$ ppm, while training on the Mixed dataset improves errors by approximately 2.5-fold to $\delta_C = 0.87$ ppm and $\delta_H = 0.060$ ppm, respectively. The model trained on the CASPER dataset suffers a dramatic loss of accuracy, with MAE increasing from $\delta_C = 0.38$ ppm and $\delta_H = 0.034$ ppm to $\delta_C = 2.35$

ppm and $\delta_H = 0.134$ ppm, a more than six-fold increase for ^{13}C and an almost four-fold increase for ^{1}H. By contrast, GeqShift trained on the Mixed dataset better maintains its accuracy, showing a much lower (over 2-fold less) increase in error. We attribute this improved generalization to the broader structural diversity present in the Mixed dataset, which enables the model to learn more transferable representations of carbohydrate NMR environments.

These results are expected as the CASPER test data, although withheld from the training data, is still within the same distribution. When adding out-of-distribution data, such as in the case of testing the model trained on only the CASPER data on the Mixed dataset, we observe a substantial increase in error. We believe this is caused by the model struggling to generalize to out-of-distribution data, especially in the case of larger polysaccharides such as tetra- and oligosaccharides which are not present in the CASPER data. The Mixed dataset consists of mostly oligosaccharides, featuring almost as many as the CASPER dataset features molecules in total. The errors produced by GeqShift trained on CASPER data for tetra- and oligosaccharides are significantly higher than for other saccharides (Table 4, Appendix B), demonstrating the importance of dataset diversity for improved generalization. Notably, training on the Mixed datasets yields an almost four-fold improvement in accuracy when predicting ^{13}C shifts for oligosaccharides (Table 5, Appendix B). This raises an important limitation: a model trained on a dataset containing only smaller saccharides struggles to generalize to more complex carbohydrates.

4.2 SugarRush

Our optimized SugarRush model shows modest improvements in most cases over the GeqShift model. It achieves better results than GeqShift for ^{1}H (Figure 11), with lower measured MAE for all classes of saccharide (Tables 4 and 5, Appendix B). It does, however, show worse performance when trained on CASPER data and performing ^{13}C predictions on the Mixed dataset (Figure 11b), with MAE increasing slightly from $\delta_C = 2.35$ ppm to $\delta_C = 2.36$ ppm. The figure shows increased variance with a larger box and longer whiskers. An opposite effect can be observed for ^{1}H where the accuracy improves and variance is reduced. When both training and testing on the Mixed dataset, the prediction accuracy is improved in every case except for ^{13}C monosaccharides, where MAE increases from $\delta_C = 0.99$ ppm to $\delta_C = 1.04$ ppm (Table 4, Appendix B). Figure 11a shows that SugarRush slightly outperforms GeqShift, achieving an MAE of $\delta_H = 0.030$ ppm, which surpasses that of $\delta_H = 0.031$ ppm as reported for GeqShift. The improvements are most pronounced for ^{1}H shifts, where the tighter distribution of errors indicates more stable predictions.

Beyond accuracy, training efficiency is an important factor for accessibility. Figure 12 compares the peak GPU memory usage of GeqShift and SugarRush during the first 200 training batches. Our model achieved better memory performance and training time than GeqShift, reducing the memory consumption from 16.26 GiB to 9.53 GiB and from 23.60 GiB to 12.85 GiB when training on the smaller CASPER and larger Mixed datasets, respectively. The introduction of automatic mixed precision results in a near two-fold reduction in memory usage on an NVIDIA GPU compared to the original GeqShift model in the case of both CASPER and Mixed training data (Figure 12). Testing on an AMD GPU shows comparable improvements for CASPER training data (Table 3, Appendix B). Notably, training the SugarRush model on the larger Mixed dataset requires less memory than training the GeqShift model on the smaller CASPER dataset. SugarRush demonstrates sub-16 GiB memory usage when training on the larger mixed dataset, enabling its use on GPUs with 16 GB of memory where the original model would otherwise not be able to run. As the size of the dataset grows, the demand placed on memory increases. We believe that it is important to minimize these requirements to allow a wider audience of researchers access to train these models without specialized compute infrastructure.

We observe an almost three-fold improvement in prediction accuracy using SugarRush trained on the Mixed dataset compared to GeqShift trained on the CASPER dataset, when testing on the Mixed dataset across all saccharide classes, while using substantially less memory. SugarRush achieves an over four-fold improvement at predicting ^{13}C shifts for oligosaccharides (Table 5, Appendix B), greater than that achieved by the GeqShift model without any optimizations.

4.3 Out-of-distribution Testing

To test robustness beyond both training datasets, we extend our testing to include two out-of-distribution polysaccharides from the *E. coli* O125 serogroup, specifically the O125ab [2] and O125ac [4] oligosaccharides containing six and five residues respectively. We observe significantly improved accuracy, roughly two-fold, with errors dropping from $\delta_C = 3.72$ ppm and $\delta_H = 0.27$ ppm with the GeqShift model trained on the CASPER dataset to $\delta_C = 1.57$ ppm and $\delta_H = 0.19$ ppm with the SugarRush model trained on the Mixed dataset as shown in the bar plots in Figure 14.

The scatter plots in Figure 14 show how SugarRush (Mixed) produces more consistent predictions with lower variance than GeqShift (CASPER). The predictions are more tightly clustered and exhibit less spread for both shift types. Notably, an outlier can be seen in the ^{1}H predictions for GeqShift (CASPER) in Figure 14a which is absent when using the model trained on the Mixed dataset. This outlier originates from the predictions for the O125ac oligosaccharide [4], specifically the sixth carbon (C6) of its *α-D-GalpNAc* residue. The experimental ^{1}H shift value was recorded as 3.80 ppm while the GeqShift (CASPER) model predicted 2.63 ppm, producing an error of 1.17 ppm. Our SugarRush (Mixed) model predicted 3.78 ppm, resulting in a substantially lower error of 0.02 ppm. ^{13}C predictions on the same residue's fifth carbon result in a larger error of 21.91 ppm from GeqShift (CASPER) compared to 3.27 ppm from our SugarRush (Mixed) model. Similar errors were produced for the *β-D-GalpNAc* and *β-D-Galp* residues of the O125ab oligosaccharide. The ^{13}C scatter plot for GeqShift (CASPER) shows a tendency for the model to overpredict ^{13}C shifts and underpredict ^{1}H shifts.

The per-atom averages in Figure 13 show lower errors and more consistent predictions from SugarRush (Mixed) compared to GeqShift (CASPER) for both ^{13}C (upper) and ^{1}H (lower) shifts across both oligosaccharides. GeqShift (CASPER) ^{13}C predictions on the O125ab oligosaccharide deviate by more than 7.5 ppm from experimental values on the fifth carbon, while the SugarRush (Mixed)

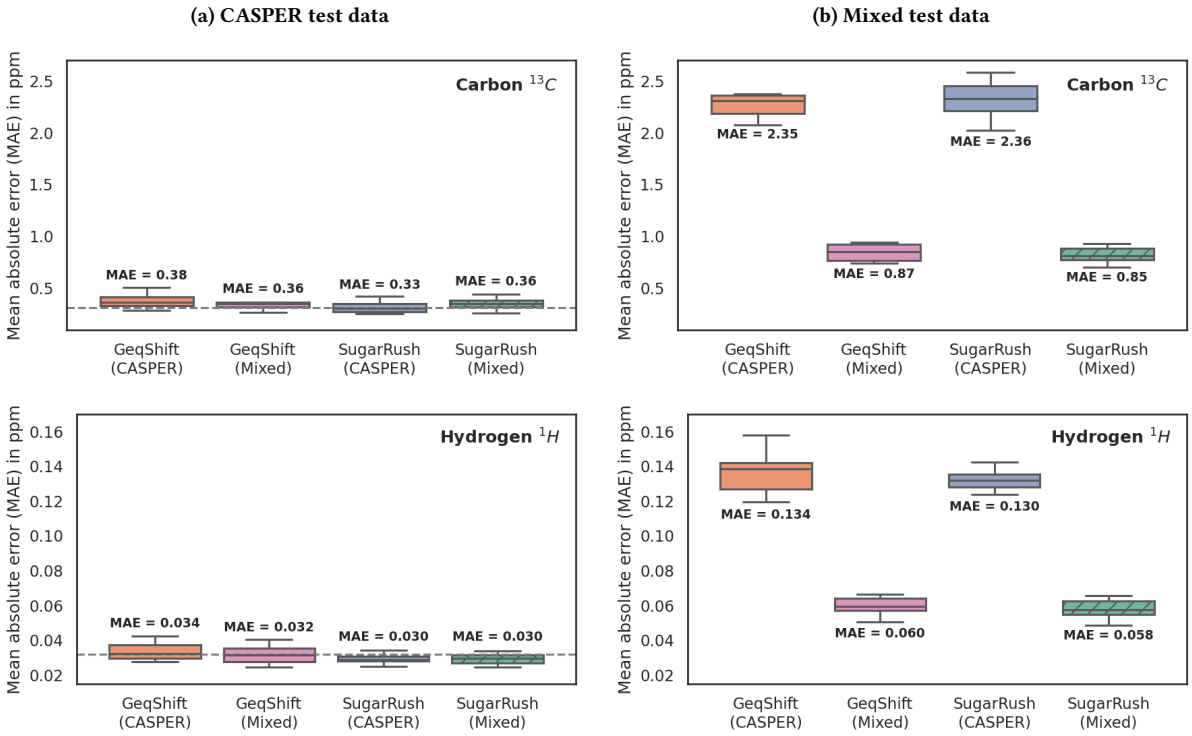


Figure 11: Comparison of test prediction accuracy on CASPER (a) and Mixed (b) test data between the models in mean absolute error for ^{13}C and ^{1}H shifts, visualized using box plots. The items in parenthesis indicate the dataset used to train the model. The dotted lines show the reported mean absolute errors, $\delta_C = 0.28$ ppm and $\delta_H = 0.031$ ppm, in GeqShift [6].

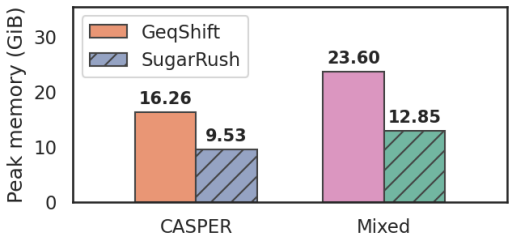


Figure 12: Model optimization: comparison of peak memory usage in gibibytes (GiB) during the first 200 training batches using an NVIDIA L40S GPU. The models are trained on the full size of the dataset without a test split.

model deviates by less than 0.7 ppm. The figures show over- and underprediction behavioral trends for GeqShift (CASPER) similar to that of the scatter plots in Figure 14, while SugarRush (Mixed) maintains predictions much closer to the experimental values. The marked increase in not only accuracy but also prediction consistency between atoms is especially important for structural elucidation of complex carbohydrates.

These results show improved generalization ability, specifically to larger oligosaccharides, using the Mixed dataset compared to the CASPER dataset. We believe this to be due to the inclusion of larger polysaccharides in the Mixed dataset as well as a more diverse spread of examples. To confirm that these polysaccharides are not within the training distribution, we used the same canonical SMILES string de-duplication method presented in Section 3.2.3. These out-of-distribution tests were performed using models trained on the full dataset without a testing split, loaded into the web API described in Section 3.4, and run on the AMD GPU listed in Section 3.5.3.

5 Conclusions

We have improved the generalization ability and general usability of the model. Increasing the amount of training data can result in increased prediction accuracy and generalization ability of machine learning models, including GNN models trained to predict chemical properties such as GeqShift. Our results show that it is possible to increase the dataset size and achieve improved results for prediction accuracy in terms of mean absolute error, as well as improved generalization ability on out-of-distribution data. SugarRush trained on the Mixed dataset achieves up to a fourfold reduction in ^{13}C errors for complex oligosaccharides, including an over two-fold improvement in prediction accuracy for the two oligosaccharides from the O125 serogroup, maintains strong generalization to unseen samples, and reduces GPU memory usage nearly twofold compared to GeqShift trained on the CASPER dataset.

Although the mixed dataset is substantially larger, we believe future developments can increase the size, diversity, and quality of the dataset further. A technique of particular interest is to improve

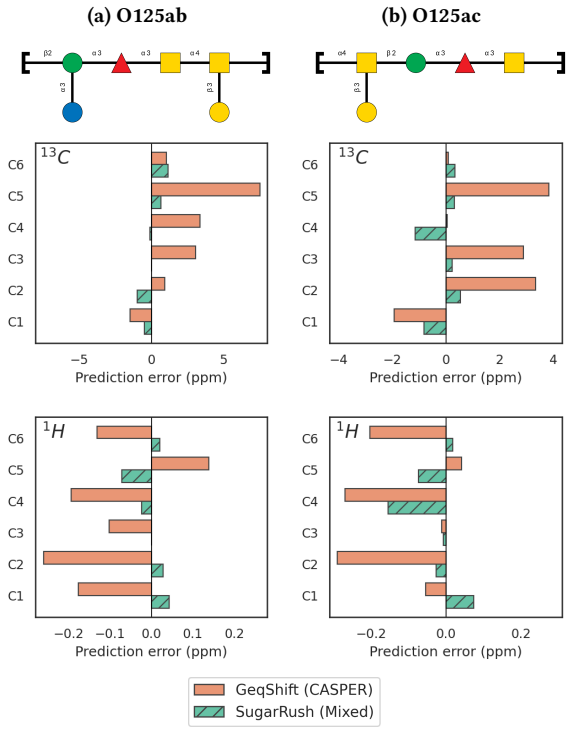


Figure 13: Comparison of prediction errors for ^{13}C (upper) and ^{1}H (lower) shifts between models for two polysaccharides from the *E. coli* O125 serogroup, O125ab (left) and O125ac (right) [2, 4]. The polysaccharides are visualized according to the SNFG standard. The plots show the mean absolute errors from experimental values on a per-atom basis.

the quality of the conformer generation by making use of more fully-featured molecular dynamics simulations using force-fields suitable for carbohydrates such as CHARMM or AMBER. This technique may additionally allow for the specification of the temperature and solvent used in the original NMR experiment, available for much of the data used, which could potentially increase conformer generation accuracy. We believe that improving the quality of the generated conformers, and thus the data augmentation method, could improve the generalization ability and accuracy of the model.

In addition, through reducing the memory consumption significantly and building an example web API, we show that it is possible to make the model more accessible to researchers. We believe that the web API demonstrates how machine learning models such as GeqShift can be deployed on contemporary machines and easily used by researchers and web services.

The challenges encountered in our attempt to reproduce the original GeqShift results highlight the importance of thorough documentation of software and parameters, ensuring that the code used for the presented results is readily available in an already-working state, as well as employing appropriate software development methods to ensure high-quality, easily understandable and maintainable software.

There remain avenues for improving the quality of conformation generation, as well as improving the performance of the model.

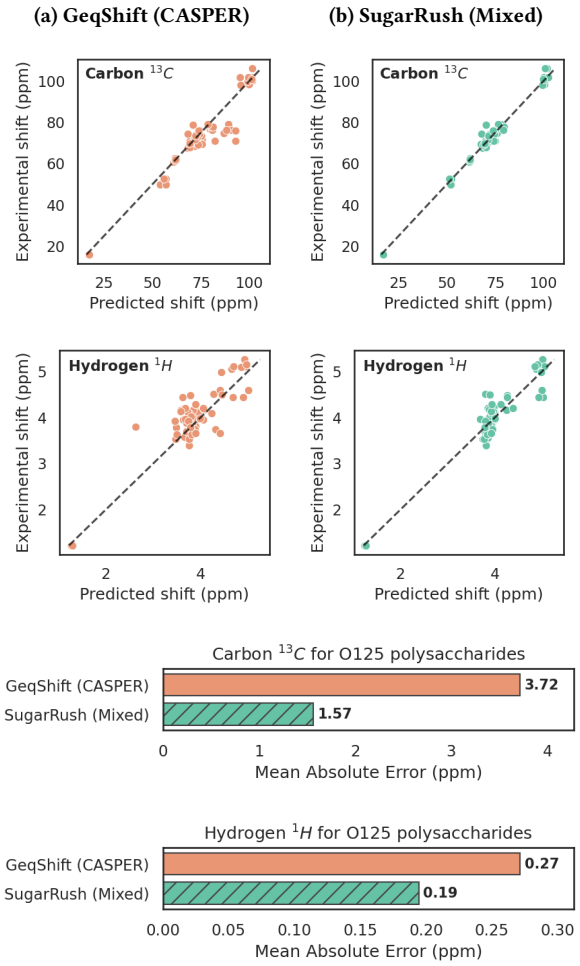


Figure 14: Comparison of prediction errors for ^{13}C (upper) and ^{1}H (lower) shifts between models for two polysaccharides from the *E. coli* O125 serogroup. Scatter plots show the relationship between experimental and predicted NMR shift values and bar plots show the average errors across both polysaccharides.

More precise data filtering techniques may further improve the model's accuracy. A more robust web API can be developed with enhanced functionality including residue mapping and alternative input formats, as well as further reduced resource consumption through an optimized model format. Integration of the web API into carbohydrate web services could spur the continued development and adoption of machine learning based NMR prediction techniques, as well as provide more accurate NMR predictions to a wider audience of scientists, accelerating fields in chemical research.

Code Availability

The code is available on the UCT GitLab repository at
https://gitlab.cs.uct.ac.za/bllcha013/sugarrush

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, Philip Toukach from the Carbohydrate Structure Database for assistance in obtaining data, and Göran Widmalm et al. for their contribution with GeqShift, upon which this research is based. Computations were performed using facilities provided by the University of Cape Town's ICTS High Performance Computing team: hpc.uct.ac.za

References

[1] Mohammed Al-Faiz, Ali Ibrahim, and Sarmad Hadi. 2019. The effect of Z-Score standardization (normalization) on binary input due the speed of learning in back-propagation neural network. *Iraqi Journal of Information & Communications Technology* 1 (02 2019), 42–48. doi:10.31987/ijict.1.3.41

[2] Andrej WEINTRAUB Alexandra KJELLBERG, Felipe URBINA and Goran WIDMALM. 1996. Structural analysis of the 0-antigenic polysaccharide from the enteropathogenic Escherichiu coli 0125. *Eur. J. Biochem* 239 (1996), 532–538.

[3] Pierre Avenas. 2012. Etymology of Main Polysaccharide Names. In *The European Polysaccharide Network of Excellence*. 13–21.

[4] Göran Widmalm Axel Furevi, Klas I Udekwu. 2022. Structural elucidation of the O-antigen polysaccharide from Escherichia coli O125ac and biosynthetic aspects thereof. *Glycobiology* 32 (2022), 1089–1100.

[5] Maria Bankestad. 2024. GeqShift: Highly accurate chemical shift prediction using a geometrically aware graph neural network. https://github.com/mariabankestad/GeqShift/. GitHub repository.

[6] Maria Bånkestad, Keven M. Dorst, Göran Widmalm, and Jerk Rönnols. 2023. Carbohydrate NMR chemical shift predictions using E(3) equivariant graph neural networks. arXiv:2311.12657 [cs.LG] https://arxiv.org/abs/2311.12657

[7] Zizhang Chen. 2024. GlycoNMR: Data repository for Glycan NMR Chemical shift. https://github.com/Cyrus9721/GlycoNMR. Retrieved July 24, 2025.

[8] Zizhang Chen, Ryan Paul Badman, Lachele Foley, Robert Woods, and Pengyu Hong. 2024. GlycoNMR: DATASET AND BENCHMARKS FOR NMR CHEMICAL SHIFT PREDICTION OF CARBOHYDRATES WITH GRAPH NEURAL NETWORKS. *cs.LG* 2311, 17134v2 (2024).

[9] Gita Cherian. 2019. *A Guide to the Principles of Animal Nutrition* (version 0.11 ed.). Oregon State University, Corvallis, OR.

[10] Ivan Yu. Chernyshov and Philip V. Toukach. 2018. REStLESS: automated translation of glycan sequences from residue-based notation to SMILES and atomic coordinates. *Bioinformatics* 34 (2018), 2679–2681.

[11] Gunicorn Contributors. 2025. Gunicorn - Python WSGI HTTP Server for UNIX. https://gunicorn.org/

[12] Pallets Contributors. 2025. Flask - The Python micro framework for building web applications. https://github.com/pallets/flask/

[13] Python Contributors. 2025. pickle - Python object serialization. https://docs.python.org/3/library/pickle.html

[14] PyTorch Contributors. 2025. PyTorch Documentation - Automatic Mixed Precision package - torch.amp. https://docs.pytorch.org/docs/stable/amp.html

[15] PyTorch Contributors. 2025. PyTorch Documentation - L1Loss. https://docs.pytorch.org/docs/stable/generated/torch.nn.L1Loss.html

[16] deric4, Gianluca Sforna, Greg Landrum, Hans De Winter, and RDKit Community. 2025. The RDKit Documentation - rdkit.Chem.rdDistGeom module. https://www.rdkit.org/docs/source/rdkit.Chem.rdDistGeom.html

[17] deric4, Gianluca Sforna, Greg Landrum, Hans De Winter, and RDKit Community. 2025. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/

[18] Carolina Fontana and Göran Widmalm. 2023. Primary Structure of Glycans by NMR Spectroscopy. *Chemical Reviews* 123, 3 (2023), 1040–1102. doi:10.1021/acs.chemrev.2c00580 arXiv:https://doi.org/10.1021/acs.chemrev.2c00580 PMID: 36622423.

[19] Will Gerrard, Lars A. Bratholm, Martin J. Packer, Adrian J. Mulholland, David R. Glowacki, and Craig P. Butts. 2020. IMPRESSION – prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* 11 (2020), 508–515. Issue 2. doi:10.1039/C9SC03854J

[20] Yanfei Guan, S. V. Shree Sowndarya, Liliana C. Gallegos, Peter C. St. John, and Robert S. Paton. 2021. Real-time prediction of 1H and 13C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* 12 (2021), 12012–12026. Issue 36. doi:10.1039/D1SC03343C

[21] Thomas A. Halgren. 1996. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* 17, 5-6 (1996), 490–519. doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291096-987X%28199604%2917%3A5/6%3C490%3A%3AAID-JCC1%3E3.0.CO%3B2-P

[22] Jongmin Han, Hyungu Kang, Seokho Kang, Youngchun Kwon, Dongseon Lee, and Youn-Suk Choi. 2022. Scalable graph neural network for NMR chemical shift prediction. *Phys. Chem. Chem. Phys.* 24 (2022), 26870–26878. Issue 43. doi:10.1039/D2CP04542G

[23] Eric Jonas, Stefan Kuhn, and Nils Schlörer. 2022. Prediction of chemical shift in NMR: A review. *Magnetic Resonance in Chemistry* 60, 11 (2022), 1021–1031. doi:10.1002/mrc.5234 arXiv:https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/mrc.5234

[24] Roman R. Kapaev, Ksenia S. Egorova, and Philip V. Toukach. 2014. Carbohydrate Structure Generalization Scheme for Database-Driven Simulation of Experimental Observables, Such as NMR Chemical Shifts. *Journal of Chemical Information and Modeling* 54, 9 (2014), 2594–2611. doi:10.1021/ci500267u arXiv:https://doi.org/10.1021/ci500267u PMID: 25020143.

[25] Roman R. Kapaev and Philip V. Toukach. 2015. Improved Carbohydrate Structure Generalization Scheme for 1H and 13C NMR Simulations. *Analytical Chemistry* 87, 14 (2015), 7006–7010. doi:10.1021/acs.analchem.5b01413 arXiv:https://doi.org/10.1021/acs.analchem.5b01413 PMID: 26087011.

[26] Roman R. Kapaev and Philip V. Toukach. 2016. Simulation of 2D NMR Spectra of Carbohydrates Using GODESS Software. *Journal of Chemical Information and Modeling* 56, 6 (2016), 1100–1104. doi:10.1021/acs.jcim.6b00083 arXiv:https://doi.org/10.1021/acs.jcim.6b00083 PMID: 27227420.

[27] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. 2020. Neural Message Passing for NMR Chemical Shift Prediction. *Journal of Chemical Information and Modeling* 60, 4 (2020), 2024–2030. doi:10.1021/acs.jcim.0c00195 arXiv:https://doi.org/10.1021/acs.jcim.0c00195 PMID: 32250618.

[28] Harvard Molecular Mechanics. 2025. CHARMM. https://www.academiccharmm.org

[29] Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. *Array* 16 (2022), 100258.

[30] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Data Augmentation Can Improve Robustness. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*. DeepMind, London, NeurIPS, Vancouver, Canada.

[31] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data Augmentation Can Improve Robustness. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 29935–29948. https://proceedings.neurips.cc/paper_files/paper/2021/file/fb4c48608ce8825b558ccf07169a3421-Paper.pdf

[32] Jerk Rönnols. 2013. *Structure, dynamics and reactivity of carbohydrates*. Ph. D. Dissertation. Stockholm University, Stockholm, Sweden.

[33] Allison Soult. 2019. Carbohydrate Structures. https://chem.libretexts.org/Bookshelves/Introductory_Chemistry/Chemistry_for_Allied_Health_(Soult)/05%3A_Properties_of_Compounds/5.02%3A_Carbohydrate_Structures. Retrieved March 26, 2025.

[34] The Widmalm Research Group, Stockholm University. 2012. CASPER. http://www.casper.organ.su.se/casper/. Retrieved March 23, 2025.

[35] 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

[36] 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.

[37] Philip V. Toukach and Ksenia S. Egorova. 2022. Source files of the Carbohydrate Structure Database: the way to sophisticated analysis of natural glycans. *Scientific Data* 9 (2022), 131.

[38] Philip V. Toukach, Ksenia S. Egorova, Yuri A. Knirel, et al. 2024. Carbohydrate Structure Database (CSDB). http://csdb.glycoscience.ru. Retrieved March 23, 2025.

[39] San Francisco University of California. 2025. The Amber Molecular Dynamics Package. https://ambermd.org

[40] Ajit Varki. 2016. Biological roles of glycans. *Glycobiology* 27, 1 (Dec. 2016), 3–49. doi:10.1093/glycob/cww086

[41] Akanksha Verma. 2025. Epochs, Batch, and Iterations in Deep Learning. https://medium.com/@akankshaverma136/epochs-batch-and-iterations-in-deep-learning-ed319565e85e. Retrieved September 8, 2025.

[42] Shuzhe Wang, Jagna Witek, Gregory A. Landrum, and Sereina Riniker. 2020. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *Journal of Chemical Information and Modeling* 60, 4 (2020), 2044–2058. doi:10.1021/acs.jcim.0c00025 arXiv:https://doi.org/10.1021/acs.jcim.0c00025 PMID: 32155061.

[43] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. 2025. A Comprehensive Survey on Data Augmentation. arXiv:2405.09591 [cs.LG] https://arxiv.org/abs/2405.09591

[44] David Weininger. 1987. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 28 (1987), 31–36.

[45] Ziyue Yang, Maghesree Chakraborty, and Andrew D. White. 2021. Predicting chemical shifts with graph neural networks. *Chem. Sci.* 12 (2021), 10802–10809. Issue 32. doi:10.1039/D1SC01895G

A Code Discrepancies and Errors

This section details the discrepancies and errors encountered in the GeqShift code, as well as the steps taken to remediate them.

A.1 Dataset Generation

A.1.1 Test Dataset Generation Error. The dataset creation code failed to create the test datasets due to a bug in the code that derives file names.

Remediation: removed the offending `[0]` list index.

A.1.2 Conformation Number Discrepancy. The dataset creation code was set to generate ten conformations for monosaccharides, one for disaccharides and one for trisaccharides. The paper states 100 conformations were used for GeqShift.

Remediation: Set all values to 100.

A.1.3 Conformation Parameter Error. The dataset creation code attempts to set a non-existent RDKit parameter, `params.maxAttempts`, which does not have any effect on the version used in GeqShift, and causes an exception on newer versions. Conformer generation often fails, producing empty conformer files which break the generation of test datasets.

Remediation: Set `params.maxIterations` instead of `params.maxAttempts`.

A.1.4 Missing Validation Set. The dataset creation code does not appear to make any provision for the 5% validation set mentioned in the paper for single conformer cases.

Remediation: ignore single-conformer cases. We focus on the final GeqShift model using 100 conformations.

A.2 Model

A.2.1 Learning Rate Discrepancy. The learning rate parameter was set to $1e-4$ in the code, while the paper states $3e-4$.

Remediation: Set the learning rate to $3e-4$ to match the stated learning rate in the paper.

A.2.2 GPU ID. The PyTorch device was hard-coded to device ID `0` which caused an exception on certain machines as the device ID may vary.

Remediation: implemented device auto-detection logic.

A.2.3 Missing Scheduler. The code is missing the ReduceLROnPlateau scheduler mentioned in the paper, used for single-conformer cases.

Remediation: ignore single-conformer cases. We focus on the final GeqShift model using 100 conformations.

A.2.4 Missing Validation. The model code does not appear to make any provision for the 5% validation set mentioned in the paper, used for single-conformer cases.

Remediation: ignore single-conformer cases. We focus on the final GeqShift model using 100 conformations.

A.2.5 Missing RMSE Results. The model code does not output the RMSE results. There does not appear to be any code to calculate the RMSE results.

Remediation: added PyTorch `MSELoss()` criterion and related code to additionally output RMSE results.

A.2.6 Missing Random Seed. The model code does not set a random seed, making the reproduction of results challenging due to stochastic behavior.

Remediation: set random seeds to fixed values where appropriate.

B Additional Figures and Tables

Additional figures and tables can be seen on the following pages.

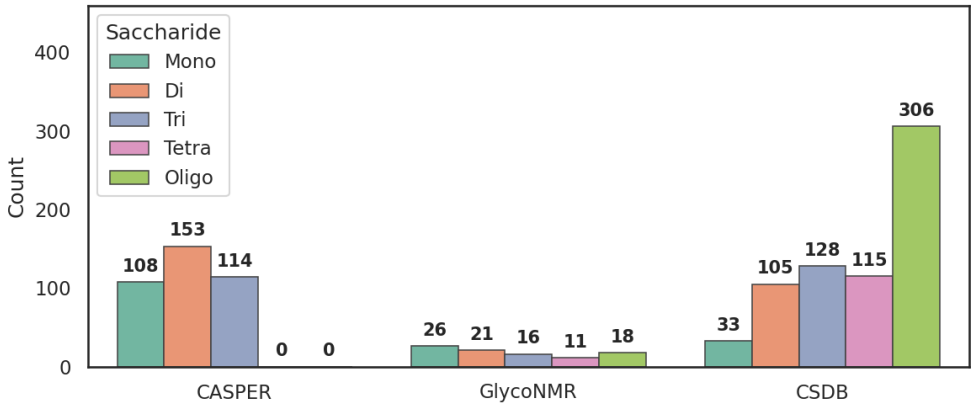


Figure 15: Comparison of base datasets used to construct the Mixed dataset, showing the number of saccharides in each class.

Dataset	Total	Monosaccharides	Disaccharides	Trisaccharides	Tetrasaccharides	Oligosaccharides	^{13}C Shifts	^{1}H Shifts
CASPER	375	108	153	114	0	0	4,955	4,856
GlycoNMR	92	26	21	16	11	18	1,466	1,232
CSDB	687	33	105	128	115	306	14,380	13,646
Mixed	1,154	167	279	258	126	324	20,801	19,734

Table 1: Comparison of processed datasets showing the number of saccharides in each class as well as the number of ^{13}C and ^{1}H NMR shifts. CASPER is the original dataset used in GeqShift. Mixed is the dataset constructed by combining the CASPER, GlycoNMR, and CSDB datasets.

Model	Training Dataset	Testing Dataset	Automatic Mixed Precision	Epochs
GeqShift (Claimed)	CASPER	CASPER	No	3
GeqShift (CASPER/CASPER)	CASPER	CASPER	No	3
GeqShift (Mixed/CASPER)	CASPER+GlycoNMR+CSDB	CASPER	No	3
GeqShift (CASPER/Mixed)	CASPER	CASPER+GlycoNMR+CSDB	No	3
GeqShift (Mixed/Mixed)	CASPER+GlycoNMR+CSDB	CASPER+GlycoNMR+CSDB	No	3
SugarRush (CASPER/CASPER)	CASPER	CASPER	Yes	6
SugarRush (Mixed/CASPER)	CASPER+GlycoNMR+CSDB	CASPER	Yes	6
SugarRush (CASPER/Mixed)	CASPER	CASPER+GlycoNMR+CSDB	Yes	6
SugarRush (Mixed/Mixed)	CASPER+GlycoNMR+CSDB	CASPER+GlycoNMR+CSDB	Yes	6

Table 2: An overview of the models with varying datasets used for training and testing and the number of training epochs as well as an indication of their use of optimizations. Models with multiple items in parenthesis represent the datasets used, with the first item indicating the training dataset and the second item the testing dataset.

Model	AMD GPU memory	NVIDIA GPU memory	Automatic Mixed Precision	Epochs
GeqShift (CASPER)	15.98 GiB	16.26 GiB	No	3
SugarRush (CASPER)	9.883 GiB	9.526 GiB	Yes	6
GeqShift (Mixed)	-	23.60 GiB	No	3
SugarRush (Mixed)	13.41 GiB	12.85 GiB	Yes	6

Table 3: Peak memory utilization per GPU in gibibytes (GiB) during training for the final models. The final models are trained on the full size of the training dataset without a test split. Dashed entries signify a failure to run the model due to insufficient available memory. Only the first 200 batches of the training phase of each model is considered. Specific GPUs used are shown in Section 3.5.3.

	Monosaccharides ^{13}C		Disaccharides ^{13}C		Trisaccharides ^{13}C	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
GeqShift (Claimed)	0.31 (0.08)	0.58 (0.18)	0.23 (0.06)	0.46 (0.19)	0.30 (0.09)	0.53 (0.16)
GeqShift (CASPER/CASPER)	0.41 (0.14)	0.80 (0.38)	0.29 (0.06)	0.68 (0.46)	0.43 (0.10)	0.89 (0.48)
GeqShift (Mixed/CASPER)	0.44 (0.17)	0.85 (0.72)	0.27 (0.05)	0.61 (0.49)	0.37 (0.07)	0.60 (0.09)
GeqShift (CASPER/Mixed)	2.18 (0.74)	8.59 (3.93)	1.52 (0.27)	5.32 (1.24)	1.59 (0.19)	4.96 (0.93)
GeqShift (Mixed/Mixed)	0.99 (0.34)	3.50 (2.39)	0.63 (0.08)	1.51 (0.34)	0.75 (0.08)	1.57 (0.38)
SugarRush (CASPER/CASPER)	0.40 (0.17)	0.97 (0.97)	0.24 (0.05)	0.59 (0.49)	0.34 (0.09)	0.71 (0.49)
SugarRush (Mixed/CASPER)	0.50 (0.20)	1.21 (1.09)	0.26 (0.05)	0.60 (0.48)	0.34 (0.07)	0.54 (0.09)
SugarRush (CASPER/Mixed)	2.15 (0.78)	8.60 (3.96)	1.50 (0.24)	5.43 (0.91)	1.57 (0.22)	5.05 (0.94)
SugarRush (Mixed/Mixed)	1.04 (0.35)	3.72 (2.37)	0.61 (0.06)	1.44 (0.30)	0.71 (0.08)	1.54 (0.29)
	Monosaccharides ^{1}H		Disaccharides ^{1}H		Trisaccharides ^{1}H	
Model	MAE	RMSE	MAE	RMSE	MAE	RMSE
GeqShift (Claimed)	0.035 (0.009)	0.057 (0.018)	0.026 (0.003)	0.044 (0.011)	0.033 (0.009)	0.052 (0.016)
GeqShift (CASPER/CASPER)	0.038 (0.010)	0.079 (0.052)	0.026 (0.003)	0.045 (0.010)	0.037 (0.007)	0.060 (0.020)
GeqShift (Mixed/CASPER)	0.036 (0.010)	0.071 (0.043)	0.025 (0.003)	0.043 (0.011)	0.034 (0.007)	0.053 (0.016)
GeqShift (CASPER/Mixed)	0.117 (0.045)	0.318 (0.136)	0.099 (0.015)	0.265 (0.052)	0.095 (0.010)	0.227 (0.036)
GeqShift (Mixed/Mixed)	0.059 (0.015)	0.144 (0.098)	0.049 (0.005)	0.106 (0.018)	0.056 (0.005)	0.113 (0.042)
SugarRush (CASPER/CASPER)	0.033 (0.008)	0.068 (0.049)	0.024 (0.004)	0.042 (0.012)	0.032 (0.006)	0.052 (0.014)
SugarRush (Mixed/CASPER)	0.033 (0.007)	0.057 (0.022)	0.025 (0.004)	0.042 (0.012)	0.032 (0.007)	0.051 (0.015)
SugarRush (CASPER/Mixed)	0.112 (0.041)	0.309 (0.144)	0.096 (0.016)	0.265 (0.056)	0.091 (0.010)	0.216 (0.043)
SugarRush (Mixed/Mixed)	0.053 (0.015)	0.125 (0.091)	0.047 (0.005)	0.100 (0.012)	0.054 (0.006)	0.114 (0.04)

Table 4: Prediction accuracy comparison for ^{13}C and ^{1}H shifts for mono-, di-, and trisaccharides across tested models. Models with multiple items in parenthesis represent the datasets used, with the first item indicating the training dataset and the second item the testing dataset. Results are shown as the ten-fold mean with standard deviation in parenthesis.

	Tetrasaccharides ^{13}C		Oligosaccharides ^{13}C	
Model	MAE	RMSE	MAE	RMSE
GeqShift (CASPER/Mixed)	2.92 (0.60)	7.44 (2.38)	3.53 (0.31)	7.81 (1.10)
GeqShift (Mixed/Mixed)	1.07 (0.65)	2.76 (3.17)	0.92 (0.10)	1.81 (0.53)
SugarRush (CASPER/Mixed)	3.05 (0.63)	7.86 (2.48)	3.52 (0.47)	7.88 (1.49)
SugarRush (Mixed/Mixed)	1.01 (0.65)	2.67 (3.21)	0.88 (0.11)	1.71 (0.42)
	Tetrasaccharides ^{1}H		Oligosaccharides ^{1}H	
Model	MAE	RMSE	MAE	RMSE
GeqShift (CASPER/Mixed)	0.154 (0.023)	0.284 (0.027)	0.208 (0.028)	0.363 (0.058)
GeqShift (Mixed/Mixed)	0.060 (0.012)	0.105 (0.019)	0.077 (0.009)	0.139 (0.019)
SugarRush (CASPER/Mixed)	0.152 (0.023)	0.279 (0.033)	0.199 (0.017)	0.340 (0.045)
SugarRush (Mixed/Mixed)	0.059 (0.012)	0.106 (0.023)	0.076 (0.010)	0.135 (0.023)

Table 5: Prediction accuracy comparison for ^{13}C and ^{1}H shifts for tetra- and oligosaccharides across tested models, excluding those for which these saccharides are absent from the test dataset. Models with multiple items in parenthesis represent the datasets used, with the first item indicating the training dataset and the second item the testing dataset. Results are shown as the ten-fold mean with standard deviation in parenthesis.