Carbohydrate molecule representations and datasets for machine learning

A Literature Review

Channing Bellamy
Department of Computer Science
University of Cape Town
Private Bag X3, Rondebosch,
7701, South Africa
bllcha013@myuct.ac.za

ABSTRACT

Carbohydrates are structurally complex molecules that are ubiquitous in nature. Their complexity poses challenges for representing them in a standard format suitable for machine learning. Stereochemistry along with other factors make carbohydrates difficult to accurately represent. Existing representational formats represent carbohydrates with varying accuracy, and datasets use differing representational formats. In this review, we examine the complexities of carbohydrate structure to understand why they can be difficult to represent. We then explore various representational formats for carbohydrates and compare available carbohydrate NMR datasets. Finally, we discuss data augmentation and synthetic data generation strategies to increase dataset size.

KEYWORDS

Carbohydrates, Carbohydrate Structure, Glycans, Chemistry, Carbohydrate Representation, Datasets, Databases, Data Augmentation, Synthetic Data Generation, Machine Learning (ML), Nuclear Magnetic Resonance (NMR)

1. INTRODUCTION

Carbohydrates are structurally complex molecules that are ubiquitous in biological systems, important in cell-cell interactions and disease processes [1]. Analyzing these molecules effectively is challenging due to their inherent complexity and large number of configurations. Vaccine development and nutrition involve carbohydrates, making their study of interest to researchers.

Representing carbohydrates digitally is important to researchers, enabling them to use computational tools to work with carbohydrates, allow for interoperability between analysis tools, and enable machine learning applications. Databases of carbohydrate structures are important for carbohydrate research, allowing for the sharing of carbohydrate information between researchers and easier access to large volumes of carbohydrate data – important in machine learning applications.

University of Cape Town, March, 2025, Cape Town, South Africa

Many machine learning (ML) pipelines benefit greatly from training on large volumes of high-quality data. Researchers use techniques such as data augmentation and synthetic data generation to improve model training. These techniques work by increasing the size of the dataset through artificially generating additional data.

There exist several carbohydrate representational formats, databases/datasets, and data augmentation methods. This literature review examines these aspects for applying machine learning to carbohydrates, identifying limitations and opportunities for future research.

2. CARBOHYDRATE STRUCTURE

To accurately model and represent carbohydrates, one must first understand their structure. Carbohydrates are structurally complex molecules consisting of only carbon (**C**), hydrogen (**H**), and oxygen (**O**) atoms [2], containing a chain of carbons, hydroxyl (-**OH**) groups, as well as an aldehyde or a ketone. An early chemical definition of carbohydrates classed them as molecules consisting of equal amounts of carbon and water, using the chemical formula $\mathbf{C_n(H_2O)_n}$, however, some carbohydrates do not adhere to this strict definition [2].

One important aspect of carbohydrate chemistry is that multiple distinct carbohydrates can emerge from the same chemical formula due to the intricacies of stereochemistry, which defines the arrangement of molecules in space.

“Saccharide” is a synonym for carbohydrate in biochemistry [3]. Carbohydrates are divided into four groups: monosaccharides, disaccharides, oligosaccharides, and polysaccharides. A residue

refers to a small or single unit of a carbohydrate such as a monosaccharide.

2.1 Saccharides

Monosaccharides

The fundamental building blocks of larger carbohydrates are monosaccharides, which can range from three to nine carbon atoms, with the simplest example being glyceraldehyde, and one of the largest being sialic acids [2]. $\mathbf{(CH_2O)_n}$ is the general formula for monosaccharides.

Stereochemistry

Stereochemistry concerns the arrangement of molecules in space. Multiple carbohydrates sharing the same chemical formula can exist, differing by their spatial molecular arrangement, and are known as isomers. Monosaccharides have two different versions depending on the orientation of the hydroxyl group on the chiral carbon farthest from the aldehyde or ketone (the chiral center or stereocenter). These are known as the **D**- and **L**- forms. Naturally occurring monosaccharides are in the **D**- form. Monosaccharides cannot interconvert between the **D**- and **L**- forms [4].

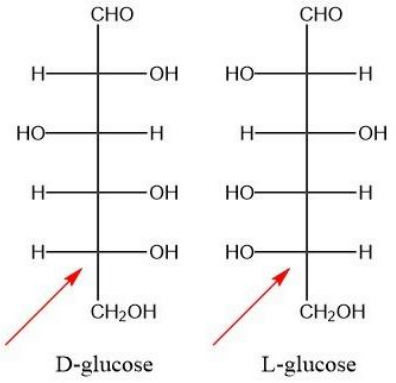


Figure 1: D-glucose and L-glucose isomers with the chiral carbon indicated by the red arrow [4].

This is a stereochemical aspect of carbohydrates. **D**-glucose and **L**-glucose are mirror images of one another. Even though the basic chemical formula for both glucose forms is the same ($\mathbf{C_6H_{12}O_6}$), the two resulting isomers differ in their spatial molecular arrangements.

Anomeric Configuration

Carbohydrates are usually present in the more common ring form as opposed to the chain form. 5-membered (furanose) or 6-membered (pyranose) rings are formed. The anomeric carbon is a carbon that 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 [4].

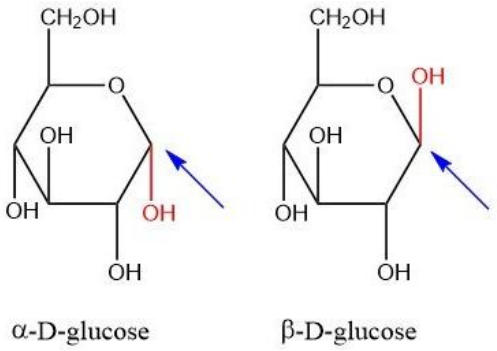


Figure 2: The alpha and beta forms of D-glucose, with the anomeric carbon indicated by the arrow, and the hydroxyl group in red [4].

Due to **D**-, **L**-, **α**-, and **β**- forms, monosaccharides have four total forms: **α**-**D**-, **α**-**L**-, **β**-**D**-, and **β**-**L**-. For example, a common sugar, glucose, can exist as **α**-**D**-glucose, **α**-**L**-glucose, **β**-**D**-glucose, and **β**-**L**-glucose.

Conformations

Carbohydrates can change their conformations – their spatial arrangement of atoms – without changing their bonds. In pyranose monosaccharides, the cyclic ring can take on one of many classes of conformations known as chair, skew, boat, half-chair, and envelope. Conformations are dynamic, differing from configurations, which are fixed unless a bond is changed. The conformation of a carbohydrate is largely affected by electronic and steric properties. For example, **α**-**D**-glucose can have multiple conformations, shifting between them depending on the environment, while remaining the same in absolute configuration. Conformations can affect recognition processes as well as the reactivity of the molecule – important for NMR chemical shift analysis [2].

Disaccharides

A disaccharide consists of two monosaccharides bonded together through a glycosidic linkage. A glycosidic linkage is a covalent bond between two monosaccharides [5].

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 3, galactose is in a **β**- anomeric configuration as the hydroxyl group on the anomeric carbon is pointing up.

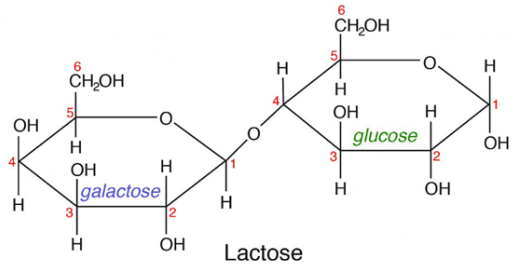


Figure 3: Lactose formed by galactose and glucose linked via a β(1→4) glycosidic bond [2].

If the **α-** anomeric configuration of galactose was used instead (hydroxyl group on the anomeric carbon point down), it would result in a different sugar, with linkage **α(1→4)**. This is true even though the chemical formula remains the same ($C_{12}H_{22}O_{11}$). Thus, stereochemistry plays a role in glycosidic bonding between saccharides.

Oligosaccharides and polysaccharides

A small number of monosaccharides, typically between 3 and 15, bonded together form an oligosaccharide. Larger numbers of saccharides form polysaccharides when bonded together. A polysaccharide can contain several thousand or more monosaccharides, forming complex carbohydrates.

Stereochemistry becomes a larger factor in polysaccharides. Since each monosaccharide unit can exist in four forms, the number of configurations increases dramatically as a polysaccharide increases in size.

3. CARBOHYDRATE REPRESENTATION

To perform computation on carbohydrates and share structures with other researchers, an appropriate standardized representational format is needed to make storage, searching, and computation easier. A suitable representation format is needed to encode the structure of carbohydrates for machine learning applications.

Several representational formats have emerged to encode carbohydrate structures. Some formats are more generalized and can work on a wider range of molecules, while others focus on carbohydrates specifically. Existing methods make tradeoffs between human readability, compactness, and depth of information encoded. Tools such as REStLESS [7] exist to convert between formats, for example to go from CSDB Linear to SMILES.

3.1 SMILES

SMILES (Simplified Molecular Input Line Entry System) is a linear chemical notation system designed for chemical information processing [6]. SMILES focuses on being compact to better utilize computer capacity [6]. It is used in many chemical databases and molecular editor software packages due to its simple syntax [8].

Molecular structures are encoded as strings. The SMILES string notation consists of a series of characters that ends with a space [6]. By following a defined set of rules, SMILES strings can be derived from a 2D molecular graph representation of the molecule. An important aspect of this is that no attempt is made to encode any 3D arrangement of the molecule's atoms [6]. The original 2D structure can be algorithmically reconstructed from the string representation by using the same rules [8]. These rules can be followed by a human, or by a computer when implemented as an algorithm in an interpreter program [9].

SMILES represents atoms with their atomic symbols. C is carbon, for example. Atoms with two-character names (Lead being Pb, for example) are represented with the first character capitalized and second character lowercase. Hydrogen atoms can be implicitly designated (hydrogen suppression) for normal valences or explicitly designated (to reduce ambiguity). Elements which are not in the organic subset, such as gold (Au), must be placed inside square brackets (e.g. [Au]) [6]. Single bonds between atoms are implicit when two atoms are next to each other in the string, with explicit bonds shown by placing "-" between the atoms. Double bonds use "=" and triple bonds use "#" [9]. Branched structures can be represented by enclosing a central atom in parenthesis. Cyclic structures can be encoded using digit pairs, where one bond is broken, and a pair of matching digits are placed immediately after each atom forming the bond [9].

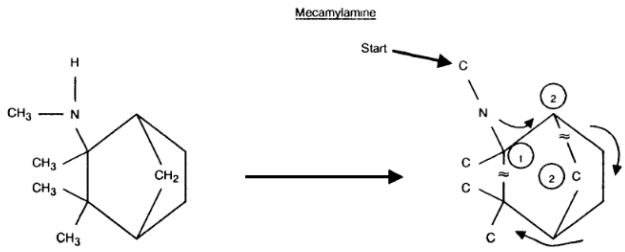


Figure 4: Example of branched cyclic structure SMILES string derivation [9].

The structure in Figure 4, Mecamylamine, has the hydrogen-suppressed SMILES string "**CNC1(C)C2CCC(C2)C1(C)C**".

Unfortunately, there are different versions of SMILES [10], and there are known differences in SMILES strings between software versions. For example, the original SMILES notation definition in [6] makes no provision for representing the stereochemistry of molecules. Various software packages began to extend SMILES with their own rules, including support for stereochemistry [11]. Chiral centers are commonly denoted by "**@**" or "**@@**". For

example, L-Alanine is **C[C@@H](N)C(=O)O**, and D-Alanine is **C[C@H](N)C(=O)O**.

SMILES struggles with ambiguity, where the same substance can have different labels, and vice-versa [11]. This is especially apparent in stereoisomers, such as the **D-** and **L-** forms of carbohydrates. Despite these challenges, SMILES is still widely used, including in machine learning applications such as GeqShift [1].

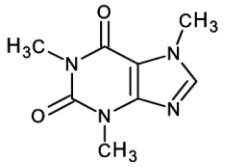
3.2 InChI

InChI (IUPAC International Chemical Identifier) is an IUPAC (International Union of Pure and Applied Chemistry) approved structure-derived tag for a chemical substance. It is based on a set of IUPAC structure conventions and rules for normalization and canonicalization [12]. InChI is non-proprietary and open-source [11].

An InChI string (or just "InChI") is produced from a graphical representation of the molecular structure using computer software. The original structure can also be regenerated from an InChI by using appropriate software [12]. InChIs are not designed to be readable by a human, instead, they act more like barcodes.

InChIs uniquely identify the structures from which they are derived [12]. Thus, the same structure will always produce the same InChI, and vice-versa. It is in this way that InChIs serve as unique chemical identifiers.

InChI uses a layered approach to represent structures, with each layer adding detail to the identifier. As a result of this, a structure drawn with a lower level of detail will be contained within the InChI of the same structure drawn with a higher level of detail [12]. InChI supports stereochemistry, encoding it inside one of its layers. For example, the InChI for a carbohydrate drawn without stereochemistry will be contained in the InChI for the same carbohydrate drawn with stereochemistry, which will have an additional stereochemical layer.



InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3 (caffeine)

Figure 5: Example of InChI for caffeine [11]. Layers are separated by a "/" followed by a lowercase character [12].

InChI's uniqueness provides unambiguity for InChIs, but unfortunately InChIs do not have sufficient support for representing ambiguities in structures [13]. More than one structure can generate the same InChI, and different InChIs could potentially be generated from the same structure due to differences in the chemist's drawing of the structure. Additionally, generating the graphical structure from an InChI is not 100% reliable, although it works more than 99% of the time [12].

3.3 CASPER

CASPER [14] is a web browser accessible tool designed to aid the classification and representation of carbohydrates, which uses its own notation. CASPER uses an incremental rule-based approach to predict 1H and 13C NMR chemical shifts of glycans [14, 15].

CASPER's notation for describing carbohydrate structures encodes anomeric configuration (**α-/β-**), stereochemical or absolute configuration (**D-/L-**), residue identity (via a 3-letter abbreviation), and ring type (f for furanose, default pyranose). For example, **α-D-Glc** represents **α-D-glucopyranose**, and **β-L-Galf** represents **β-L-galactofuranose**. Glycosidic linkages are also explicitly defined: **α-D-Gal(1→4)α-D-GlcOMe** represents a **Gal-Glc** disaccharide with a **α(1→4)** linkage [17]. Branches can be represented by square brackets, and multiple branches are sorted according to substitution position. Repeating units are represented by leaving the linkage open (e.g. "**→4)aDGlc(1→6)aDGal(1→"**). Substituents, groups of atoms replacing one or more atoms, are added in alphabetical order, with substitution positions and multiplicity specified [17].

The CASPER notation covers carbohydrate residues, it is not atomic. This means that it is specific to carbohydrates and cannot represent arbitrary molecular structures.

CASPER notation is supported by the CarbBuilder software. CarbBuilder is a tool to generate 3D structures of carbohydrates in the Protein Data Bank (PDB) format from CASPER notation strings. CarbBuilder supports a pre-defined set of monosaccharides and substituents, and thus cannot generate structures of other arbitrary carbohydrates [18].

CASPER has been used in machine learning in GeqShift [1], where the CASPER notation was converted into SMILES format through the use of conversion software.

3.4 GlycoCT

GlycoCT is a structured encoding format using a connection table (CT) approach. It was developed as part of the EUROCarbDB project to address the limitations of existing carbohydrate sequence formats having differing capabilities to encode carbohydrate structures [19].

GlycoCT comes in 2 variants, a condensed form consisting of a unique compact linear string, as well as a more verbose XML format for machine use. The condensed form can be used as a unique identifier, similar to InChI, for carbohydrate structures, including cases where ambiguities are present in the structures [19]. The XML format is more suited towards machine use and data exchange. Both variants use the same definitions and sequences can be easily transformed between them [19].

Monosaccharides are defined using IUPAC conventions, preventing the restriction of GlycoCT to a limited set of monosaccharides. Uncertain linkages, ambiguous monosaccharides, and repeating units are supported by GlycoCT [19]. Like CASPER, GlycoCT is focused on carbohydrates and therefore cannot represent arbitrary molecular structures. GlycoCT is only complete in the scope of carbohydrate residues [13].

A unique feature of GlycoCT is its use of strict sorting rules that ensure a unique encoding for each glycan sequence, which enables its use as a primary key in databases and allows for the search of exact structures. These sorting rules determine the order of appearance of different elements in the sequence, resulting in a unique string representation for each sequence [19]. Sorting works through a hierarchical set of rules which order structures in an unambiguous manner [19]. Due to these sorting rules, a particular carbohydrate will only have one representation in GlycoCT. Researchers extracted and translated CarbBank's monosaccharide namespace to GlycoCT, which resulted in a 65% reduction in the number of distinct residues [19].

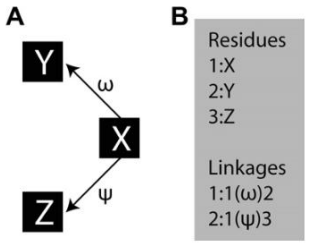


Figure 6: Connection table in GlycoCT, with (B) containing the residue and linkage lists. [19].

GlycoCT uses a connection table (CT) based approach instead of a linear encoding scheme. A trisaccharide is depicted in Figure 6 (A), where X, Y, and Z represent residues (monosaccharide units, for example). Linkages between the residues are indicated with ω and ψ. (B) shows the connection table containing a residue and linkage list. The residue list encodes the residues X, Y, and Z, while the linkage list encodes the connectivity between residues [19].

For example, a condensed representation of a linear trisaccharide:
RES
1b:x-dglc-HEX-1:5
2b:a-dman-HEX-1:5
3b:a-dgal-HEX-1:5
LIN
1:1o(2+1)2d
2:2o(2+1)3d

GlycanBuilder exists to build GlycoCT descriptions for carbohydrates without requiring the user to learn how to directly write the GlycoCT format manually [19].

GlycoCT is used in several other projects, including the GlyTouCan repository [20] and REStLESS conversion software [7].

3.5 WURCS

The Web3 Unique Representation of Carbohydrate Structures (WURCS) is a linear notation developed specifically to support the representation of carbohydrates in the Semantic Web [21].

WURCS aims to address the limitations of previous formats such as GlycoCT. The researchers noted that while GlycoCT is quite close to satisfying their requirements, it is non-linear and would be difficult to represent certain monosaccharide structures in GlycoCT format [21].

WURCS encodes glycan structures in a string, making it suitable for use as a Uniform Resource Identifier (URI) in databases and the Semantic Web. Like GlycoCT, WURCS ensures a unique identifier for glycans so that any structure is distinctly notated [21]. The WURCS string consists of a sorted list of BMUs (monosaccharides) followed by a sorted list of MLUs (linkages). Uncertain glycosidic linkages and repeating units are supported by WURCS, like GlycoCT, as well as cyclic structures. Ambiguous glycosidic bonds are supported in WURCS, where backslashes indicate possible attachment sites [21].

An example of a chemical structure that cannot be represented by a unique linear string in any other major carbohydrate format is the structure in Figure 7 with PDB ID 1A78. It has the WURCS string
"**WURCS=1.0/2,1/[12112h|1,5][12112h|1,5]1+1:1,2+1:1*S***".

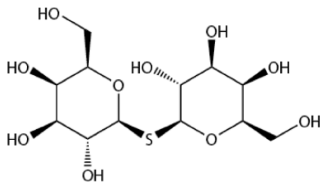


Figure 7: Chemical structure that cannot be represented in any other major carbohydrate format. [21].

A limitation of WURCS is that it does not consider ambiguous monosaccharide structures. WURCS 2.0 was developed to address this limitation, adding anomeric information and an updated format to handle ambiguous monosaccharides as well as give a shorter and simpler representation for carbohydrate structures [22]. Software tools exist to validate and convert between WURCS and other formats and chemical structures [22]. WURCS 2.0 is used in the GlyTouCan repository [20].

3.6 CSDB Linear

The CSDB (Carbohydrate Structure Database) Linear representation is an unambiguous, human- and machine-readable linear string representation for carbohydrates [13].

CSDB Linear encodes molecules as directed graphs, with the vertices being residues and the edges being linkages between residues. CSDB Linear supports stereochemistry, ambiguity, and other features that alternative notations such as WURCS already support. A unique feature of the notation is that it can encode atypical residues – ones that are missing from monomer vocabulary by using SMILES notation inside the string through a substitution operation [13].

Example of a CSDB Linear string containing an atypical residue: "**aDRibf(1−3)Subst // Subst = questin = SMILES COC1=C{3}C(O)=CC(C(C2=C3{8}C(O)=CC(C)=C2)=O)=C1C3=O"**.

CSDB Linear can be converted between different formats including SMILES, GlycoCT, and WURCS through the CSDB interface. A validation tool is also available at the CSDB website [23]. CSDB Linear is used in the CSDB database.

Format	Approach	Universal	Unique	Stereochemistry	Ambiguity
SMILES	Atomic	Yes	No	Yes (some versions)	No
InChI	Atomic	Yes	Yes	Yes	No
CASPER	Residue string	No (carb-only)	No	Yes	No
GlycoCT	Connectivity table	No (carb-only)	Yes	Yes	Partial
WURCS 2.0	Connectivity Table	No (carb-only)	Yes	Yes	Yes
CSDB Linear	Graph	No (carb-only)	Yes	Yes	Yes

Table 1: Summary of representational formats

4. CARBOHYDRATE DATASETS

A large dataset containing high-quality carbohydrate structure and corresponding NMR spectra is important for training machine learning models on carbohydrate NMR prediction. There currently exist several databases with NMR spectra, each varying in size and completeness. Each database additionally encodes data in different formats, another factor to consider.

4.1 CSDB (Carbohydrate Structure Database)

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 [23].

Data export to various formats and encoding schemes is available on request, and web access is free. Structure data is annotation with assigned NMR spectra and other information where available. CSDB makes use of the CSDB Linear notation [13].

CSDB source data files can be obtained and imported into a different database system, such as SQL [24]. This flexibility is helpful for machine learning applications.

Example of a data record from CSDB, featuring NMR spectra, with information irrelevant to the project removed:

ID: 4676
ST1: -6)[Ac(1-2)]aDGlcpN(1-3)[bDGlcp(1-4),Ac(1-2)]aLFucpN(1-3)[xXEtN(1-P-6),Ac(1-2)]bDGlcpN(1-2)bDGlcp(1-
NMRH: #2,3,3,2_Ac // #2,3,3_aDGlcpN 5.05 3.98 3.73 3.72 3.95 4.06 // #2,3,4_bDGlcp 4.63 3.45 3.53 3.45 3.42 3.76-3.97 // #2,3,2_Ac // #2,3_aLFucpN 5.00 4.46 4.10 4.08 4.47 1.31 // #2,6,0_xXEtN 4.17 3.32 // #2,6_P // #2,2_Ac // #2_bDGlcpN 4.91 3.91 3.77 3.61 3.62 4.13-4.25 // #_bDGlcp 4.55 3.55 3.55 3.42 3.44 3.73-3.91 //
NMRC: #2,3,3,2_Ac // #2,3,3_aDGlcpN 99.0 54.5 72.4 70.5 72.2 69.4 // #2,3,4_bDGlcp 104.2 75.0 77.5 70.9 76.8 61.9 // #2,3,2_Ac // #2,3_aLFucpN 98.6 50.4 71.4 79.1 68.9 16.7 // #2,6,0_xXEtN 63.2 41.3 // #2,6_P // #2,2_Ac // #2_bDGlcpN 102.2 57.0 78.9 69.5 75.9 66.0 // #_bDGlcp 103.0 81.2 77.4 70.8 76.8 61.6 //

The large amount of carbohydrate structure and NMR shift data in CSDB is of interest.

4.2 GlycoNMR

The research paper [25] states that this is the first large-scale curated carbohydrate-specific NMR dataset, aiming to address the scarcity of appropriate carbohydrate data for machine learning. It includes GlycoNMR.Exp (containing experimental ^{13}C and ^{1}H NMR shifts) and GlycoNMR.Sim (containing simulated shifts using GODESS). It contains 2,609 carbohydrate structures and 211,543 annotated NMR chemical shifts. Researchers note that a lack of raw spectra is a limiting factor, as experimentalists typically only report the peak positions of chemical shifts [25].

Each data record in the GlycoNMR dataset contains the NMR chemical shifts for the carbohydrate structure in CSV format, using the CSDB notation for carbohydrate representation. This CSV format can be easily used in machine learning pipelines.

GlycoNMR's training readiness and large number of NMR shifts are noteworthy. The addition of simulated shifts expands the dataset, although it may introduce inaccuracies as the simulated shifts are not real experiments.

4.3 NMRShiftDB2

NMRShiftDB2 [26] is an open-source database for organic molecule NMR shifts, which includes carbohydrates. It contains both experimentally recorded as well as predicted NMR chemical shifts. The core of the databases features fully assigned spectra with raw data as well as peak lists, which are peer reviewed by a board of reviewers [26].

NMRShiftDB2 contains 271,668 structures, 68,467 measured NMR spectra, and 396,583 calculated NMR spectra as of March 2025 [26]. The structures are encoded in multiple formats, including SMILES and InChI.

Although this database does not focus on carbohydrates specifically, the large amount of data may be useful in training, generalizing the model.

4.4 CCMRD (Complex Carbohydrate Magnetic Resonance Database)

The CCMRD contains more than 400 solid-state NMR spectra of carbohydrate molecules [13]. It is freely accessible via the web. CCMRD only accepts high-resolution data from peer-reviewed publications, checking each record for reliability [28]. The researchers state that their efforts can ultimately lead to spectral analysis using deep-learning neural networks.

Example of a data record from CCMRD, with the residue followed by the solid-state NMR chemical shifts:
a-?-GlcpNAc
C1:104.0;C2:54.8;C3:73.4;C4:82.9;C5:75.6;C6:60.6;C7:173.0;C8:22.6;H1:4.6;H2:3.3;H3:3.1;H4:3.0;H5:3.0;H6:3.1;H7:;H8:1.1

This database is small, but contains solid-state NMR spectra, a unique kind of NMR that works for insoluble complex carbohydrates [28].

4.5 GeqShift Dataset [1]

The dataset used for GeqShift [1] is based on published data used by CASPER. It contains 375 carbohydrate structures with their accompanying NMR chemical shifts, giving 5,356 1H and 4,713 13C shifts. The dataset is small, containing just 375 carbohydrates.

Dataset	Carbohydrate-specific	Structures	Spectra	Approx. spectra coverage
CSDB	Yes	32,937	19,728	60%
GlycoNMR	Yes	2609	211,543 (shifts)	Unknown
NMRShiftDB2	No	271,668	68,467 (measured)	~25%
CCMRD	Yes	400+	400+	100%
GeqShift Dataset	Yes	375	375	100%

Table 2: Comparison of datasets.

5. 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 combined 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. Data augmentation is especially important when existing datasets are small, improving model robustness [30].

5.1 Augmentation Strategies

Conformation Sampling

Molecules can have many conformations. A single confirmation may not fully model factors that affect NMR chemical shifts. GeqShift [1] employs conformation sampling using RDKit. Training on multiple conformations (100 per molecule) reduced the mean absolute error (MAE) from 0.55 to 0.34 for ^{13}C chemical shifts. This method also improves the model's generalization to previously unseen structures.

Simulated NMR Spectra

Tools exist to simulate or predict NMR spectra, including GODDESS from CSDB [23], CASPER [14], and GeqShift [1]. The GlycoNMR dataset makes use of simulated NMR spectra through GODDESS to augment the dataset. This results in over 200,000 NMR chemical shifts available for training. The accuracy of this method may vary depending on the accuracy of the simulated or predicted shifts.

5.2 Synthetic Data Generation

Synthetic data is data that has been artificially generated to expand the dataset. Synthetic data generation can improve model robustness by exposing it to a wider range of scenarios [31]. Synthetic data has been used extensively, especially when real data is lacking. It has enormous potential for improving machine learning [31].

Synthetic data generation can be done using both standard methods (those not involving machine learning) and deep learning methods. Deep learning methods can better learn the underlying patterns in the data compared to standard methods and therefore provide higher-quality synthetic data in most cases [32]. Synthetic generation of carbohydrate structures, followed by simulating their NMR spectra, or using the existing GeqShift [1] model to predict their spectra may prove to be useful.

6. CONCLUSIONS

Carbohydrate molecules have complex structures that are difficult to accurately represent. This poses a challenge for machine learning involving carbohydrates, especially NMR chemical shift prediction. Some molecule representational formats fail to capture the intricacies of carbohydrate molecules. Although carbohydrate-specific formats like GlycoCT, WURCS 2.0, and CSDB Linear notation, along with conversion tools [32], have been developed to overcome these challenges, tradeoffs exist, and representational format usage remains fragmented between datasets.

ML model performance depends on the availability and quality of datasets containing carbohydrate NMR shifts. There appears to be a lack of large high-quality datasets suitable for training machine learning models on for NMR chemical shift prediction.

Data augmentation has been proven to be useful. Expanding an existing dataset using data augmentation methods may provide substantial improvements in the performance of machine learning methods as ML models benefit greatly from training on large high-quality datasets. Data augmentation methods have been successfully used for carbohydrate data, showing promising results. Synthetic data generation has high potential for improving machine learning model robustness, and its application to carbohydrate data may be worth further exploration.

ACKNOWLEDGMENTS

This work is done during an Honours study at the University of Cape Town.

REFERENCES

[1] Bånkestad, Maria and Dorst, Kevin M. and Widmalm, Göran and Rönnols, Jerk,. 2024. Carbohydrate NMR chemical shift prediction by GeqShift employing E(3) equivariant graph neural networks. RSC Adv. 14, 36 (2024), 26585-26595.

[2] Jerk Rönnols. 2013. Structure, dynamics and reactivity of carbohydrates. Thesis. Stockholm University, Stockholm, Sweden.

[3] Pierre Avenas. 2012. Etymology of Main Polysaccharide Names. The European Polysaccharide Network of Excellence (Nov. 2012), 13-21.

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

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

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

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

[8] Jaroslaw Polanski, Johann Gasteiger. 2015. Computer Representation of Chemical Compounds. Springer Science, Handbook of Computational Chemistry (Jan. 2017), 1997-2039.

[9] Eric Anderson, Gilman D. Veith, and David Weininger. 1987. SMILES: A Line Notation and Computerized Interpreter for Chemical Structures. US EPA Environmental Research Laboratory EPA/600/M-87/021 (Aug. 1987), 4 pages.

[10] Steven M Bachrach. 2012. InChI: a user's perspective. Journal of Chemoinformatics 4:34 (2012), 3 pages.

[11] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. 2015. InChI, the IUPAC International Chemical Identifier. Journal of Cheminformatics 7:23 (2015), 34 pages.

[12] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. 2013. InChI – the worldwide chemical structure identifier standard. Journal of Cheminformatics 5:7 (2013), 9 pages.

[13] Philip V. Toukach and Ksenia S. Egorova. 2020. New Features of Carbohydrate Structure Database Notation (CSDB Linear), As Compared to Other Carbohydrate Notations. 2020. J. Chem. Inf. Model. 60 (2020), 1276-1289.

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

[15] Per-Erik Jansson, Lennart Kenne, Göran Widmalm. 1991. CASPER: A Computer Program Used for Structural Analysis of Carbohydrates. J. Chem. Inf. Comput. Sci., 31 (1991), 508-516.

[16] Alexander Loß, Roland Stenutz, Eberhard Schwarzer, Claus-W. von der Lieth. 2006. GlyNest and CASPER: two independent approaches to estimate 1H and 13C NMR shifts of glycans available through a common web-interface. Nucleic Acids Research, 34 (2006), 733-737.

[17] The Widmalm Research Group, Stockholm University. 2012. CASPER manual. (January 2012). Retrieved March 23, 2025 from http://www.casper.organ.su.se/casper/manual.pdf

[18] Michelle M. Kuttel, Jonas Ståhle, and Göran Widmalm. 2016. CarbBuilder: Software for Building Molecular Models of Complex Oligo- and Polysaccharide Structures. Journal of Computational Chemistry 37 (2016), 2098-2105.

[19] S. Herget, R. Ranzinger, K. Maass, C.-W v. d. Lieth. 2008. GlycoCT – a unifying sequence format for carbohydrates. Elsevier, Carbohydrate Research 343 (2008), 2162-2171.

[20] Michael Tiemeyer, Kazuhiro Aoki, James Paulson et. al. 2017. GlyTouCan: an accessible glycan structure repository. Glycobiology 27, 10 (2017), 915-919.

[21] Kenichi Tanaka, Kiyoko F. Aoki-Kinoshita, Masaaki Kotera, Hiromichi Sawaki, Shinichiro Tsuchiya, Noriaki Fujita, Toshihide Shikanai, Masaki Kato, Shin Kawano, Issaku Yamada, and Hisashi Narimatsu. 2014. WURCS: The Web3 Unique Representation of Carbohydrate Structures. J. Chem. Inf. Model., 54 (Jun. 2014), 1558-1566.

[22] Masaaki Matsubara, Kiyoko F. Aoki-Kinoshita, Nobuyuki P. Aoki, Issaku Yamada, and Hisashi Narimatsu. 2017. WURCS 2.0 Update To Encapsulate Ambiguous Carbohydrate Structures. J. Chem. Inf. Model., 57 (2017), 632-637.

[23] Philip V. Toukach, Ksenia S. Egorova, Yuri A. Knirel, et al. 2024. Carbohydrate Structure Database (CSDB). (July 2023). Retrieved March 23, 2025 from http://csdb.glycoscience.ru

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

[25] Zizhang Chen, Ryan Paul Badman, Lachele Foley, Robert Woods, Pengyu Hong. 2023. GlycoNMR: DATASET AND BENCHMARKS FOR NMR CHEMICAL SHIFT PREDICTION OF CARBOHYDRATES WITH GRAPH NEURAL NETWORKS. arXiv, cs.LG (2023), 2311.17134v2.

[26] Stefan Kuhn. 2024. NMRShiftDB2 – Open NMR Database on the Web. (August 2024). Retrieved March 23, 2025 from https://nmrshiftdb.nmr.uni-koeln.de/

[27] Xue, Alex , Malitha, Uluc, Tuo. 2025. Complex Carbohydrate Magnetic Resonance Database (CCMRD). (March 2025). Retrieved March 23, 2025 from http://www.ccmrd.org/

[28] Kang X, Zhao W, Dickwella Widanage MC, Kirui A, Ozdenvar U, Wang T. 2020. CCMRD: a solid-state NMR database for complex carbohydrates. J Biomol NMR 74 (May 2020), 239-245.

[29] Alhassan Mumuni and Fuseini Mumuni. 2022. Data augmentation: A comprehensive survey of modern approaches. Array 16, 100258 (2022), 27 pages.

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

[31] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, Wenqi Wei. 2024. Machine Learning for Synthetic Data Generation: A Review. arXiv, cs.LG (2024), 2302.04062v9.

[32] Alvaro Figueira and Bruno Vaz. 2022. Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics 10, 2733 (2022), 41 pages.