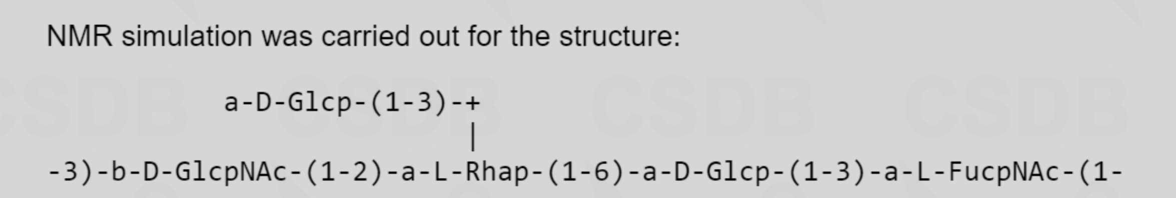
# *Annotation example of a molecule that is present in both Glycosciences and GODESS datasets*

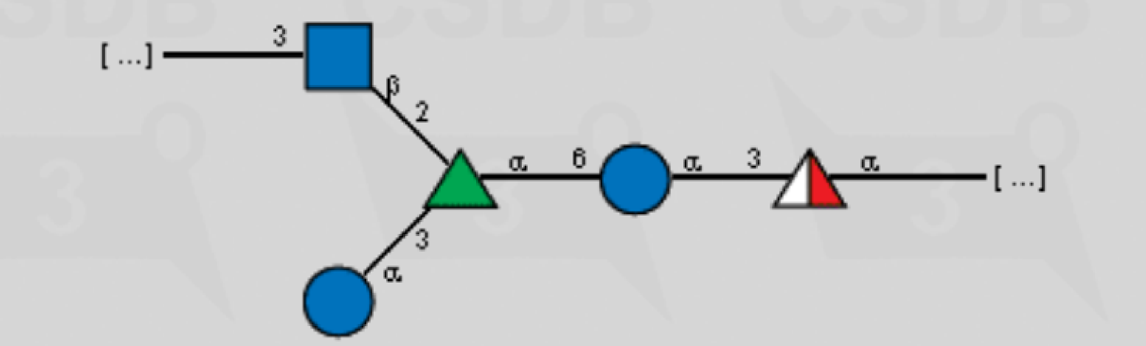
## Section 1: Carbohydrate Summary and Overview of Mismatch Correction

For this ML pipeline annotation example we will discuss the following molecule (screenshots with a gray background extracted from http://csdb.glycoscience.ru/ ) using the input notation: *"-3)[Ac(1-2)]bDGlcpN(1-2)[aDGlcp(1-3)]aLRhap(1-6)aDGlcp(1-3)[Ac(1-2)]aLFucpN(1-"* when using GODESS. This molecule is a helpful illustration of our annotation process as it shows several common issues in experimental data that led us to make our carbohydrate-specialized annotation pipeline for ML models.



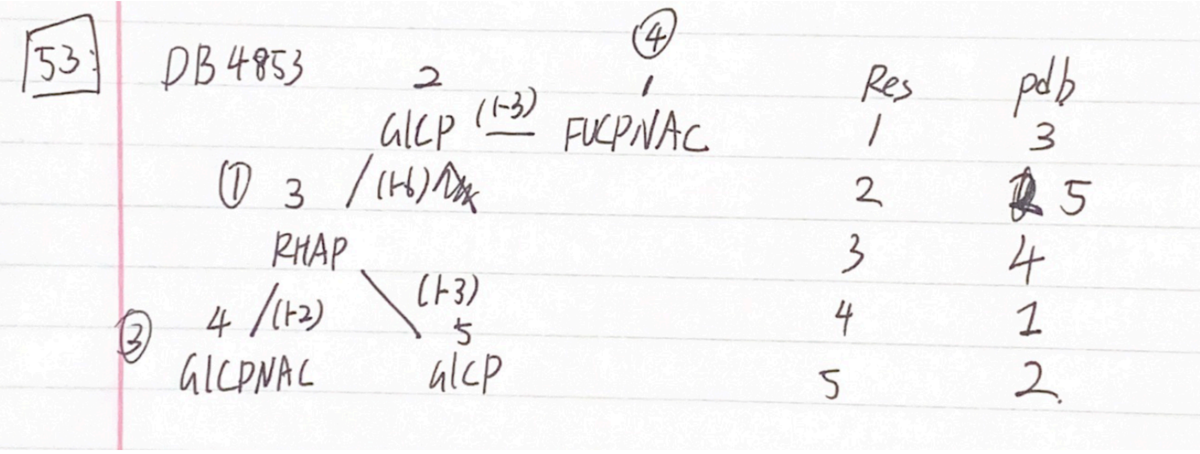
NMR simulation was carried out for the structure:

```
                a-D-Glcp-(1-3)-+
                               |
-3)-b-D-GlcpNAc-(1-2)-a-L-Rhap-(1-6)-a-D-Glcp-(1-3)-a-L-FucpNAc-(1-
```



▲ **a-L-Rhap;** ● **a-D-Glcp;** ◭ **a-L-FucpNAc;** ■ **a-D-GlcpNAc**

As can be seen from the handwritten notes pdf on the github (example extracted below this paragraph) there are ordering mismatches between the residue list in the NMR file ("Res") and the structure file ("pdb") as seen in the columns on the right in the figure below:



53
DB 4853
④
2
1
GlCP
(1-3)
FUCPNAC
① 3
/(1-6)
RHAP
③ 4
/(1-2)
(1-3)
5
GlCPNAC
GlCP

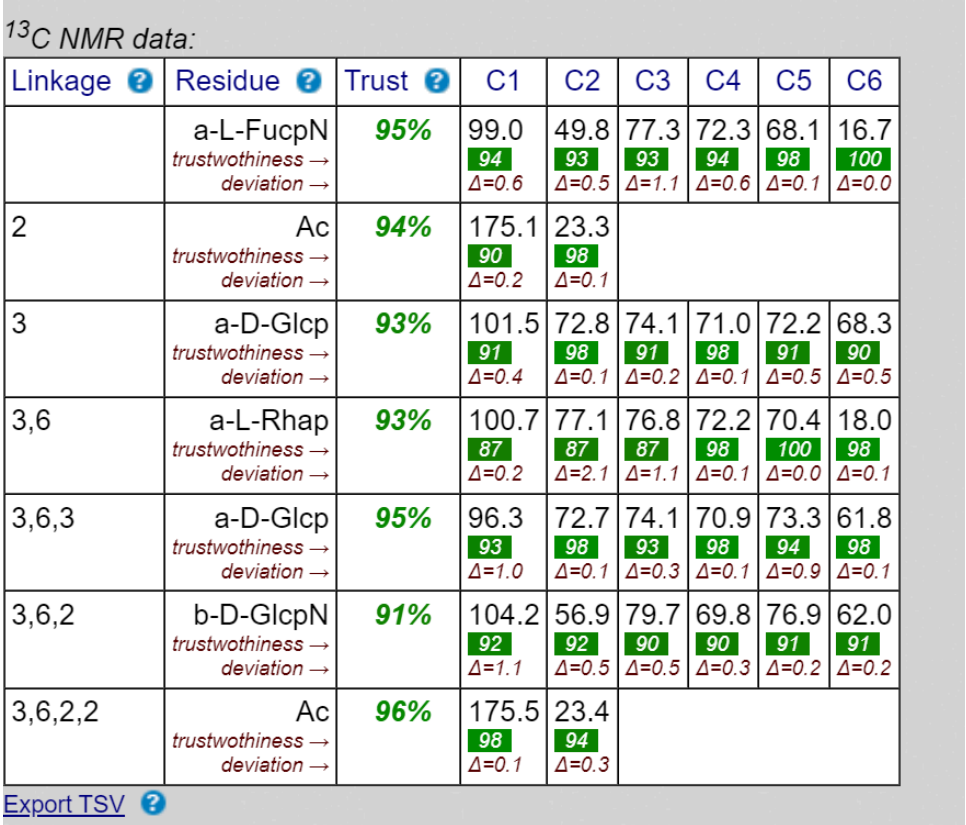
| Res | pdb |
|---|---|
| 1 | 3 |
| 2 | ~~2~~ 5 |
| 3 | 4 |
| 4 | 1 |
| 5 | 2 |

This ordering mismatch causes several issues:

(1) The baseline stem and linkages will be mismatched. For example, stem (residue) #1 in the PDB file is FUC (a-L-FucpNAc). However, the first residue in the NMR file is RAM (a-L-Rhap). Thus a 1-1 mapping simply based on default ordering within files will not work.

(2) “FUC” as a PDB label sometimes means Fucp as a residue name in these files, but here it means FucpNAc and in other files “FUC” labels other Fucp-based stems. Thus a global lookup table cannot be used for all carbohydrate entries due to labeling ambiguity in the PDB notation.

(3) There are two GLC labels in the PDB file, and further both are a-D-Glcp with (1-3) linkages. Thus residue and linkage labels alone cannot be used for matching the residues between the NMR and PDB files. As can be seen from the theoretical prediction in GODESS (and also in the experimental data below, **Section 3**), the different positions of these GLC residues leads to different chemical shifts and they are not interchangeable:



$^{13}C$ NMR data:

| Linkage | Residue | Trust | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|---|---|
| | a-L-FucpN<br>trustwothiness →<br>deviation → | 95% | 99.0<br>94<br>Δ=0.6 | 49.8<br>93<br>Δ=0.5 | 77.3<br>93<br>Δ=1.1 | 72.3<br>94<br>Δ=0.6 | 68.1<br>98<br>Δ=0.1 | 16.7<br>100<br>Δ=0.0 |
| 2 | Ac<br>trustwothiness →<br>deviation → | 94% | 175.1<br>90<br>Δ=0.2 | 23.3<br>98<br>Δ=0.1 | | | | |
| 3 | a-D-Glcp<br>trustwothiness →<br>deviation → | 93% | 101.5<br>91<br>Δ=0.4 | 72.8<br>98<br>Δ=0.1 | 74.1<br>91<br>Δ=0.2 | 71.0<br>98<br>Δ=0.1 | 72.2<br>91<br>Δ=0.5 | 68.3<br>90<br>Δ=0.5 |
| 3,6 | a-L-Rhap<br>trustwothiness →<br>deviation → | 93% | 100.7<br>87<br>Δ=0.2 | 77.1<br>87<br>Δ=2.1 | 76.8<br>87<br>Δ=1.1 | 72.2<br>98<br>Δ=0.1 | 70.4<br>100<br>Δ=0.0 | 18.0<br>98<br>Δ=0.1 |
| 3,6,3 | a-D-Glcp<br>trustwothiness →<br>deviation → | 95% | 96.3<br>93<br>Δ=1.0 | 72.7<br>98<br>Δ=0.1 | 74.1<br>93<br>Δ=0.3 | 70.9<br>98<br>Δ=0.1 | 73.3<br>94<br>Δ=0.9 | 61.8<br>98<br>Δ=0.1 |
| 3,6,2 | b-D-GlcpN<br>trustwothiness →<br>deviation → | 91% | 104.2<br>92<br>Δ=1.1 | 56.9<br>92<br>Δ=0.5 | 79.7<br>90<br>Δ=0.5 | 69.8<br>90<br>Δ=0.3 | 76.9<br>91<br>Δ=0.2 | 62.0<br>91<br>Δ=0.2 |
| 3,6,2,2 | Ac<br>trustwothiness →<br>deviation → | 96% | 175.5<br>98<br>Δ=0.1 | 23.4<br>94<br>Δ=0.3 | | | | |

Export TSV

## Section 2: Our annotation solution

In this specific example, the PDB file ordering is more correct. Carbohydrate residue order is usually read right to left, though with ambiguity in which branches to proceed through first. Here we will match both files to the #1-5 ordering: **a-L-FucpNAc (FUC), a-D-Glcp {1} (GLC), a-L-Rhap (RAM), a-D-GlcpNAc (NAG), a-D-Glcp {2} (GLC).** As there is only one FUC, NAG, and RAM, it is straightforward to reorder and match these across the PDB and NMR file.

For the GLC ambiguity, we will use the button SWECON rows in the PDB file (see **Section 3)**, and the linkage column in the NMR shift file (see **Section 4**):

*PDB*
SWECON 1   2  1  A-D-GLCP-(1-3)-A-L-FUCPNAC **-> This is GLC {1}, by inspection**
SWECON 2   3  2  A-L-RHAP-(1-6)-A-D-GLCP
SWECON 3   4  3  B-D-GLCPNAC-(1-2)-A-L-RHAP
SWECON 4   5  3  A-D-GLCP-(1-3)-A-L-RHAP **-> This is GLC {2}, by inspection**

*NMR*
a-D-Glcp          3,3,2,3 **-> This is GLC {2}, based on inspection and molecules deeper in the chain have more linkages listed**
a-D-Glcp          3,3,until **-> This is GLC {1}**

## Section 3: PDB File

***(abbreviated to focus on the ordering)***

*Note:* "HETATM    1  C1  FUC     1…" This red number is the residue ordering number in the file below

```
HEADER    CARBOHYDRATE
COMPND    UNNAMED
AUTHOR    CREATED BY SWEET-II ON WWW.GLYCOSCIENCES.DE
LINK       O3  FUC    1                  C1  GLC    2
LINK       O6  GLC    2                  C1  RAM    3
LINK       O2  RAM    3                  C1  NAG    4
LINK       O3  RAM    3                  C1  GLC    5
HETATM    1  C1  FUC     1      8.030   7.339  -6.263  1.00  0.00          C
HETATM    2  C2  FUC     1      9.375   7.534  -6.988  1.00  0.00          C
…
HETATM   29  C1  GLC     2     12.353   5.919  -7.557  1.00  0.00          C
HETATM   30  C2  GLC     2     13.646   6.357  -8.238  1.00  0.00          C
…
HETATM   50  C1  RAM     3     10.116   0.895  -8.913  1.00  0.00          C
HETATM   51  C2  RAM     3      8.686   0.770  -9.500  1.00  0.00          C
…
HETATM   69  C1  NAG     4      6.633   0.105  -8.394  1.00  0.00          C
HETATM   70  C2  NAG     4      5.800  -1.188  -8.120  1.00  0.00          C
…
HETATM   97  C1  GLC     5      6.614   1.104 -11.994  1.00  0.00          C
HETATM   98  C2  GLC     5      5.113   0.839 -12.051  1.00  0.00          C
…
CHLDEF    1    46   29    9    3   15    0 -51.0  -3.9 -165.8
CHLDEF    2    60   50   39   34   33   43  51.4 -160.0  -2.0
CHLDEF    3    96   69   56   51   61    0  37.0  15.0  14.1
CHLDEF    4   115   97   57   52   62    0 -42.0 -25.9  79.4
CHLNAM    1 A-D-GLCP-(1-3)-A-L-FUCPNAC
CHLNAM    2 A-L-RHAP-(1-6)-A-D-GLCP
CHLNAM    3 B-D-GLCPNAC-(1-2)-A-L-RHAP
CHLNAM    4 A-D-GLCP-(1-3)-A-L-RHAP
SWECON 1   2  1  A-D-GLCP-(1-3)-A-L-FUCPNAC
SWECON 2   3  2  A-L-RHAP-(1-6)-A-D-GLCP
SWECON 3   4  3  B-D-GLCPNAC-(1-2)-A-L-RHAP
SWECON 4   5  3  A-D-GLCP-(1-3)-A-L-RHAP
MASTER       0    0    0    0    0    0    0    0  118    0  118    0
END
```

**Section 4: NMR File**

***(H shifts are first, then the residues are re-listed for C shift in the same order)***

```
MHz    300
Temperature   353
Solvent       D2O
Residue       Linkage Proton
PPM    JFrom  JTo     Hz
a-L-Rhap       3,until  H1      4.84    1         2          2
a-L-Rhap       3,until  H2      4.16    2         3          3.5
a-L-Rhap       3,until  H3      3.82    3         4          9
a-L-Rhap       3,until  H4      3.27    4         5          9
a-L-Rhap       3,until  H5      3.65    5         6          6
a-L-Rhap       3,until  CH3     1.22                         0
a-D-Glcp       3,3,until         H1     5.05    1          2          3.5
a-D-Glcp       3,3,until         H2     3.63    2          3          9.5
a-D-Glcp       3,3,until         H3     3.74    3          4          9.5
a-D-Glcp       3,3,until         H4     3.42    4          5          9.5
a-D-Glcp       3,3,until         H5     3.95    5          6          2.5
a-D-Glcp       3,3,until         H61    3.76    6           6'        12.5
a-D-Glcp       3,3,until         H62    3.69    5           6'        4.5
b-D-GlcpNAc   2,3,until         H1     4.71    1          2          8
b-D-GlcpNAc   2,3,until         H2     3.84    2          3          9
b-D-GlcpNAc   2,3,until         H3     3.51    3          4          9
b-D-GlcpNAc   2,3,until         H4     3.45    4          5          9
b-D-GlcpNAc   2,3,until         H5     3.37                          0
b-D-GlcpNAc   2,3,until         H61    3.88                          0
b-D-GlcpNAc   2,3,until         H62    3.68                          0
a-L-FucpNAc   3,2,3,until       H1     4.97    1          2          3.5
a-L-FucpNAc   3,2,3,until       H2     4.29    2          3          10
```

| | | | | | | |
|---|---|---|---|---|---|---|
| a-L-FucpNAc | 3,2,3,until | H3 | 3.86 | 3 | 4 | 4 |
| a-L-FucpNAc | 3,2,3,until | H4 | 3.81 | 4 | 5 | 2 |
| a-L-FucpNAc | 3,2,3,until | H5 | 4.34 | 5 | 6 | 6.5 |
| a-L-FucpNAc | 3,2,3,until | CH3 | 1.15 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | H1 | 4.97 | 1 | 2 | 3.5 |
| a-D-Glcp | 3,3,2,3,until | H2 | 3.43 | 2 | 3 | 9.5 |
| a-D-Glcp | 3,3,2,3,until | H3 | 3.66 | 3 | 4 | 9.5 |
| a-D-Glcp | 3,3,2,3,until | H4 | 3.37 | 4 | 5 | 9.5 |
| a-D-Glcp | 3,3,2,3,until | H5 | 3.83 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | H61 | 3.83 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | H62 | 3.65 | | | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| a-L-Rhap | 3,until C1 | 100.5 | C1 | H1 | 174 | |
| a-L-Rhap | 3,until C2 | 75.1 | | | 0 | |
| a-L-Rhap | 3,until C3 | 75.6 | | | 0 | |
| a-L-Rhap | 3,until C4 | 71.8 | | | 0 | |
| a-L-Rhap | 3,until C5 | 70 | | | 0 | |
| a-L-Rhap | 3,until C6 | 17.5 | | | 0 | |
| a-D-Glcp | 3,3,until | C1 | 96.1 | C1 | H1 | 169 |
| a-D-Glcp | 3,3,until | C2 | 72.4 | | | 0 |
| a-D-Glcp | 3,3,until | C3 | 74.2 | | | 0 |
| a-D-Glcp | 3,3,until | C4 | 70.6 | | | 0 |
| a-D-Glcp | 3,3,until | C5 | 72.48 | | | 0 |
| a-D-Glcp | 3,3,until | C6 | 61.6 | | | 0 |
| b-D-GlcpNAc | 2,3,until | C1 | 102.78 | C1 | H1 | 163 |
| b-D-GlcpNAc | 2,3,until | C2 | 56.3 | | | 0 |
| b-D-GlcpNAc | 2,3,until | C3 | 80 | | | 0 |
| b-D-GlcpNAc | 2,3,until | C4 | 69.7 | | | 0 |
| b-D-GlcpNAc | 2,3,until | C5 | 76.8 | | | 0 |
| b-D-GlcpNAc | 2,3,until | C6 | 61.8 | | | 0 |
| a-L-FucpNAc | 3,2,3,until | C1 | 98.9 | C1 | H1 | 172 |
| a-L-FucpNAc | 3,2,3,until | C2 | 48.98 | | | 0 |
| a-L-FucpNAc | 3,2,3,until | C3 | 77.5 | | | 0 |
| a-L-FucpNAc | 3,2,3,until | C4 | 72.1 | | | 0 |
| a-L-FucpNAc | 3,2,3,until | C5 | 67.7 | | | 0 |
| a-L-FucpNAc | 3,2,3,until | C6 | 16.2 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | C1 | 101.4 | C1 | H1 | 172 |
| a-D-Glcp | 3,3,2,3,until | C2 | 72.4 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | C3 | 73.8 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | C4 | 70.5 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | C5 | 72 | | | 0 |
| a-D-Glcp | 3,3,2,3,until | C6 | 67.4 | | | 0 |