GlycoNovo: De novo structural sequencing framework for N-linked glycans

Qianqiu Zhang[1], Zeping Mao[1], Weiping Sun[2], Xiyue Zhang[2], Ngoc Hieu Tran[2], Lei Xin[2], Baozhen Shan[2] and Ming Li[1]

University of Waterloo, Waterloo, Ontario, Canada
Bioinformatics Solutions Inc., Waterloo, Ontario, Canada

Introduction

N-linked glycosylation is one of the most common and sophisticated post-translational modification (PTM) of proteins. The biological properties of glycopeptides can be greatly affected by the structures of the attached glycans. The use of LC/MS has been demonstrated as an effective method for studying glycans attached to proteins. Due to the topological complexity, de novo structural sequencing of glycopeptides remains challenge in glycoproteomics. Recent de novo sequencing methods aim to identify new glycan structures by modifying existing glycans or using heuristically defined rules. Such methods suffer from scalability and unable to generalize to other species. We propose GlycoNovo, a two-stage framework that builds glycan structures without predefined rules and remain consistent high performance in various tissues and species.

Methods

Workflow

GlycoNovo, a database-independent glycan de novo method for determining the structure of glycans, composes two stages. In the first stage, GlycoNovo derives the compositions of glycans from mass spectra by dynamic programming algorithms. In the second stage, it constructs the topology of the glycan based on derived composition. At each iteration, the model attempts to extract information for predicting the next monosaccharide, starting from the root (i.e. asparagine (Asn) residue of the peptide).

Dynamic programming algorithm to determine the glycan composition

Each spectrum is first searched by peptide database to obtain peptide sequence. The glycan mass is then deduced from the precursor mass and peptide mass. Dynamic programming algorithm is applied to compute the glycan composition based on the glycan mass and the spectrum. Five common monosaccharides are considered: Hex, HexNAc, Fuc, NeuAc, and NeuGc. Note that we consider signature ions to determine the presence of NeuAc and NeuGc for each spectrum. All peaks with mass values greater than the peptide mass (i.e., glycopeptide Y-ions) are used to compute the composition. A table is constructed to store the cumulative intensity and the deduced monosaccharide residues in the paths from peptide mass to precursor mass. When traceback over the table to obtain the most-matched composition, the path has the highest sum of intensity is selected.

Deep learning model to construct the glycan tree

Once the composition determined, the glycan tree is constructed from the asparagine (Asn) residue of the peptide (i.e., the root to leaves by adding various monosaccharides iteratively). At each iteration, a combination of monosaccharides is selected by the deep learning model and attached to a leaf node of the sub-tree obtained from the previous iteration. The iterative process of adding such combinations will result in a branched topology of the glycan tree where multiple monosaccharides can be linked to leaf nodes in the tree. The set of all possible combinations are derived from the training data.

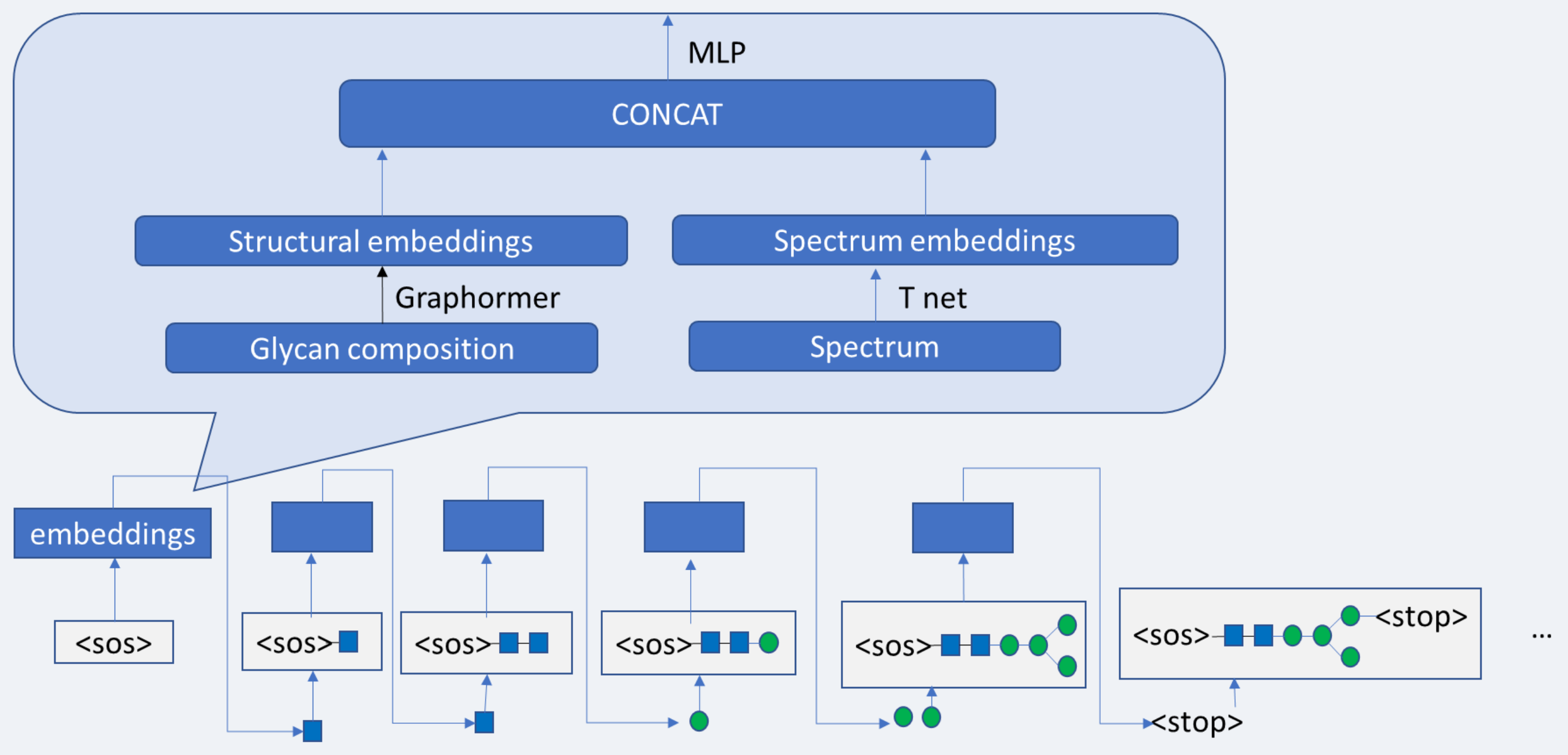


Figure 1. The overall architecture of structure generation model. In a manner similar to RNN settings, our approach involves the extraction of information from both the spectrum and the current substructure at each time step. Each node within the substructure represents compositions derived from dynamic programming. For instance, in the initial time step consists of node features of two Hexes and three HexNAcs, the HexNAc node in the second time step consists of two Hexes and two HexNAcs. These combinations of monosaccharides are integrated into the substructure. The generation process concludes either when all monosaccharides in composition have been utilized or when all child nodes are connected by a sign..

The deep learning model consists of two neural network. One for encoding the topologic knowledge of substructure at each iteration and the other for capturing the matched glycopeptide Y ions between the candidate trees and the spectrum.

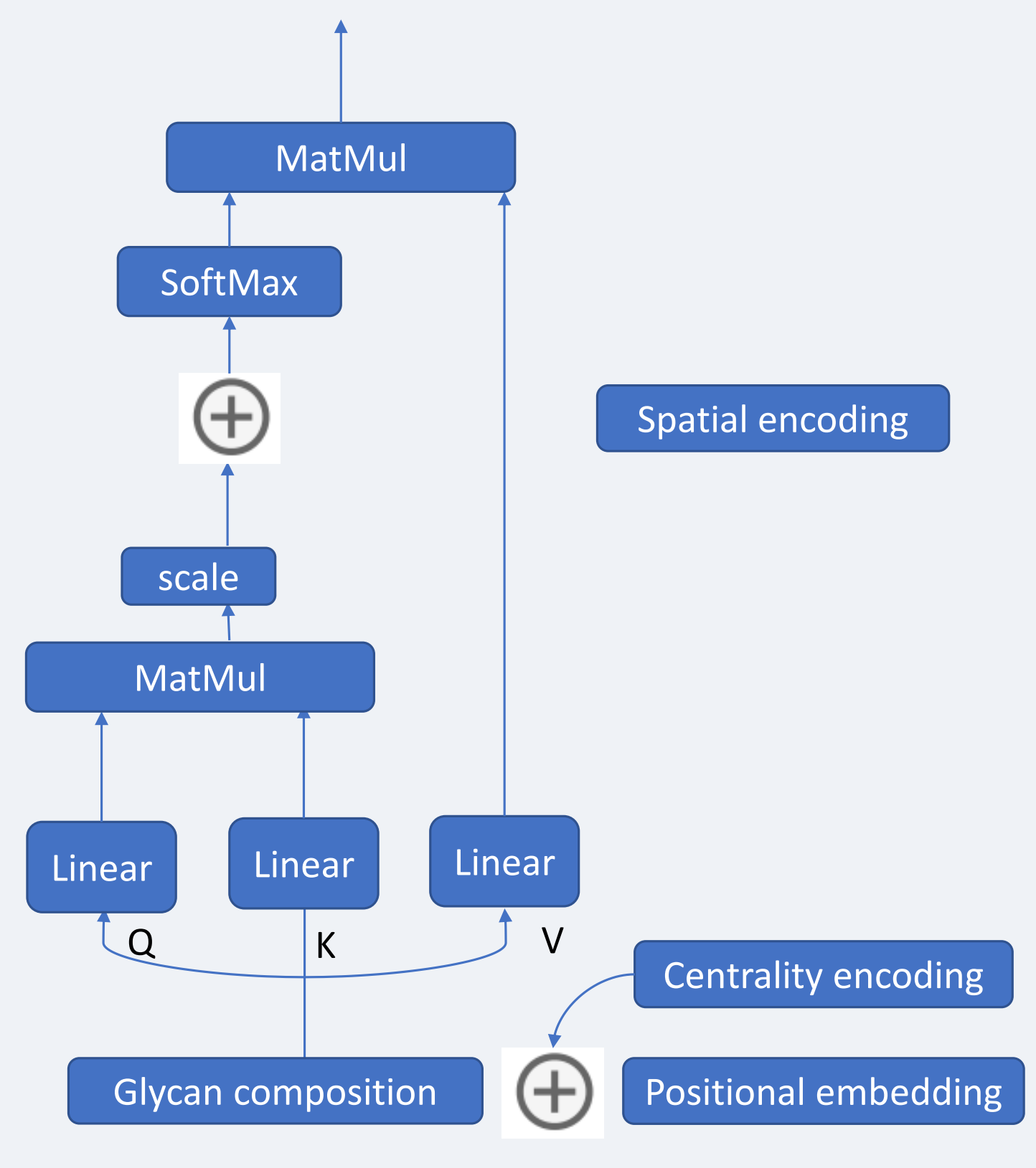


Figure 2. Neural network for encoding glycan sub-structure. We took advantage of the state of the art, Graphormer, to encode structure at each time step. Each node in structure is labeled by breadth-first-search order starting from the root attached to the peptide. As a result, we use positional embedding introduced in the transformer to encode the label and centrality encoding that encodes the in and out degrees of each node.

Results

GlycoNovo has better and consistent de novo glycan sequencing accuracy

GlycoNovo, as the first deep learning model for glycan de novo sequencing, is more efficient and robust than methods relied on fixed rules and predefined criteria. Rule-based models usually lead to uncertainty and inconsistencies when dealing with new data.

We evaluated the performance of GlycoNovo on mass spectrometry/mass spectrometry (MS/MS) data from five mouse tissues (PXD005411, PXD005412, PXD005413, PXD005553, and PXD005555). Our model was trained on a subset of data from four out of five mouse tissues and tested on the remaining tissue. To further assess the model's ability to generalize to new data, we remove the glycan structures appeared in testing set from training set.

We compare our model to StrucGP which is a rule-based de novo sequencing method. GlycoNovo achieved average accuracies of 32%, 83%, and 89% at the three levels structure, fragment ions, and composition, respectively, whereas the accuracies of StrucGP were 23%, 84%, and 85%. While both tools showed comparable accuracies of fragment ions and composition, the structure accuracy of GlycoNovo was substantially higher than that of StrucGP in average and also across all five tissues. The performance of StrucGP indicates that it is overfitting to one tissue and exhibiting fluctuations on other tissues, while our model shows a more consistent accuracy.

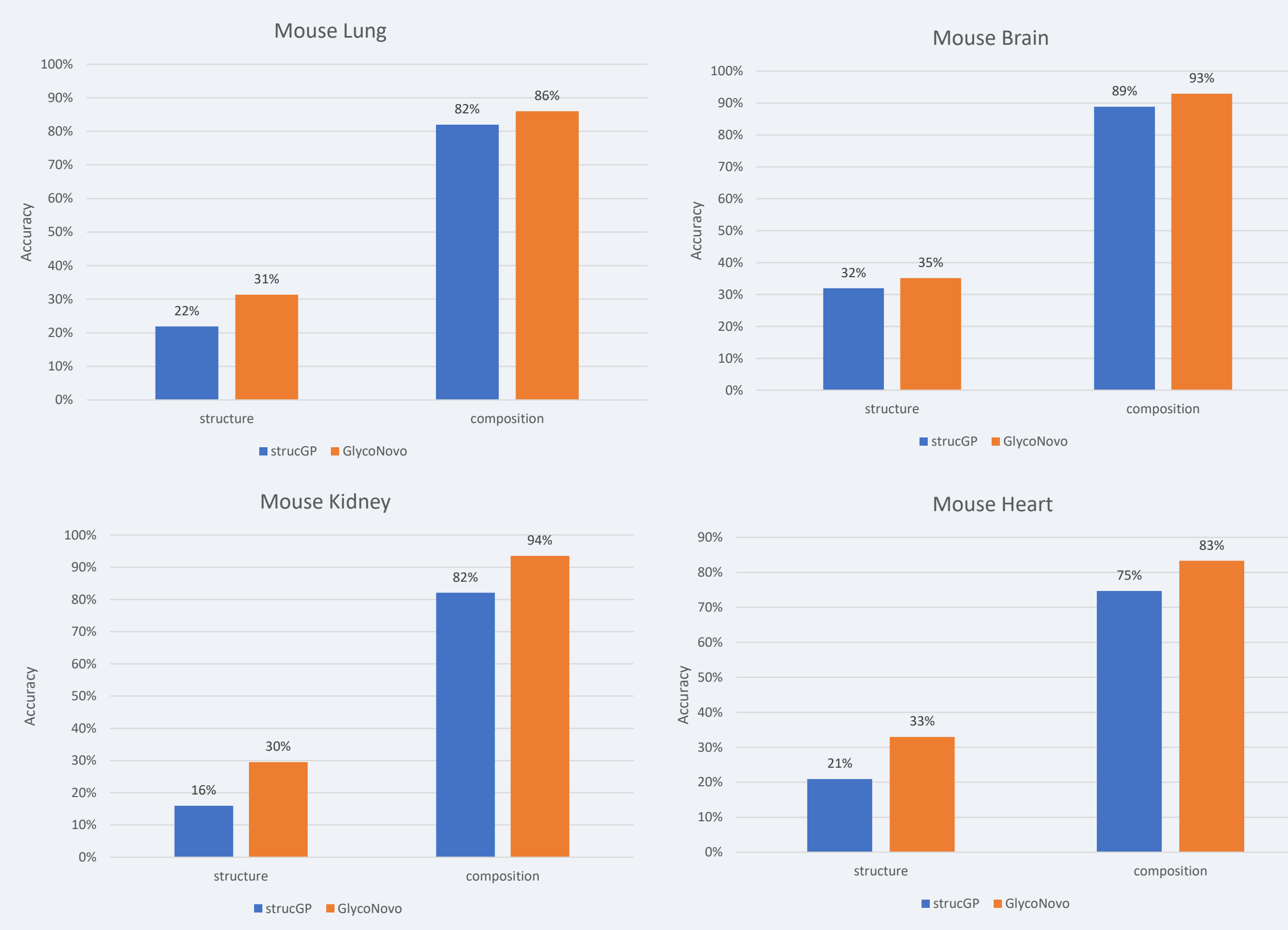


Figure 3. Comparison between GlycoNovo and StrucGP

These results demonstrate the advantage of GlycoNovo’s deep learning model to learn and predict the tree structures of de novo glycans. Following are spectrum examples to further illustrate GlycoNovo’s robustness. Figure 4 shows an examples of de novo glycans that were only discovered by GlycoNovo and were not found in the database. Figure 5b and 5c further show an example of different de novo glycans predicted by GlycoNovo and StrucGP on the same spectra. The glycans predicted by GlycoNovo matched with the respective database search results. In the examples (mouse brain, fraction 1, scan 39012), StrucGP predicted a glycan with three Fucoses, which did not likely represent a correct structure. Despite the limited information provided by the spectrum, our deep learning model is able to give more symmetric topology that aligns better with the biological requirements.

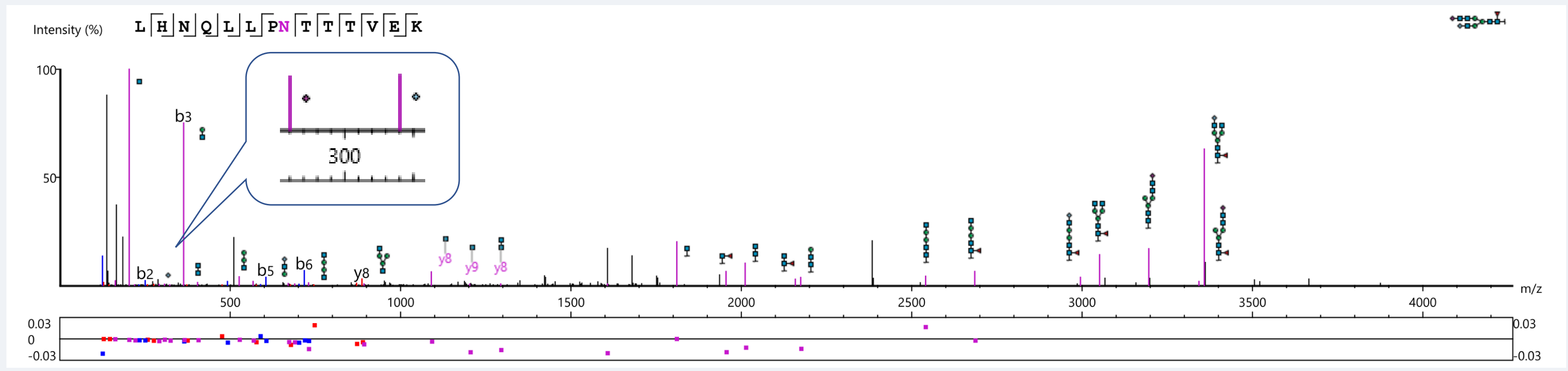
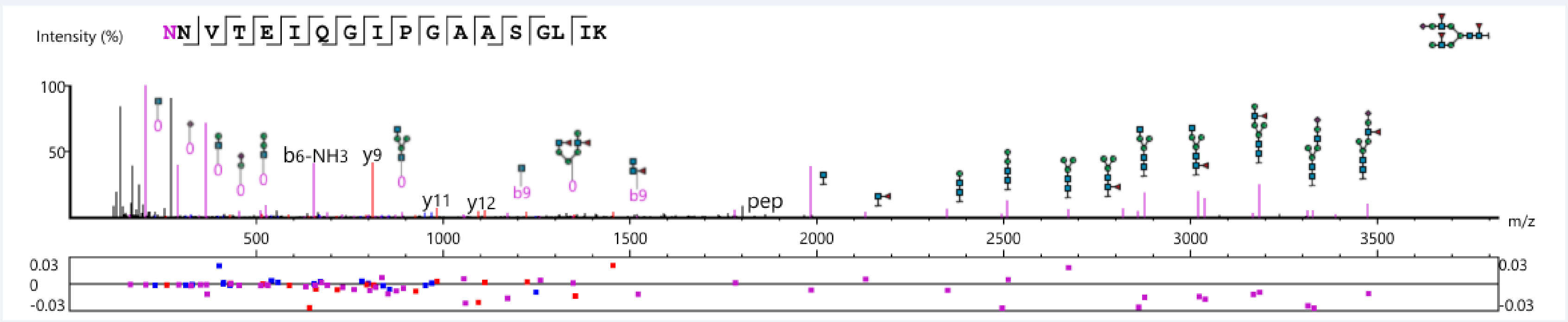


Figure 4. GlycoNovo discovered plausible new structure. The spectrum shows signature ions of both NeuAc and NeuGc. No glycans in the database match both these ions and precursor mass. De novo sequencing technique is able to build plausible new structure that can match the spectrum.

a. StrucGP: (HexNAc)4(Hex)5(Fuc)3(NeuAc)1



b. GlycoNovo: (HexNAc)4(Hex)5(Fuc)1(NeuAc)2

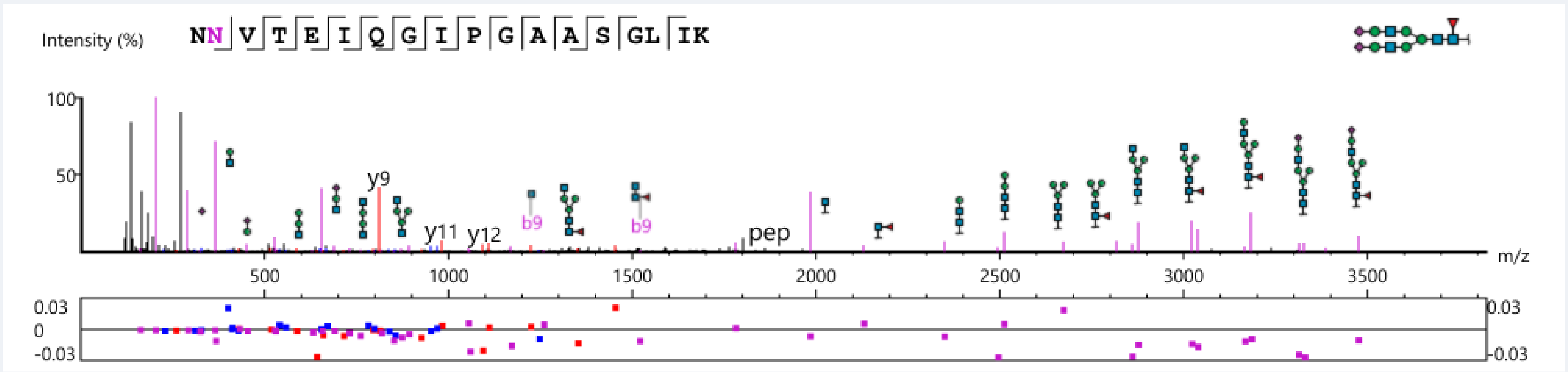


Figure 5. Glycopeptide spectrum annotation for N-linked glycosylation. The matched peptide b/y ions are marked in blue/red, while glycan ions are marked as purple on the spectrum

Contact

Qianqiu Zhang q359zhan@uwaterloo.ca

Summary

- GlycoNovo is a robust de novo sequencing tool for analyzing the structure of glycans.
- This study highlights the capability of deep learning techniques in automating the identification of unknown glycan structures, without human expertise.

The authors declare no competing financial interest.