Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-024-05505-4

RESEARCH PAPER

Deep structure-level N-glycan identification using feature-induced structure diagnosis integrated with a deep learning model

Suideng Qin[1] · Zhixin Tian[1]

Received: 10 June 2024 / Revised: 30 July 2024 / Accepted: 14 August 2024

Abstract

Being a widely occurring protein post-translational modification, N-glycosylation features unique multi-dimensional structures including sequence and linkage isomers. There have been successful bioinformatics efforts in N-glycan structure identification using N-glycoproteomics data; however, symmetric “mirror” branch isomers and linkage isomers are largely unresolved. Here, we report deep structure-level N-glycan identification using feature-induced structure diagnosis (FISD) integrated with a deep learning model. A neural network model is integrated to conduct the identification of featured N-glycan motifs and boosts the process of structure diagnosis and distinction for linkage isomers. By adopting publicly available N-glycoproteomics datasets of five mouse tissues (17,136 intact N-glycopeptide spectrum matches) and a consideration of 23 motif features, a deep learning model integrated with a convolutional autoencoder and a multilayer perceptron was trained to be capable of predicting N-glycan featured motifs in the MS/MS spectra with previously identified compositions. In the test of the trained model, a prediction accuracy of 0.8 and AUC value of 0.95 were achieved; 5701 previously unresolved N-glycan structures were assigned by matched structure-diagnostic ions; and by using an explainable learning algorithm, two new fragmentation features of $m/z = 674.25$ and $m/z = 835.28$ were found to be significant to three N-glycan structure motifs with fucose, NeuAc, and NeuGc, proving the capability of FISD to discover new features in the MS/MS spectra.

Keywords N-Glycan structure · N-Glycoproteomics · FISD · Neural network · Featured motif

Introduction

Protein N-glycosylation functions as a significant post-translational modification and plays important roles in many biological processes [1–3] and pathology processes such as cancers [4–6]. N-Glycoproteomics using liquid chromatography coupled with tandem mass spectrometry (LC–MS/MS) has remarkably facilitated the analysis of N-glycoproteins in the fields of disease diagnosis, therapy, and prognosis [7, 8]. The structures of the N-glycan moiety on the N-glycoproteins bring non-negligible impact on the functions [9], indicating the need of structure-level analysis of the N-glycans.

Because of the high heterogeneity and structural diversity of the N-glycans [10], precise identification of N-glycan structures from MS/MS spectra is challenging. In monosaccharide identification, many tools (such as Byonic [11], Mascot [12], GlycoPeptideSearch [13], GPQuest [14], GlycoPAT [15], pGlyco3 [16], MSFragger-Glyco [17]) have been developed. Feature patterns of N-glycan sequence and motifs (such as fucosylated core and branch, bisecting GlcNAc, sialic acid terminal) can be observed in tandem MS spectra, and structure identification is thus possible [18–21]. In this regard, StrucGP [22] is capable of identifying precise N-glycan structures using a de novo strategy; PEAKS GlycanFinder [23] achieved similar performance using a deep learning model-assisted sequencing method; and GPSeeker [24] established a strategy utilizing the theoretical unique fragments (named structure-diagnostic ions) of A/B/C/X/Y/Z types computed among all sequence isomers of the same monosaccharide composition. However, symmetric “mirror” branch isomers and linkage isomers are largely un-resolved.

Published in the topical collection featuring *Current Progress in Glycosciences and Glycobioinformatics* with guest editors Joseph Zaia and Kiyoko F. Aoki-Kinoshita.

✉ Zhixin Tian
zhixintian@tongji.edu.cn

[1] School of Chemical Science & Engineering, Shanghai Key Laboratory of Chemical Assessment and Sustainability, Tongji University, Shanghai 200092, China

Published online: 30 August 2024

Here, we report deep structure-level N-glycan identification using feature-induced structure diagnosis (FISD) integrated with a deep learning model. A neural network model is integrated to conduct the identification of featured N-glycan motifs and boosts the process of structure diagnosis as well as the distinction of linkage isomers. By adopting publicly available N-glycoproteomics datasets of five mouse tissues (17,136 intact N-glycopeptide spectrum matches) and a consideration of 23 motif features, a deep learning model integrated with a convolutional autoencoder (CAE) and a multilayer perceptron (MLP) was trained to be capable of predicting N-glycan featured motifs in the MS/MS spectra with previously identified compositions. In the test of the trained model, a prediction accuracy of 0.8 and AUC value of 0.95 were achieved; 5701 previously un-resolved N-glycan structures were assigned by matched structure-diagnostic ions; and by using explainable learning algorithm, two new fragmentation features of $m/z = 674.25$ and $m/z = 835.28$ were found to be significant to three N-glycan structure motifs with fucose, NeuAc, and NeuGc, proving the capability of FISD to discover new features in the MS/MS spectra.

Methods

Data acquisition and processing The N-glycoproteomics raw datasets of the five mouse tissues were downloaded from the PRIDE Archive [25] and Proteome Xchange [26] using accession numbers of PXD005411, PXD005412, PXD005413, PXD005553, and PXD005555 [27]. The N-glycoproteomics raw datasets of the four standard N-glycoproteins and the StrucGP identification results of these standard N-glycoproteins together with the aforementioned five mouse tissues were downloaded using the accession number of PXD025859 [22] covering the treatments of Untreated, De-sialylated, Cut HexNAc, Cut Gal Cut HexNAc, and Cut Fucose. All the raw datasets were converted to the.mzML format using ProteoWizard MSConvert (version 3.0.21335-35327b7) [28, 29] with the parameters of mzML output format, 64-bit binary encoding precision, Vendor Peak Picking Filter at MS levels 1 ~ 2, and the default Title Maker Filter.

Acquisition of the intact N-glycopeptide identification results The mouse proteome fasta database (17,202 entries) was downloaded from Uniprot with the UniProtKB accession of UP000000589 and reviewed filtration. All raw data from the five mouse tissues were searched using pGlco3 (version 3.1) and MSFragger-Glyco (FragPipe v21.1). The parameters adopted for pGlyco3 are the following: the file type was set to raw, the fragmentation method was set to HCD, trypsin was used as the enzyme and max missing cleavage was set to 2, carbamidomethyl on C was set as fixed peptide modification and oxidation on M was set as dynamic modification, max variable modification on peptides was set to 3, peptide length was set from 6 to 50, peptide mass range was set from 400 to 5000, the glycan database was set to “pGlyco-N-Mouse-Large.gdb” with “N-Glycan” set as glycan type, the precursor tolerance was set to 10 ppm and the fragment tolerance was set to 20 ppm, the other parameters were kept as default. The parameters for FragPipe were the same as pGlyco3 except the following settings: “glyco-N-HCD” was set as the workflow, the mouse proteome database was set with added 50% decoys using the Database module, the PTM-Shepherd was activated and set to “Glyco Search,” the other parameters were kept as default.

Acquisition of the sulfated intact N-glycopeptide identification results using GlycReSoft The GlycReSoft tool was downloaded in 0.4.22 version. The peptide backbones were extracted from the GPSMs used in the model training and testing, and were compiled into a fasta file in which each peptide was deposited as a single protein (824 peptides, Supplementary Material S1). All glycan compositions from the intersection of pGlyco3 mouse N-glycan database, MSFragger-Glyco, and our glycan structure definition rules were adapted by adding 1 ~ 3 sulfates iteratively for each composition (also restricted that the number of sulfate should not be more than HexNAc, 4272 glycan compositions, Supplementary Material S1). For building the glycopeptide search space, carbamidomethyl on C was set as constant peptide modification and oxidation on M was set as variant modification. No enzyme was chosen because the fasta file already presented cleaved peptide sequences. Other parameters for building the glycopeptide search space were set as default. All 25 mouse five tissue mzML data files were added to the workspace using preset configurations of “LC–MS/MS Glycoproteomics,” leaving other parameters to remain default. The glycopeptide search parameters were set to default except a 500 Batch Size and a 500 Minimum Search Mass. Other key default parameters included 10 ppm MS1 Error Tolerance and 20 ppm MS2 Error Tolerance. The search results were filtered before being used in the estimation process with: q-value < 0.05, is_best_match = True and possesses the highest ms2_score in the MS2 scan.

Definition of the putative N-glycan branches In FISD, all branch structures were derived from pre-defined 18 base branch structures, which in the text format [30] were constructed from the combinatorial structures of branch, sequence, linkage and N-glycan types in the glycan structures. The glycan structure code formatting is illustrated in Supplementary Fig. S1. For example, the LacNAc structure “GlcNAc-β 4–1-Gal” was encoded as “Y41L” with the linkage numbers remained; “T” code was specified for NeuGc;

a question mark "?" was specified as an uncertain linkage number and the *α* 2,3 and *α* 2,6 linkages of the terminal sialic acid residues were coded together as "?2S." Among the 18 base branch structures (Table 1), 12 were determined using the mouse N-glycan branches provided by StrucGP in the study of low HCD spectra, and six came from consideration of the existing biosynthesis-based summaries [31–35]. All possible branch structures, annotated as the same "source" (Supplementary Table S1), were generated by detaching one monosaccharide residue a time from each of the base branches. For the intact N-glycan structure, the types of complex and hybrid were taken into consideration. For complex N-glycans, a max of four antennae was stipulated with the fixed "locations" of *β* 1,2 and *β* 1,4 on the *α* 1,3 mannose residue of the N-glycan core, together with *β* 1,2 and *β* 1,6 on the *α* 1,6 core mannose residue. Then, all the generated branches were iteratively connected to the fixed four "locations" to construct the intact N-glycan structure following the following specificities. If two branches were connected to the same core mannose residue, they should be from the same group (i.e., from the same base branch); if there was only one branch connected to the *α* 1,3 or *α* 1,6 core mannose residue, the branch should always be determined as *β* 1,2 linkage; for the core structures, fucosylation and bisecting GlcNAc were also adopted into the exhaustive combination. For hybrid N-glycans, a max of four mannose residue was limited on the *α* 1,6 core mannose residue.

Definition of the N-glycan motif features Thirty N-glycan motif features were defined based on the base branches and core structures (Table 2). According to the structure rules, a single intact N-glycan structure might contain more than one defined motif features. A 30-dimension vector was utilized to represent the presences of the defined features in a single N-glycan. Each dimension stood for a fixed defined feature and the values 0 (or "False") and 1 (or "True") represent absence and presence of the feature, respectively. All the 30-dimension feature vectors were calculated for each of the possible intact N-glycan structure (Supplementary Table S1). After the evaluation of the intact N-glycopeptide spectrum matches (GPSMs) from the training set, seven features were removed because of insignificant presence and 23 features were kept for further test evaluation.

Construction of the deep learning neural network model for the prediction of defined motif feature structures The first set of GPSMs for training and testing the model was extracted from the common ones (same peptide sequences, modifications, and glycan compositions) simultaneously reported by pGlyco3, MSFragger-Glyco (FragPipe), and StrucGP in the five mouse tissues. The true labels for the training set, that is, the confirmed total 30 motif features regarded as the true positive references of each GPSM in the training set, were calculated using the reported N-glycan structures from StrucGP, leaving no true positive references for the motif features related to the six base branches that exceeded the collections of StrucGP. In order to obtained

Table 1 Base branches and corresponding derived branch structures defined in the N-glycan structure rules, the text coding format follows the previous work [30], where a question mark "?" stands for an uncertain linkage number and the code "T" stands for NeuGc

Group	Base branch structure	Derived branch structure
1	Y	/
2	Y41L	Y
3	Y(31F)41L	Y41L; Y31F; Y
4	Y(31F)41L?2S	Y41L?2S; Y(31F)41L; Y41L; Y31F; Y
5	Y(31F)41L?2T	Y41L?2T; Y(31F)41L; Y41L; Y31F; Y
6	Y(31F)41L?1F	Y41L?1F; Y(31F)41L; Y41L; Y31F; Y
7	Y(31F)41L31L	Y41L31L; Y(31F)41L; Y41L; Y31F; Y
8	Y(31F)41V	Y41V; Y31F; Y
9	Y41V?2S	Y41V; Y
10	Y(62S)41L?2S	Y41L?2S; Y(62S)41L; Y41L; Y
11	Y(62T)41L?2S	Y41L?2S; Y(62T)41L; Y41L; Y
12	Y(62T)41L?2T	Y41L?2T; Y(62T)41L; Y41L; Y
13	Y(62S)41L32S82S	Y41L32S82S; Y(62S)41L?2S; Y41L?2S; Y41L; Y
14	Y41L31Y	Y41L; Y
15	Y41L31Y(31F)41L	Y41L31Y41L; Y41L31Y31F; Y41L31Y; Y41L; Y
16	Y41L31Y(31F)41L?2S	Y41L31Y41L?2S; Y41L31Y(31F)41L; Y41L31Y41L; Y41L31Y31F; Y41L31Y; Y41L; Y
17	Y41L31Y(31F)41L?2T	Y41L31Y41L?2T; Y41L31Y(31F)41L; Y41L31Y41L; Y41L31Y31F; Y41L31Y; Y41L; Y
18	Y41L31Y(31F)41L?1F	Y41L31Y41L?1F; Y41L31Y(31F)41L; Y41L31Y41L; Y41L31Y31F; Y41L31Y; Y41L; Y

Table 2 Defined N-glycan motif features

Name	Abbreviation	Structure	Removed	Index
Core Fucosylation	CrFuc	Y(61F)	N	1/1
Bisecting GlcNAc	Bisect	Y41M(41Y)	N	2/2
Branch Fucosylation	BrcFuc	Y(31F)	N	3/3
Terminal Sialylation NeuAc	TmSiaAc	Y41L?2S	N	4/4
Terminal Sialylation NeuAc Fucosylated	TmSiaAcF	Y(31F)41L?2S	N	5/5
Terminal Sialylation NeuGc	TmSiaGc	Y41L?2T	N	6/6
Terminal Sialylation NeuGc Fucosylated	TmSiaGcF	Y(31F)41L?2T	N	7/7
Terminal Fucosylation	TmFuc	Y41L?1F	N	8/8
Terminal Fucosylation Fucosylated	TmFucF	Y(31F)41L?1F	N	9/9
Terminal Digalactose	TmDiGal	Y41L31L	Y	10/-
Terminal Digalactose Fucosylated	TmDiGalF	Y(31F)41L31L	Y	11/-
N,N′-Diacetyllactosamine	LacDiNAc	Y41V	N	12/10
N,N′-Diacetyllactosamine Fucosylated	LacDiNAcF	Y(31F)41V	N	13/11
Terminal Sialylation LacDiNAc	TmSiaDN	Y41V?2S	N	14/12
Branch Sialylated NeuAc	BrcSiaAc	Y(62S)41L	N	15/13
Terminal Sialylation NeuAc with Branch Sialylated NeuAc	TmSiaAbrcAc	Y(62S)41L?2S	N	16/14
Branch Sialylated NeuGc	BrcSiaGc	Y(62T)41L	N	17/15
Terminal Sialylation NeuGc with Branch Sialylated NeuAc	TmSiaAcBrcGc	Y(62S)41L?2T	Y	18/-
Terminal Sialylation NeuGc with Branch Sialylated NeuGc	TmSiaGcBrcGc	Y(62T)41L?2T	N	19/16
Terminal DiNeuAc	TmDiSiaAc	Y41L32S82S	N	20/17
Terminal DiNeuAc with Branch Sialylated NeuAc	TmDiSiaAcBrcAc	Y(62S)41L32S82S	N	21/18
Extended LacNAc	ExLacNAc	Y41L31Y41L	N	22/19
Extended LacNAc Fucosylated	ExLacNAcF	Y41L31Y(31F)41L	N	23/20
Terminal Sialylation NeuAc with Extended LacNAc	TmSiaAcExLN	Y41L31Y41L?2S	N	24/21
Terminal Sialylation NeuAc with Extended LacNAc Fucosylated	TmSiaAcExLNF	Y41L31Y(31F)41L?2S	Y	25/-
Terminal Sialylation NeuGc with Extended LacNAc	TmSiaGcExLN	Y41L31Y41L?2T	N	26/22
Terminal Sialylation NeuGc with Extended LacNAc Fucosylated	TmSiaGcExLNF	Y41L31Y(31F)41L?2T	Y	27/-
Terminal Fucosylation with Extended LacNAc	TmFucExLN	Y41L31Y41L?1F	Y	28/-
Terminal Fucosylation with Extended LacNAc Fucosylated	TmFucExLNF	Y41L31Y(31F)41L?1F	Y	29/-
Hybrid Category	Hybrid	Y41M61M61M	N	30/23

true labels for all motif features, further selection was carried out from the GPSMs that co-reported by pGlyco3 and MSFragger-Glyco in five mouse tissues as supplements to the training set. The GPSMs were selected by matching the fragments in a manually defined ion list (Supplementary Table S2a) created using the in silico fragmented ions within the type of B/C and ions produced by two times of fragmentation named as type BY/BZ/CY. Oxonium ions with *m/z* = 274.0927 for motif feature structures containing NeuAc and *m/z* = 290.0876 for NeuGc were also taken into account. If all the fragments in the list were matched, the corresponding motif feature was determined as truly presented in the spectrum and the GPSM could be finally selected as a supplement. The second set of GPSMs was from the results reported by StrucGP in the four standard glycoproteins. The third set of GPSMs was from GPSMs common to pGlyco3 and MSFragger-Glyco but not StrucGP in the five mouse tissues, where structures (other than those of StrucGP) with any of the defined motif features were also generated and included.

All GPSMs with the corresponding peptide backbones and monosaccharide compositions were embedded into a high-dimensional matrix before inputted to the CAE layers of the neural network model. For each GPSM, a fragment *m/z* list was computed by merging all theoretical fragments of all N-glycan structures with the same monosaccharide composition. Both the canonical fragments of B/Y/C/Z and internal fragments of BY/BZ/CY/YY/YZ/ZZ are included for consideration. As for charge states, + 1 and + 2 for Y/Z/YY/YZ/ZZ ions, + 1 for B/C/BY/BZ/CY ions, and + 2 for ions with *m/z* larger than 900 were calculated. All experimental isotopic peaks were then assigned to the merged fragment *m/z* list, glycan-related signals were collected, the intensity of each fragment peak was normalized to the summed intensity of the Y1 ion and the precursor ion in the MS2 spectrum with charges 1 and 2, and finally, the normalized intensities were embedded to the input matrix at the specified location based on the target motif feature, ion type, and monosaccharide composition of the matched fragment.

Oxonium ions were also considered as an individual fragment ion type in the input matrix. The matrix was shaped as dimensions of $30 \times 5 \times 6 \times 6 \times 5 \times 4 \times 4$, where the first dimensional "30" represents the first 30 defined motif features; the second dimensional "5" represents the fragmentation types of oxonium, B/C, BY/BZ/CY, Y/Z, and Y/Z/YY/YZ/ZZ, respectively; and the remaining five dimensions sequentially represent the difference values of HexNAc, Hex, Fuc, NeuAc, and NeuGc at the monosaccharide composition level between the composition of matched fragments and target featured motif. The difference value started from -2 to 3, 3, 2, 1, 1 respectively. For example, a B fragment HexNAc(1)Hex(2)Fuc(1) matched against the motif feature TmSiaAc (abbreviation referred to Table 2) with index of 4 and composition of HexNAc(1)Hex(1)NeuAc(1) would be placed to location (4, 2, 3, 4, 4, 2, 3) with the composition difference calculated as HexNAc(0)Hex(1)Fuc(1)NeuAc(-1) NeuGc(0). For each defined motif feature, if it was not contained in the N-glycans generated from the identified composition, it would be blocked and excluded from the embedding step, while all contained features would iteratively become the target feature to conduct spectrum embedding. It is worth stressing that this "ignorance" only takes place in the embedding step, aiming at preventing the embedding step from introducing illogical data to the model (for example, inputting NeuAc-relevant data from a spectrum that in fact contains no NeuAc). On the other hand, however, it would certainly introduce false negatives to the model prediction results due to the identified composition might be a false positive. In the prediction and validation steps, all motif features will be predicted regardless of the identified composition and the theoretically blocked features, indicating that we adopted a feature-unblocked model predictive method. The fragment list generating method and data embedding method made full use of the precise candidate N-glycan structures, and the original features were kept as many as possible to transfer the spectrum data into fragment correlations stored in the spatial structure of the input matrix.

The CAE model was trained and responsible for reducing the dimensionality of the embedded matrix data and decoding the convoluted data to retrieval the input data. The model consisted of several 3D convolution layers and used ReLu as activation function with a drop-out layer to realize regularization handling. A weighed mean squared error (MSE) loss was designed as the loss function for training model and function as the 0.2 total MSE combined with a 0.8 MSE between all non-zero values so that it would lower the impact of large proportions of zero values in the input matrix (source code provided as supplemental information). The inputted $30 \times 5 \times 6 \times 6 \times 5 \times 4 \times 4$ matrix was convoluted to a 128-dimension vector through the CAE model and can be decoded to rebuild the input data.

The MLP model was trained using the encoded 128-dimension vector and true labels of defined motif features to perform prediction of the 23 features. The model consisted of several fully connected layers (FC) and used ReLu as activation function with a drop-out layer to realize regularization handling. The last output FC layer was set to 23 dimensions and a weighted binary cross-entropy (BCE) loss was designed as the loss function for model training. The weight values were calculated based on the presence proportions of each of the 23 motif features (source code available at GitHub repository https://github.com/QinSD-TJU-Chem-BAMS/FISD-Core-Scripts-and-Datasets, Supplementary Material S2). Before MLP training, a random over-sampling strategy was adopted to balance the true label distributions of the motif features, in which the least six presented features were targeted and GPSMs that contained any of these six features were collected and copied to reach the proportion of 20% in the total GPSMs. Each copy introduced a random change of less than 2% to all the copying values in encoded GPSM data to avoid overfitting. In the dataset splitting step, an iterative stratification algorithm was utilized to keep all the true labels equally distributed in both the training set and test set.

Evaluating the performance of the deep learning model on the prediction of defined N-glycan motif features The model was evaluated using a *K*-fold method on the training split dataset and tested on the test split dataset from the GPSMs of the five mouse tissues. For each GPSM, the predictive probabilities of each single defined N-glycan motif feature were produced by the model. The presence of all the 23 features was finally determined using the output 23 probabilities and all candidate feature vectors deduced from the N-glycan structure rules based on the identified composition of the GPSM. Features that remained 0 or "False" in all candidate feature vectors were removed from this GPSM. Then, for each candidate feature vector, the remaining features were compared between the predicted probability and candidate label to calculate the customized BCE value (Eq. 1),

$$\mathrm{BCE} = -\frac{1}{N}\sum\nolimits_{i}^{N}[y_i \ln\left(P_i\right) + (1 - y_i)\ln(1 - P_i)] \tag{1}$$

where N is the number of remaining features, y_i is the 0 or 1 value of feature i in the candidate feature vector, and P_i is the predictive probability of feature i. The candidate feature vector with the lowest BCE value was determined as the final matched feature vector. The accuracies and F1 scores at the single feature level were calculated using the predicted 0 or 1 label value of the predicted feature vector. At the total feature level, or equally for the entire feature vector, the binary tag was defined by whether the sum of the label values of features remained from the aforementioned selection was larger than the remaining feature count. Then, the predictive probability of the entire feature vector was calculated by the average value of the predictive probability of the positive label around all remaining features. The ROC curve data and AUC values of the model performance on the test split dataset were calculated using the defined binary label and the calculated probability of the entire feature vector in each GPSM. There were two criteria for estimating the accuracy and F1 score at the total dataset level. The final trained model was evaluated on the GPSMs from the four standard glycoproteins and the same algorithm was used to calculate the ROC curve data and AUC value.

Evaluation of the FISD strategy The structure diagnosis step was proceeded by computing all candidate N-glycan structures based on the given feature vector and identified composition. Then, theoretical fragments of each candidate were calculated using the precise glycan structure considering the ion type of A/X/B/Y/C/Z, where the A/X type fragments could be calculated only when given the detailed linkage numbers between monosaccharide residues. Then, the structure-diagnostic ions for each candidate, that is, the unique fragments of a specific glycan structure that have different mass from all the other fragments generated by its isomers, were determined by checking whether the ion formula was shared by other candidates. Finally, all candidate N-glycans were assigned in the MS/MS spectrum by matching their structure-diagnostic ions, generating results of single or multiple glycans with matched structure-diagnostic ions. Certainly, there were glycans possessing no theoretical structure-diagnostic ions or glycans with no matched structure-diagnostic ions. The former condition was handled that all those glycans were kept as plausible identifications waiting for a better motif feature grouping strategy or other methods to conduct distinguishment and validation. The latter condition was marked as a failure match. For the branch location isomers, there were almost no B/Y/C/Z type fragments that could be unique to the other isomers, emphasizing the importance of A/X type fragments in structure diagnosis step for they would theoretically change when the location or linkage number of the branch structure change.

When a spectrum was assigned to have more than one N-glycan with matched structure-diagnostic ions, a diagnosis score was calculated for each N-glycan (Eq. 2),

$$\text{score} = \ln(\frac{n}{N} + 1) \sum_{i}^{n} A_i \tag{2}$$

where n is the number of matched structure-diagnostic ion regardless of the charge states, N is the number of theoretical structure-diagnostic ions, and A_i is the abundance or intensity of the ion i matched MS/MS peak. The final selected N-glycans as the structure-level identification result were determined using an iterative algorithm to figure out the top ranked N-glycan group whose average score was required to be at least three times as the highest score out of the top ranked group. Moreover, filtered N-glycans with matched structure-diagnostic ions were still kept in the output results for further study inspiration.

Explainable learning of the neural network model The CAE model was trained to decode the convoluted vector data to the input dimensions, reducing the neural network embedded data to the data with practical significance. The same weighted MSE loss function as used in training the CAE model was applied to evaluate the loss in the decoding process. The method of Shapley additive explanations (SHAP) was used in the MLP model to quantify the contributions of each of the input 128 dimensions from upper CAE layers. Three infrequently presented N-glycan motif features were selected as target features to perform explainable learning on the GPSMs from the test split dataset of mouse five tissues and four standard glycoproteins with the presence of target features in true labels. Then, the SHAP values were calculated for each target feature acquiring the contribution level of the 128-dimension encoded vectors of each GPSM. Then, a fixed interference of a 100 times expansion was added to the top 4 prominent contributed vector dimensions filtered by the SHAP values. The interfered vector and original vector were both decoded to the embedded matrix format by the CAE model. The two decoded matrices were compared and the values with 5 times expanded or 0.2 times narrowed would be considered as activated elements and the corresponding matrix spatial locations were extracted to reduce to monosaccharide compositions and fragment types. By collecting the activated values in the decoded embedded matrices in all target GPSMs against the target features, the significances of fragments reduced from activated locations in the decoded matrices were evaluated using the presence frequencies and manual check of the fragment types and monosaccharide compositions.

Results

Development of the FISD strategy for the deep structure-level identification of N-glycans on intact N-glycopeptides A series of pre-determined N-glycan structure rules (Table 1) and a list of pre-defined N-glycan motif features (Table 2) is needed and applied to FISD. Detailed glycan structure code formatting followed our previous work [30] and is displayed in Supplementary Fig. S1. FISD consists of a feature-induced step and a structure diagnosis step (Fig. 1). The inputs are intact N-glycopeptide MS/MS spectra with identified peptide backbones and glycan compositions from other tools or software. The feature-induced step was first proceeded, where a trained deep learning neural network model integrating a CAE and a MLP was adapted and applied to predict and identify the existence of the N-glycan motif features from the spectra. The identified motif features were then transferred to the glycan structure rules to generate all candidate structural isomers of which the number was much smaller than those computed from merely the given compositions, prominently increasing the numbers of the theoretical structure-diagnostic ions for each isomer and thus enabling the structure diagnosis step at such deep level. For the structure diagnosis step, all generated candidate structures were evaluated for the possession of theoretical structure-diagnostic ions and were annotated in the corresponding MS/MS spectra to finally confirm the structure, resulting in either single, multiple, or theoretically indistinguishable structural identifications in a single MS/MS spectrum.

The customized N-glycan structure rules took branch location isoforms and carbon linkage numbers into consideration, making it cover an ultra-deep isomer diversity. Moreover, the linkage numbers such as the β 1,4 linkage presented in the traditional LacNAc (GlcNAc-β 4–1-Gal) branch facilitated the diverse in silico fragmentation of A/X type ions, which also enhanced the depth of theoretical unique fragments. The rules and pre-determined branch structures were filtered and deduced from existing summaries and confirmed studies [1, 22, 31–36], including all the 17 mouse N-glycan branches confirmed from the low-energy HCD experiments by StrucGP [22]. As a result, a total of 32 branch structures derived from 18 base branches attached with 30 motif features were adopted (Table 2), exhaustively constructing 217,896 N-glycans of complex and hybrid types majorly targeting the mouse species and mammals (Supplementary Table S1, also deposited in Supplementary Material S2), as well as covering 1424 mouse N-glycan compositions out of 1600 compositions of complex and hybrid type provided by pGlyco [27] and simultaneously used by the MSFragger-Glyco (FragPipe) [17].

The pre-defined N-glycan motif features were designed to fit for the structure rules under the criteria of representing adequately featured N-glycan motifs that simultaneously group the isomers and are potentially capable of generating distinct fragments or fragmentation patterns with detectable regularity. Each N-glycan structure might contain more than one motif feature and could be grouped into specific feature group based on its contained feature motifs. The number of the features was finally reduced from 30 to 23 for the over-imbalanced distributions in the training spectra datasets.

In the feature-induced step, a deep learning neural network model was designed for predicting the motif features contained in the N-glycan at the GPSM level. By submitting the MS/MS spectrum data with the corresponding peptide backbone and monosaccharide composition results by N-glycoproteomics identification tools, the model would embed the spectrum data into a high-dimensional matrix based on the pre-identified composition and sent it to the input layer of the convolutional neural network. After the CAE and MLP layers processed, a 23-dimension vector was

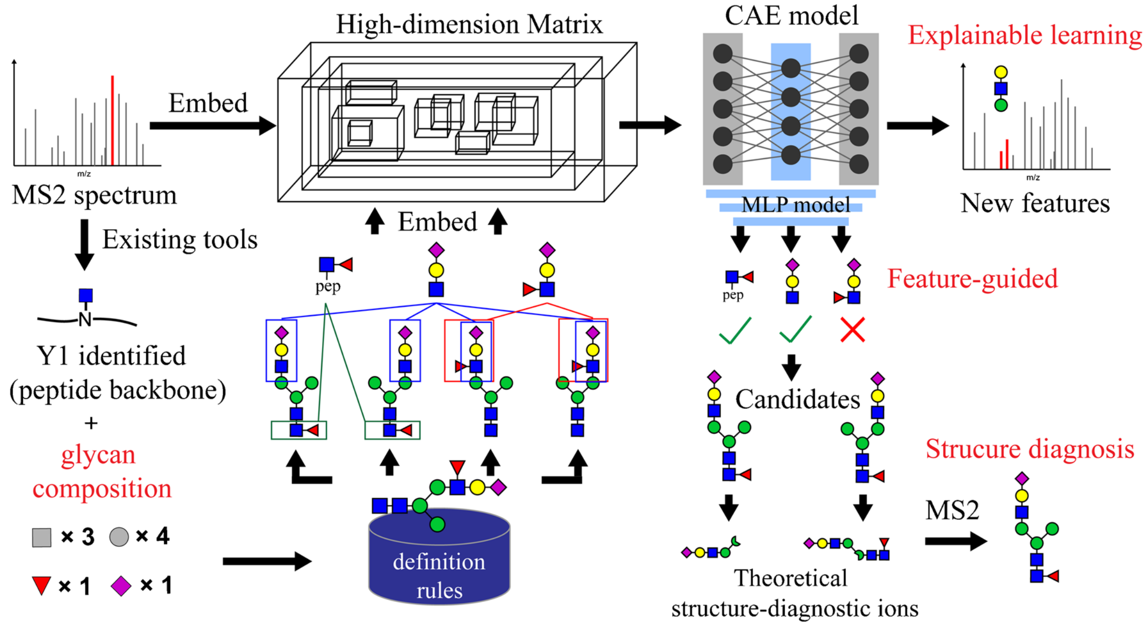


Fig. 1 The workflow of the FISD strategy

outputted as the prediction probabilities of all 23 pre-defined features. The final prediction of the features was generated by referring to the structure rules to find the best match of all features.

With the identified motif features of the N-glycan and structure rules, all possible intact glycan structures were constructed with a much smaller number of isomers than those generated from just a given identified composition, leading to a considerable proportion of isomers that could theoretically produce unique fragments as structure-diagnostic ions. Finally, the structure diagnosis step was proceeded by matching theoretical structure-diagnostic ions in the experimental MS/MS spectrum to conclude the precise N-glycan structure. The result could be either a single match, multiple matches, or plausible structures without theoretical structure-diagnostic ions.

Moreover, the CAE and MLP models were capable of explainable learning, which offered the calculated contributions and importance of each input spectrum signal to the predicted N-glycan motif features, suggesting that the FISD strategy was able to discover new signal patterns presented in the experimental spectra that closely related to the structure feature and broaden the usage of indirect fragments or fragmentation patterns in the identification procedure.

Training and estimation of the deep learning model for predicting the motif features of N-glycans The N-glycoproteomics MS/MS spectra datasets for training and testing the model were acquired from mouse five tissues [27] with co-identified intact N-glycopeptide spectrum matches (GPSMs) reported by pGlyco3 [16], MSFragger-Glyco (FragPipe), and StrucGP. The spectra of four standard glycoproteins acquired by the StrucGP developer team [22] were also used to conduct extra model validation. A total of 17,136 GPSMs from all the five replicate LC–MS/MS experiments of the five mouse tissues (brain, heart, kidney, liver, and lung, total 25 raw files, Supplementary Table S2b) were selected as training and testing datasets, covering all defined motif features with most of which reported to have the same identifications of peptide sequences, modifications, and glycan compositions across pGlyco3, MSFragger-Glyco, and StrucGP, while a small number of the GPSMs were not reported by StrucGP due to their assignments to the featured N-glycan branches which exceeded the branch collections used by StrucGP. The true labels of the N-glycan motif features of each GPSM were deduced from the identified intact N-glycan structure in StrucGP results, while a small number without StrucGP support were confirmed using the fragment list (Supplementary Table S2a), regardless of the model-driven feature prediction. This labeling method would obviously introduce false positive labels to the training data due to the inevitable false positive structure identification presented by StrucGP. Therefore, using motif features (small partial glycan structures) as label rather than the intact glycan structure would to a large extent ensure the correction of feature motif labels and exclude the impact of reported incorrect intact glycan structure, making full use of the existing high-performance tools. Still and all, few false positive labels were inevitably introduced and it remained as a known issue. As a result, a total of 2943 GPSMs from the four standard glycoproteins reported from StrucGP [22] were also adopted as testing data, including samples with different treatments (Supplementary Table S2c).

The model was first trained on 80% of the GPSMs from the five mouse tissues using a five-fold *K*-fold estimation strategy. The motif feature labels were considered to be symmetrically split and the remaining 20% of the GPSMs were used as a test set. A random over-sampling method was adapted to further balance the label distributions of the features with minor proportions. At the single feature level, the model achieved ultra-high accuracies at about 0.95 ~ 0.99 and F1 scores > 0.85 for most of the motif features in both *K*-fold circulations and predictions on the 20% test set (Fig. 2a), revealing a high performance of the deep learning model in predicting single motif feature. The prediction accuracy of all pre-defined motif features together in a single spectrum was estimated in two different criteria. Criterion 1 is identical to presented true features, which required that all 23 features should be correctly predicted in a single spectrum; and criterion 2 covers all presented true features, allowing extra positive prediction of features that true labels do not provide but reject any of the missing prediction of features that are positively provided by true labels. The compiled predictive probability of all features was calculated using a structure rules-based method. As a result, an average of 0.81 criterion 1 accuracy accompanied with an average F1 score of 0.77 and an average of 0.87 criterion 2 accuracy with an average F1 score of 0.83 were reached around the five circulations of *K*-fold estimation. A 0.80 accuracy of criterion 1 with an F1 score of 0.76 and a 0.87 accuracy of criterion 2 with an F1 score of 0.84 were acquired in the 20% test set. The receiver operating characteristic (ROC) curve was also calculated (Fig. 2b) to further evaluate the model performance. The area under the ROC curve value was 0.95 on the test set for calculation using the compiled predictive probabilities at the level of all features.

The final model was trained on all the GPSMs from the five mouse tissues and was evaluated on the GPSMs from the four standard glycoproteins. An accuracy of 0.66 with an F1 score of 0.63 on criterion 1 and an accuracy of 0.85 with 0.81 F1 score on criterion 2 were acquired, showing divergences from part of the feature prediction. The ROC curve is also calculated in Fig. 2b with an AUC value of 0.84, revealing a considerable performance on the majority of the motif feature prediction.

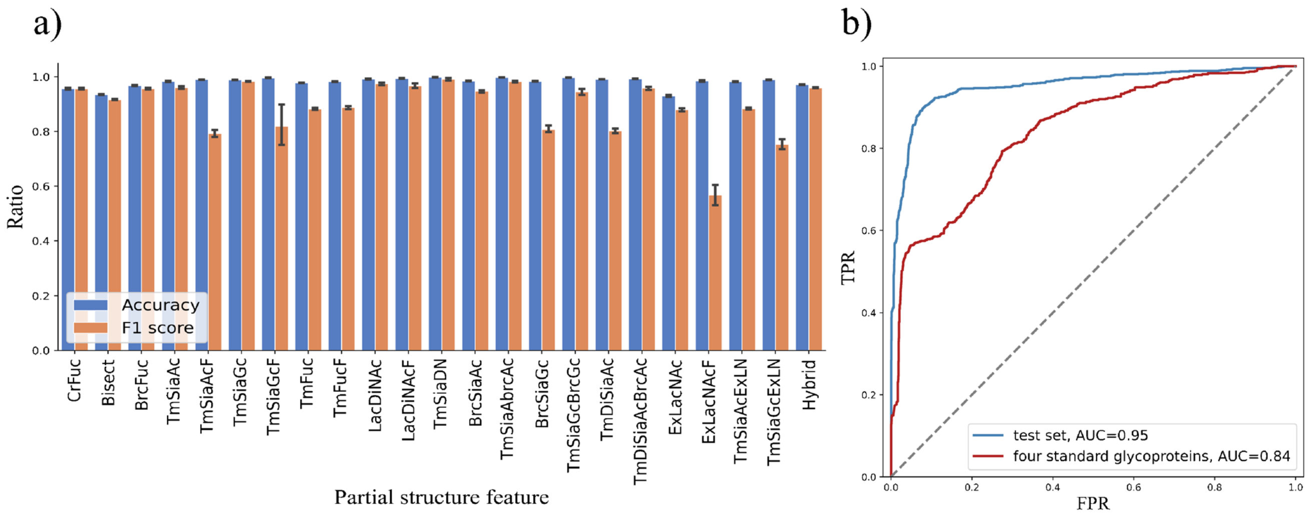


Fig. 2 Evaluation of the model performance: **a** the accuracies and F1 scores of all 23 motif features at the single feature level in the *K*-fold circulation; **b** the ROC curves of the model evaluated on the test split dataset and four standard glycoproteins

Considering the inevitable false positive glycan composition identifications from the adopted existing tools, especially the mass substitutions caused by sulfate (e.g., one sulfate and two HexNAc have equivalent mass to three Hex), we adopted GlycReSoft tool [37, 38], which was capable of identifying sulfated N-glycopeptides in LC–MS/MS data, to estimate this false positive in the train/test split GPSMs as well as the subsequent structure diagnosis step. The original search results of the mouse five tissues are deposited in Supplementary Material S1. The false positive criterion was determined as a GPMS with the same peptide backbone but different glycan composition reported in GlycReSoft. As a result, 49 out of 13,792 GPSMs (0.36%) in the train set split and 8 out of 3703 in the test set split (0.22%) were detected as sulfated mass substitution false positives (Supplementary Fig. S2a and Supplementary Table S3) with a few GPSMs detected as totally different glycopeptides. The total accuracy of the neural network model for predicting feature motifs reduced from 0.8022 to 0.8011 (different GP identification)/0.8018 (only glycan mass substitution), F1 score from 0.7635 to 0.7621/0.7628, and ROC-AUC value from 0.9517 to 0.9462/0.9473, when considering the false positives. In the structure diagnosis step, the false positives took the similar proportions within a low rate (0.097 ~ 0.47%, Supplementary Fig. S2b). The results exhibited a low impact but still these false positives remained inevitable and became a tough drawback of the FISD strategy, leaving a key factor that influences the model performance.

Evaluation of the structure diagnosis step The performance of the structure diagnosis strategy based on the identified motif features without the participation of the deep learning model was first evaluated on the GPSMs from the four standard glycoproteins by taking the motif features contained in the reported N-glycan structures of each spectrum as true labels. All candidate N-glycan structures were generated using the structure definition rules, and then, their corresponding theoretical structure-diagnostic ions were computed and annotated in MS/MS spectra to obtain the final structure identification. The identified N-glycan structures were compared to the structures reported by StrucGP (Fig. 3a left, Supplementary Table S4a) and would result in three conditions: (1) the structure diagnosis strategy reported results covered the StrucGP results; (2) different reports but the structure identified by StrucGP was computed to have no theoretical structure-diagnostic ions; (3) entirely different reports. The comparison result exhibited that about 2/3 of the reported glycan matches were identical or theoretically not conflicted, but there were still a large proportion that the structure diagnosis strategy diverged. Considering that the structure diagnosis step was distinguishing a deeper level N-glycan isomers, the reported structures with annotated structure-diagnostic ions still remained competitive as shown in several spectrum annotation results (Fig. 4a) and more annotated spectra are presented in Supplementary Fig. S3. Comparison results that exhibited the outperformance of FISD strategy are also displayed in Supplementary Figs. S4–S6, including feature motif distinguishment and structure diagnosis annotations. Moreover, there were parts of the GPSMs identified to contain multiple N-glycan structures, proving it capable of assigning isomers in a single MS/MS spectrum.

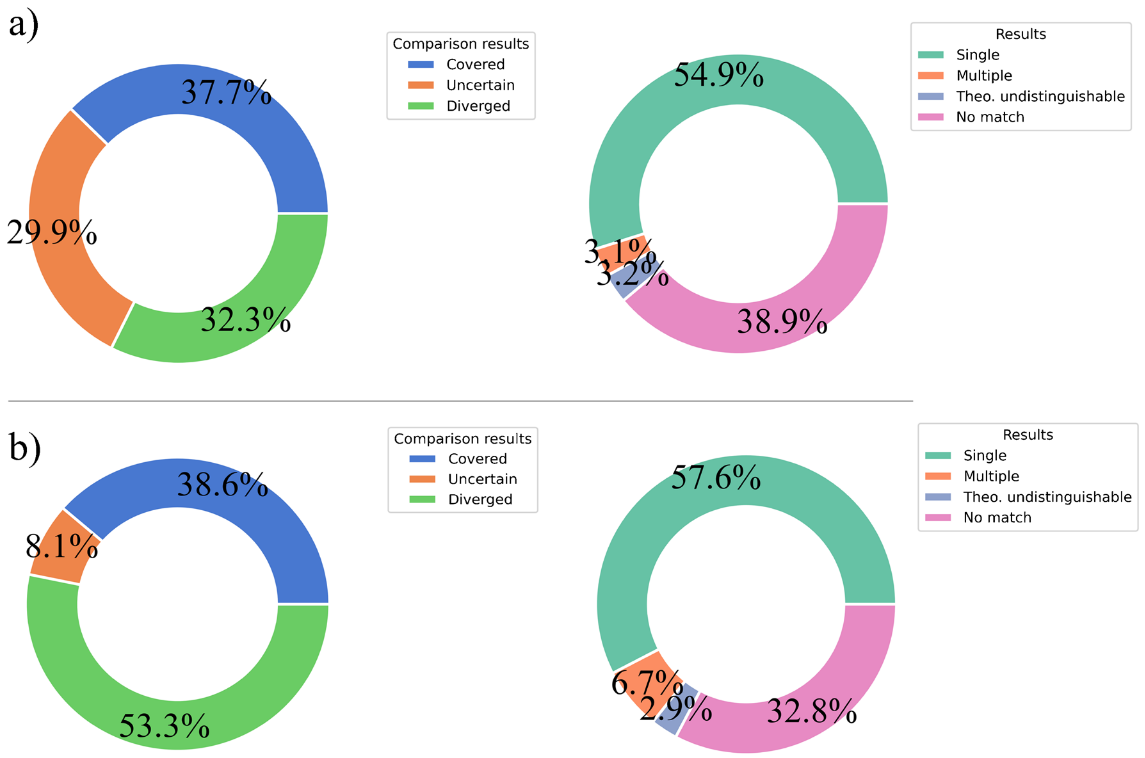
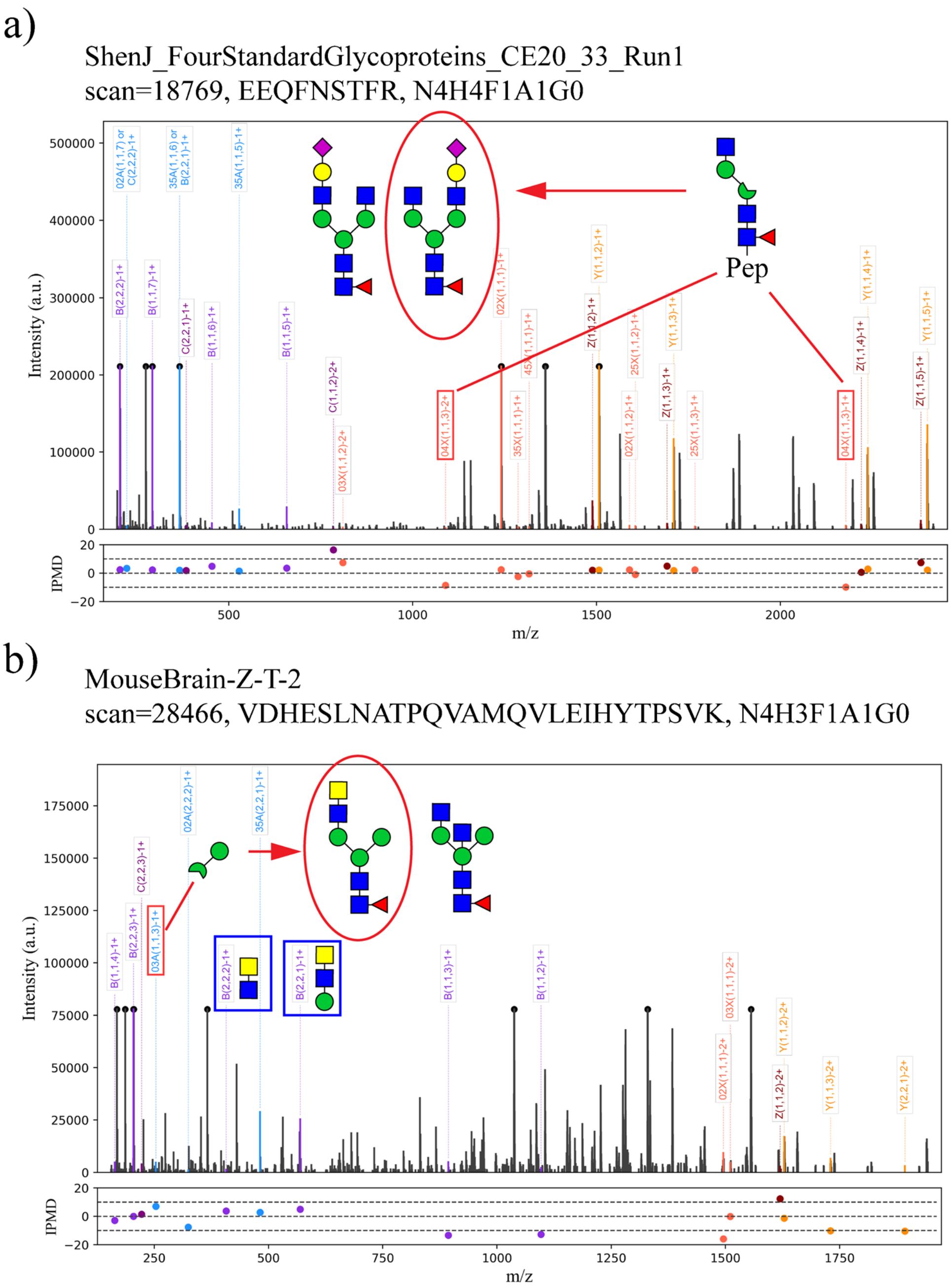


Fig. 3 Comparison of the N-glycan structure reported by structure diagnosis step with the existing results, where the "Uncertain" label stands for the spectrum in which structures reported by StrucGP were evaluated to have no theoretical structure-diagnostic ions: (**a left**) the comparison result on the dataset of four standard glycoproteins without the participation of the deep learning model; (**a right**) the overall statistics of the structure assignment conducted by the structure diagnosis step on the GPSMs from four standard glycoproteins, where the label "Single" stood for the GPSMs finally assigned to one N-glycan structure, the label "Multiple" stood for the GPSMs assigned to more than one structure, the "Theo. undistinguishable" label indicated GPSMs with all candidate N-glycans producing no theoretical structure-diagnostic ions; (**b left**) the comparison result on the dataset of the test split set of the predicted feature vectors using the deep learning model; (**b right**) overall statistics of the structure assignment on the test split dataset

Evaluation of the entire protocol of the FISD strategy The trained model on the 80% GPSMs from the five mouse tissues was temporarily used for the feature-induced step to conduct a comparison identification on the GPSMs of the test set. The predicted motif features and reported N-glycan structures were compared to those reported by StrucGP in the test set (Fig. 3b, Supplementary Table S4b). The result suggested that it gained stronger divergence due to the approximate 0.8 accuracy of motif feature prediction at the total feature level. Differentially identified motif features would lead to distinguishment of differentially grouped candidate N-glycan isomers. However, considering the depth of the isomerism, the FISD strategy still provided remarkable feature identifications with evident spectrum signals (Fig. 4b). There were also some of the GPSMs predicted to be containing more than one N-glycan structure at the single spectrum level, indicating that the FISD strategy was able to accomplish multiple N-glycan structure assignment at an ultra-deep level.

The trained model on the total GPSMs from the five mouse tissues was adopted as the final model for the feature-induced step. The FISD protocol was evaluated on a total of 9689 GPSMs from the five mouse tissues with identical peptide backbone and monosaccharide composition identified by both pGlyco3 and MSFragger-Glyco but not StrucGP (Supplementary Table S4c). The results showed that a total of 5344 MS/MS spectra were identified with singly assigned N-glycan structures while 357 spectra were identified with multiply assigned structures at an ultra-deep isomeric distinguishment level with matched structure-diagnostic ions (Fig. 5a), revealing new N-glycan structures that were not previously reported; and the FISD strategy was qualified for identifying new N-glycan structures at a deep distinguishment level within more generalized conditions.

Utilization of explainable learning to discover new fragmentation regularities and new significant patterns related to featured N-glycan motifs in MS/MS spectra The MLP



ShenJ_FourStandardGlycoproteins_CE20_33_Run1
scan=18769, EEQFNSTFR, N4H4F1A1G0

MouseBrain-Z-T-2
scan=28466, VDHESLNATPQVAMQVLEIHYTPSVK, N4H3F1A1G0

Fig. 4 Annotation of the identified N-glycan structures in single MS/MS spectra. **a** Using only structure diagnosis strategy without the prediction of motif features. The feature vector was calculated from the structure reported by StrucGP and paired "mirror" structures were identified respectively by StrucGP and the structure diagnosis step. The latter provided with key matching of structure-diagnostic ion 04X(1,1,3) to support the deeper level identification. **b** Intact FISD, including deep learning prediction of the feature vector, was applied to the test split dataset from the five mouse tissues. Note that the predicted features (containing a LacDiNAc feature) were different from the true labels (containing a Bisect feature), revealing the competitive predictive results in the diverged reports. The nomenclature of the fragment ions was adapted from our previous work [30], with the 3 numbers listed in the bracket respectively standing for the branch depth, branch index in the branch depth, and the monosaccharide index in the branch. The "right-most" antenna is defined as the one-depth branch

model was capable of explainable learning producing the contributions and significances of the input data to the predicted probabilities of N-glycan motif features. With the assistance of decoding processes in the CAE model and manual checking, a total of three infrequently identified motif features including fucosylated LacNAc with NeuAc (abbreviated as TmSiaAcF), NeuGc (TmSiaGcF), and fucose as terminal monosaccharide (TmFucF) were targeted for discovery of new featured fragmentation patterns in the 9689 GPSMs containing corresponding identified motif features which were used for assigning new N-glycan structures from the five mouse tissues (Fig. 5b, Supplementary Table S5). The fragment composed of HexNAc(1)Hex(2)Fuc(1)-H_2O with $m/z = 674.25$ was marked as significant to the TmSiaAcF and TmFucF motifs, while fragment composed of HexNAc(1)Hex(2)NeuGc(1)-H_2O with $m/z = 835.28$ was marked significant to TmSiaGcF. Both fragments were not yet taken into consideration as the marker signals for these three motif features, suggesting the usage of explainable methods in the feature-induced step of the FISD strategy was capable of discovering new features or significantly marked fragmentation patterns and reporting new potential marker signals for N-glycan structures.

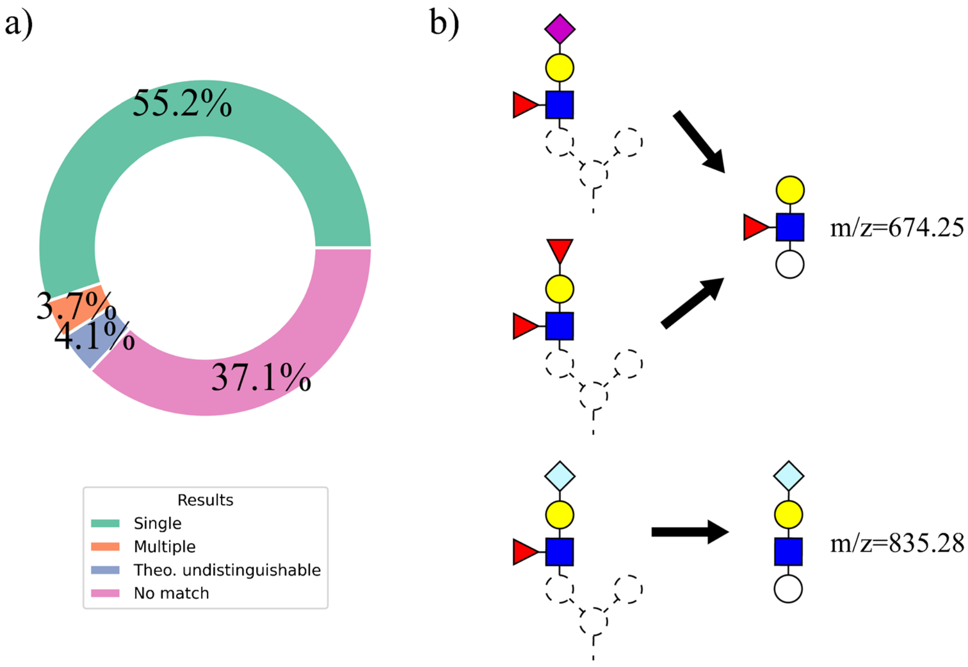


Fig. 5 Applications of the FISD strategy: **a** the identification statistics on the GPSMs from the mouse five tissues co-reported by pGlyco3 and MSFragger-Glyco, in which no specific N-glycan structure was previously assigned; **b** the marked significant fragment patterns reported by the explainable learning on the trained CAE and MLP model, which were potentially stable features for the three infrequent motifs

Discussion

The FISD strategy accomplished the identification of N-glycan structures at an ultra-deep level considering the branch location isomers and an extreme large amount of branch structure combinations in the N-glycan structure rules. By introducing the motif feature-induced strategy, our previously reported structure diagnosis method could finally be adapted to large N-glycan systems. The deep learning method for predicting motif features also performed well at the single feature level with F1 scores of most features > 0.85 and reached considerable accuracy of about 0.8 at the total feature level in a single MS/MS spectrum. Incorrect identified compositions reported by existing tools would introduce false positives to the model prediction. Though it was proved to exhibit a low impact under our experiment conditions in both the proportion rates and accuracies, it still remained an inevitable and non-negligible drawback. More fragment types were taken into consideration by this model especially the internal fragments of BY/BZ/CY/YY/YZ/ZZ types, broadening the theoretical patterns used in the N-glycan identification. Fragment correlations and abundance signals were also embedded into the neural network, boosting the utilization of indirect and hidden patterns in the MS/MS spectra; and the remarkable generalization ability of the neural network is to further promote the identification. Although several results reported by the FISD strategy diverged from existing reports both at the motif features and intact N-glycan structures, our results provided competitive proofs in the experimental MS/MS spectra to further support our identifications of features and intact structures, suggesting a deeper structure-level vision of the intact N-glycopeptide spectra. Finally, in the application of the strategy to the GPSMs from the five mouse tissues without previously reported N-glycan structures, 5701 spectra were assigned to specific deep-level structures, making an outstanding supplement of deep N-glycan structure identification to the dataset. By utilizing an explainable learning algorithm, the FISD strategy obtained the ability of discovering new fragmentation patterns and new significant ions related to featured N-glycan motifs by excavating the prominently contributed elements in the neural network.

There are still limits and drawbacks in the current FISD strategy. (1) The neural network model is under inadequate optimization. More confidently N-glycan structure-level characterized GPSMs are needed to train the model to “learn” more about each N-glycan feature. Also, many hyperparameters still remain to be optimized. (2) There is a lack of methodology for verifying these deep structure-level glycan identifications in a large-scale glycoproteomics context regardless of MS/MS spectrum interpretation. Potential strategies such as conducting accompanied glycomic investigation and using deep learning-based glycan structure identification tool (e.g., CandyCrunch [39]) might be feasible. Also checking the reported structures with glycan-mediated interactions or biological functions resources (e.g., works of Daniel Bojar et al. [40]). However, these two strategies still rely on spectrum interpretation or lack hard evidences for detailed glycan structures. (3) The performance of the integrated neural network model relies deeply on the quality of the GPSMs reported by existing tools. Any false positive glycan composition identification especially mass substitutions will introduce false labels to the training data, causing inevitable false positive that inherited from the existing tools. Moreover, the feature-blocking method we adopt in the embedding step also suffers from these introduced false positives. Though we have conducted estimation to show its experimentally low impact, it still remains as an inevitable issue in the model performance and accuracy control. (4) The strategy lacks a method to evaluate the glycan-level FDR at the deep level such as evaluating branch location isomers. Though the glycan-level FDR has been controlled by the tools that provide with the identified glycan composition, the FISD strategy still needs to make full use of the structure-diagnostic ions to estimate a deeper level FDR and search for new features at the deep level as the prediction model does.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-024-05505-4.

Author contribution ZT conceived the study, SQ carried out the study and wrote the draft, and ZT revised and finalized the manuscript.

Funding This research was financially supported by the National Natural Science Foundation of China (22074105, 21775110) and the Shanghai Science and Technology Commission (14DZ2261100).

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Reily C, et al. Glycosylation in health and disease. Nat Rev Nephrol. 2019;15(6):346–66.
2. Ohtsubo K, Marth JD. Glycosylation in cellular mechanisms of health and disease. Cell. 2006;126(5):855–67.
3. Grunwald P. Glycosylation in health and disease. Carbohydrate-Modifying Biocatalysts. 2012;397–476. https://doi.org/10.1201/b11409.
4. Rasheduzzaman M, et al. Protein glycosylation in head and neck cancers: From diagnosis to treatment. Biochim Biophys Acta-Rev Cancer. 2020;1874(2):188422.
5. Shen H, Lee CY, Chen CH. Protein glycosylation as biomarkers in gynecologic cancers. Diagnostics. 2022;12(12):3177.
6. Marsico G, et al. Glycosylation and integrin regulation in cancer. Trends in Cancer. 2018;4(8):537–52.
7. Chen ZW, Huang JF, Li LJ. Recent advances in mass spectrometry (MS)-based glycoproteomics in complex biological samples. Trac-Trends Anal Chem. 2019;118:880–92.
8. Wang Y, Xiao KJ, Tian ZX. Quantitative N-glycoproteomics using stable isotopic diethyl labeling. Talanta. 2020;219:121359.
9. Thaysen-Andersen M, Packer NH, Schulz BL. Maturing glycoproteomics technologies provide unique structural insights into the glycoproteome and its regulation in health and disease. Mol Cell Proteomics. 2016;15(6):1773–90.
10. Rudd PM, Dwek RA. Glycosylation: Heterogeneity and the 3D structure of proteins. Crit Rev Biochem Mol Biol. 1997;32(1):1–100.
11. Medzihradszky KF, et al. Intact N- and O-linked glycopeptide identification from HCD Data using byonic. Mol Cell Proteomics. 2014;13(8):S36–S36.
12. Bollineni RC, et al. Large-scale intact glycopeptide identification by Mascot database search. Sci Rep. 2018;8:2117.
13. Chandler KB, et al. Exploring site-specific N-Glycosylation microheterogeneity of haptoglobin using glycopeptide CID tandem mass spectra and glycan database search. J Proteome Res. 2013;12(8):3652–66.
14. Eshghi ST, et al. GPQuest: A spectral library matching algorithm for site-specific assignment of tandem mass spectra to intact N-glycopeptides. Anal Chem. 2015;87(10):5181–8.
15. Liu G, et al. A Comprehensive, Open-source platform for mass spectrometry-based glycoproteomics data analysis. Mol Cell Proteomics. 2017;16(11):2032–47.
16. Zeng WF, et al. Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3. Nat Methods. 2021;18(12):1515–23.
17. Polasky DA, et al. Fast and comprehensive and glycoproteomics analysis with MSFragger-Glyco. Nat Methods. 2020;17(11):1125–32.
18. Lee HK, et al. Selective identification of a-Galactosyl epitopes in glycoproteins using characteristic fragment ions from higher-energy collisional dissociation. Anal Chem. 2020;92(19):13144–54.
19. Dang LY, et al. Recognition of bisecting glycans on intact glycopeptides by two characteristic ions in tandem mass spectra. Anal Chem. 2019;91(9):5478–82.
20. Li J, et al. Site-specific glycoproteomic analysis reveals upregulated sialylation and core fucosylation during transient regeneration loss in neonatal mouse hearts. J Proteome Res. 2020;19(8):3191–200.
21. Acs A, et al. Distinguishing core and antenna fucosylated glycopeptides based on low-energy tandem mass spectra. Anal Chem. 2018;90(21):12776–82.
22. Shen JC, et al. StrucGP: de novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. Nat Methods. 2021;18(8):921–9.
23. Sun WP, et al. Glycopeptide database search and de novo sequencing with PEAKS GlycanFinder enable highly sensitive glycoproteomics. Nat Commun. 2023;14(1):4046.
24. Xiao KJ, Tian ZX. GPSeeker enables quantitative structural N-Glycoproteomics for site- and structure-specific characterization of differentially expressed N-Glycosylation in hepatocellular carcinoma. J Proteome Res. 2019;18(7):2885–95.
25. Perez-Riverol Y, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47(D1):D442–50.
26. Deutsch EW, et al. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. 2023;51(D1):D1539–48.
27. Liu MQ, et al. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. Nat Commun. 2017;8:438.
28. Adusumilli R, Mallick P. Data conversion with ProteoWizard msConvert. Proteomics: Methods Protoc. 2017;1550:339–68.
29. Chambers MC, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol. 2012;30(10):918–20.
30. Xiao KJ, et al. Large-scale identification and visualization of N-glycans with primary structures using GlySeeker. Rapid Commun Mass Spectrom. 2018;32(2):142–8.
31. Kornfeld R, Kornfeld S. Assembly of asparagine-linked oligosaccharides. Annu Rev Biochem. 1985;54:631–64.
32. Lowe JB, Marth JD. A genetic approach to mammalian glycan function. Annu Rev Biochem. 2003;72:643–91.

33. Kronewitter SR, et al. The development of retrosynthetic glycan libraries to profile and classify the human serum N-linked glycome. Proteomics. 2009;9(11):2986–94.
34. Schjoldager KT, et al. Global view of human protein glycosylation pathways and functions. Nat Rev Mol Cell Biol. 2020;21(12):729–49.
35. Bagdonaite I, et al. Glycoproteomics. Nat Rev Methods Prim. 2022;2(1):48.
36. Lowe JB. Glycosylation, immunity, and autoimmunity. Cell. 2001;104(6):809–12.
37. Klein J, Carvalho L, Zaia J. Application of network smoothing to glycan LC-MS profiling. Bioinformatics. 2018;34(20):3511–8.
38. Maxwell E, et al. GlycReSoft: A software package for automated recognition of glycans from LC/MS data. PLoS ONE. 2012;7(9):e45474.
39. Urban J, et al. Predicting glycan structure from tandem mass spectrometry via deep learning. Nat Methods. 2024;21(7):1206–15.
40. Bojar D, et al. Deep-learning resources for studying glycan-mediated host-microbe interactions. Cell Host Microbe. 2021;29(1):132-144.e3.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.