Western Graduate&PostdoctoralStudies

Western University
Scholarship@Western

Electronic Thesis and Dissertation Repository

7-13-2023 10:30 AM

Decoy-Target Database Strategy and False Discovery Rate Analysis for Glycan Identification

Xiaoou Li,

Supervisor: Kaizhong Zhang, *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Computer Science
© Xiaoou Li 2023

Follow this and additional works at: https://ir.lib.uwo.ca/etd

 Part of the Bioinformatics Commons, and the Theory and Algorithms Commons

Recommended Citation

Li, Xiaoou, "Decoy-Target Database Strategy and False Discovery Rate Analysis for Glycan Identification" (2023). *Electronic Thesis and Dissertation Repository*. 9581.
https://ir.lib.uwo.ca/etd/9581

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.

Abstract

In recent years, the technology of glycopeptide sequencing through MS/MS mass spectrometry data has achieved remarkable progress. Various software tools have been developed and widely used for protein identification. Estimation of false discovery rate (FDR) has become an essential method for evaluating the performance of glycopeptide scoring algorithms. The target-decoy strategy, which involves constructing decoy databases, is currently the most popular utilized method for FDR calculation. In this study, we applied various decoy construction algorithms to generate decoy glycan databases and proposed a novel approach to calculate the FDR by using the EM algorithm and mixture model.

Keywords

Tandem mass spectrometry, false discovery rate, target-decoy search strategy

Summary for Lay Audience

In recent years, an increasing number of glycopeptide identification software has been developed, capable of scoring glycopeptides and identifying tandem mass spectrometry data. However, due to the potential mistakes in the results, false discovery rate (FDR) estimation plays a key role in evaluating the confidence of correctness. Applying the decoy-target approach is one of the most effective methods for calculating FDR, which requires building a decoy database. In this study, we explored a novel method for generating decoy databases based on the probability of glycan composition in the target database, and then compared it with other decoy construction methods. Meanwhile, since the distribution of target matches could be a mixture of the correct matches and incorrect matches, we created a new FDR estimation approach by using the EM algorithm with a mixture model.

Acknowledgments

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Professor Zhang Kaizhong. Without his profound expertise, rich research experience, and patient guidance, I would not have been able to complete my thesis and my master's research. His care and encouragement have continuously supported my academic and personal growth, and I feel fortunate and honored to have completed my studies under his supervision.

I would also like to thank Weiping Sun from Bioinformatics Solutions Inc and Saito Shun for their help and support in using the GlycanFinder software, as well as the staff members from the Computer Science Department at the University of Western for their assistance.

Finally, I would like to express my gratitude to my family, friends, and Snow for their encouragement and support during difficult times and when I faced obstacles.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

1.1 Glycoproteomic

Glycoproteomics is a rapidly advancing field that focuses on the comprehensive identification and characterization of glycosylation on proteins at the proteome level. Glycosylation, a prevalent and vital post-translational modification (PTM), involves the covalent attachment of glycans to proteins, which modulates the glycan structures observed on individual proteins and regulates their functions within the cell. This ubiquitous PTM not only plays a critical role in elucidating cell development, intercellular communication, and interactions but also holds significant potential for the treatment of various diseases such as Alzheimer's disease and cancer [1, 2].

Protein glycosylation represents a preponderant and essential PTM, arising from the enzymatic activity of glycosyltransferases that orchestrate the formation of glycosidic bonds. The interplay between glycosyltransferases, carbohydrate transporters, and glycosidases intricately fine-tunes the glycan structures observed on individual proteins, thereby modulating their biological activities.

Glycoproteomic analysis performed at the protein level allows for the comprehensive identification and characterization of glycosylation. However, the structural diversity and heterogeneity of glycosylation sites make the analysis of glycosylation more intricate than simpler PTMs. Due to its intricate nature, a single protein may exhibit hundreds of possible glycan attachments, with N-linked and O-linked glycosylation sites being modified by a wide range of different glycans [3]. Consequently, the study of glycosylation necessitates a comprehensive understanding of the intricate interplay between the various factors influencing glycan structures, the deployment of advanced analytical methods and technologies to capture and comprehend these complex structures.

1.1.1 Monosaccharides

Monosaccharides are simple carbohydrate molecules that cannot be broken down by hydrolysis into smaller carbohydrate units. They are often referred to as "sugars" due to their basic structure. Chemically, they are aldehydes or ketones with two or more hydroxyl groups. Generally, monosaccharides have a chemical formula of $C_x(H_2O)_y$.

Monosaccharides are categorized based on three primary characteristics: the position of carbonyl group, the number of carbon atoms, and the chiral nature. For instance, monosaccharides with an aldehyde carbonyl group are called aldoses, while those with a ketone carbonyl group are called ketoses. Monosaccharides can also be classified based on the number of carbon atoms, such as triose (3), tetrose (4), pentose (5), hexose (6), heptose (7). In addition, there are several minor monosaccharides, including mannose, galactose, xylose, and arabinose [4, 5, 6].

1.1.2 N-linked/O-linked glycans:

Glycans are compounds consisting of many monosaccharides linked glycosidically, and they are integral components of numerous biological processes. Among the various types of glycans, N-linked and O-linked glycans are the most extensively studied. N-linked glycans are typically released from glycoproteins through enzymatic digestion, while O-linked glycans are commonly released using chemical methods [7]. The precise characterization of glycans not only enhances our understanding of various cellular processes but also provides insights into the underlying mechanisms governing disease pathogenesis, enabling the identification of novel therapeutic targets.

Typically, N-glycans possess a shared core structure that consists of three mannose residues linked to two GlcNAc residues. This fundamental core structure can be modified by substituting its atoms or molecules with each other to form different derivatives, resulting in the generation of diverse branching patterns and a plethora of linkages [2]. N-glycans can be classified into three distinct groups: high-mannose, complex, and hybrid glycans. The high-mannose N-linked glycans exhibit a core structure that encompasses multiple mannose residues. In contrast, complex N-linked glycans exhibit a core structure decorated with a diverse range of monosaccharides. The hybrid N-linked glycans, as the

name suggests, exhibit a core structure that carries mannose residues on one side and complex residues on the other [8].

In contrast, O-glycosylation lacks a singular consensus sequence, as it encompasses a broad spectrum of glycan types [9, 10]. Notably, there are currently eight characterized core structures for mucin-type O-glycans, which exhibit diverse variations in length and branching antennae [11].

This thesis focuses only on N-linked glycans, and all tests were conducted on N-linked glycan database.

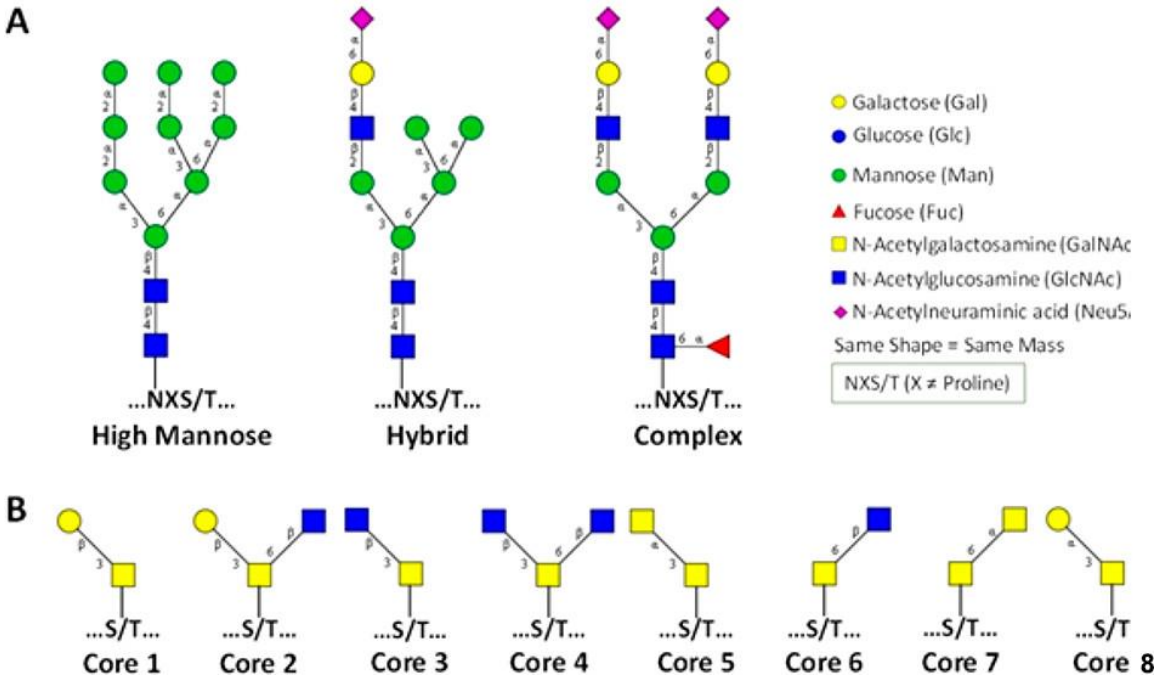


Figure 1 Variation in the *N*- and *O*-glycosylation in proteins. (A) Three types of *N*-glycans. (B) Eight core structures of *O*-glycans. [2]

1.1.3 Format for glycans:

There are different ways to represent glycans, including GlycoCT and linear representation.

GlycoCT is a multi-line format for representing glycan structures and compositions. It was published in 2008 [12]. The format was designed to be easily readable and compressed while also ensuring a single representation for each glycan structure.

GlycoCT consists of two main sections: the entity list and the linkage list. The entity list contains all the residues present in the glycan, with each residue assigned a unique number. The linkage list specifies the connectivity between residues using these numbers as addresses.

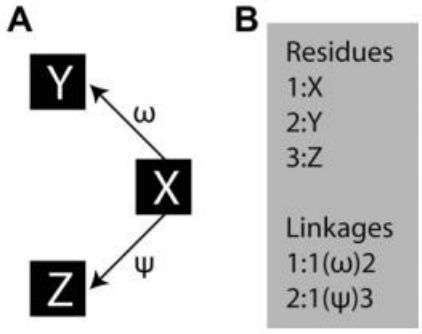


Figure 2 The general idea of GlycoCT [12]

Besides, GlycoCT also includes the repeating unit and underdetermined terminal. The repeating unit is referenced and specified in the REP-section. It can also link to further sections by applying underdetermined units.

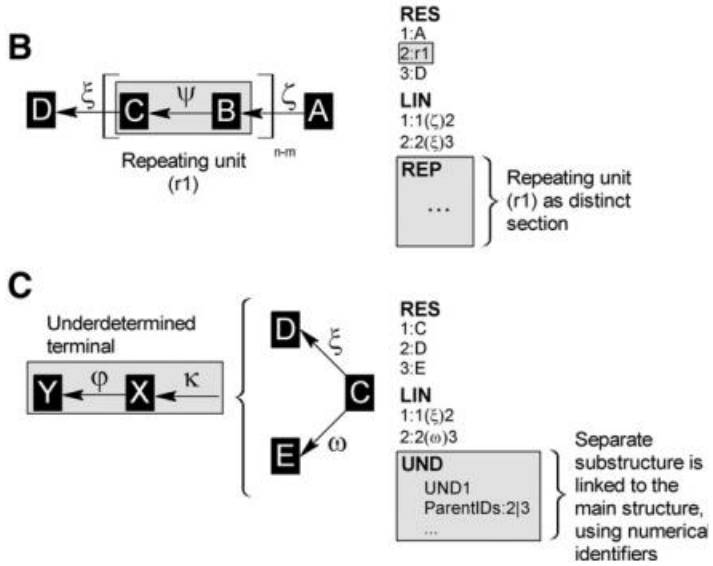


Figure 3 Repeating unit and Underdetermined terminal [12]

GlycoCT employs five attributes in its residue list to represent various features of glycan structures. These attributes include the anomeric carbon configuration, the three-letter code representing the stem type with its configuration, the chain length denoting the number of carbons, the positions involved in ring formation, and additional modifier information.

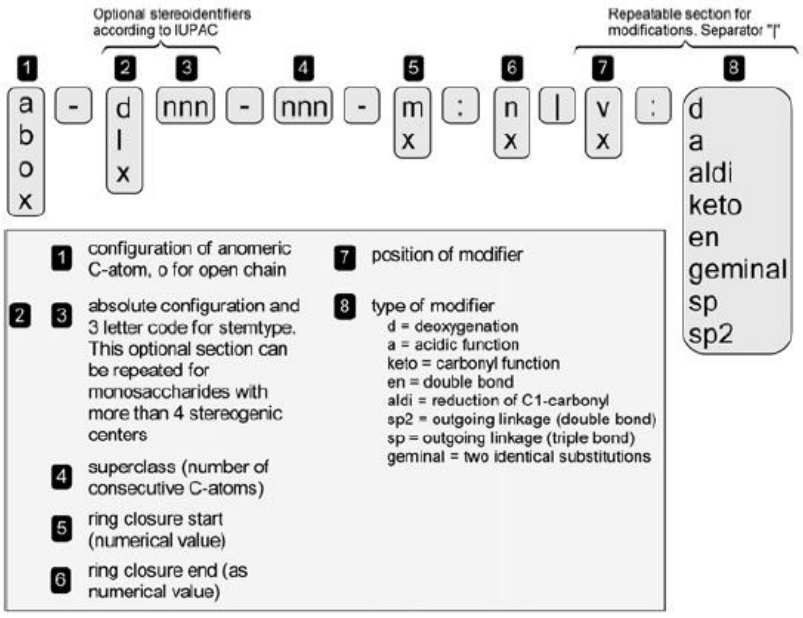


Figure 4 GlycoCT format [12]

On the other hand, in the linkage information, we are able to represent canonical linkage number and residue number, atom replacement identifier, attachment position of both parent side and child side.

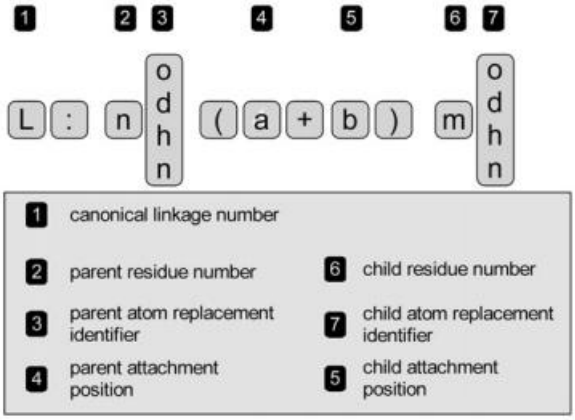


Figure 5 Linkage information [12]

At the same time, glycans can be represented using a linear representation, such as a preorder traversal tree. Therefore, glycans in GlycoCT format can be transformed into linear notation.

It is noteworthy that in this study, information such as residue number and atom replacement identifier were not considered, as the glycan score was calculated only based on the distance of ion mass lists.

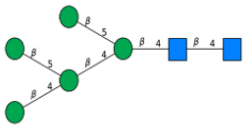
GlycoCT	Glycan structure	VLR tree
RES 1b:b-dglc-HEX-1:5 2s:n-acetyl 3b:b-dglc-HEX-1:5 4s:n-acetyl 5b:b-dman-HEX-1:5 6b:b-dman-HEX-1:5 7b:b-dman-HEX-1:5 8b:b-dman-HEX-1:5 9b:b-dman-HEX-1:5 LIN 1:1d(2+1)2n 2:1o(4+1)3d 3:3d(2+1)4n 4:3o(4+1)5d 5:5o(4+1)6d 6:6o(4+1)7d 7:6o(5+1)8d 8:5o(5+1)9d		HexNAc(HexNAc(Hex(Hex)(Hex(Hex(Hex)(Hex)))))

Figure 6 Glycan example

1.2 Mass spectrometry

1.2.1 Mass spectrometry

Mass spectrometry is an analytical technique for determining the mass-to-charge ratio (m/z) of molecules in a sample. This powerful tool is widely used in various applications, including identification of unknown compounds based on the molecular weight, quantification of known compounds, and elucidation of molecular structure and chemical properties.

Modern mass spectrometers are mainly composed of three parts: ionization source, mass analyzer, and ion detection system [13]. The mass spectrometer initially produces ions from a liquid sample, once ionized, the ions can be sorted and separated in the mass analyzer based on its mass-to-charge ratio (m/z), where m denotes the relative molecular weight of the ion in Daltons and z represents the absolute value of its charge number in electrons. Then the abundance of ions at each m/z value is detected and converted into an electrical signal. The detector subsequently processes the electrical signal and transmits it to a computer for analysis. The output of the mass spectrometer is typically presented as a histogram, where the X-axis denotes the m/z value, and the Y-axis represents the intensity at each m/z value.

1.2.2 Tandem MS

In instrumental analysis, tandem mass spectrometry (MS/MS or MS^2) is a technique that involves coupling two or more mass analyzers together, along with an additional reaction step, to separate and detect more key parameters.

In MS/MS, the molecules in a sample are first ionized, and the ions are then separated by m/z using the first spectrometer (MS1). Next, ions with a specific m/z ratio are selected and induced to fragment into smaller ions. These fragment ions are then introduced into the second mass spectrometer (MS2), which further separates and detects them based on their m/z ratio. This fragmentation step allows for the identification and

separation of ions with similar m/z ratios that may not be easily distinguished in single conventional mass spectrometers.

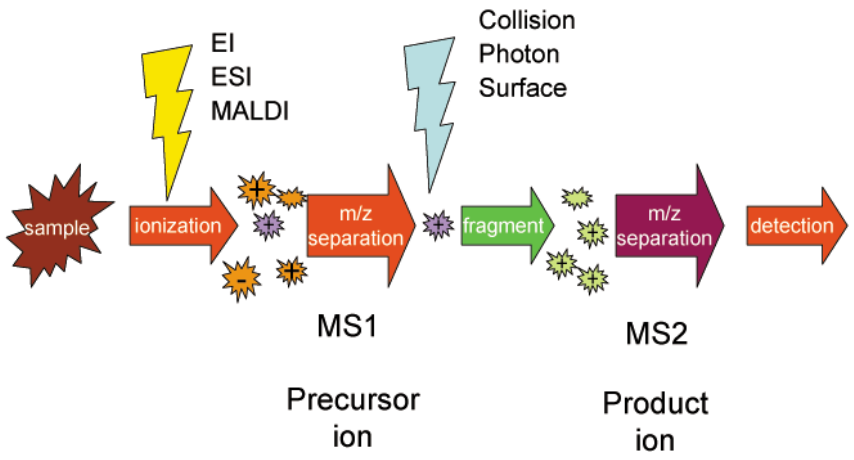


Figure 7 MS/MS

1.2.3 Dissociation methods:

Nowadays, various MS peptide fragmentation methods have been developed, and they have two major classes based on energy deposition: vibration-based and electron-based methods. Vibration-based methods include low-energy collision-induced dissociation (CID), high-energy collision dissociation (HCD), and infrared multiphoton dissociation (IRMPD). Electron-based methods include electron capture dissociation (ECD), electron transfer dissociation (ETD), and ultraviolet photodissociation (UVPD) [14].

CID is the most commonly used fragmentation technique, it typically results in only a single fragmentation event for precursor ions. Alternatively, HCD induces multiple collisions with a collision gas occurring under high-pressure conditions. This leads to the activation of ions in a multi-step process, resulting in more complex spectra compared to CID. Both of methods tend to produce spectra with predominantly b-type and y-type amino acid sequence-specific ions.

ECD and ETD are two techniques used to fragment molecular ions in gas phase during mass spectrometry analysis. In ECD, low energy electrons are directly introduced to trapped ions, causing simultaneous fragmentation and generating more complete ion

series. Similarly, ETD also involves transferring electrons to peptides, resulting in random cleavage along the peptide backbone while leaving side chains intact, and both methods generate c and z type ions.

Additionally, hybrid fragmentation, which involves the combination of multiple dissociation methods, has been utilized to generate a wider range of ion types. This approach enhances the characterization of glycopeptides by enabling the identification of more complex structural features [15][16].

1.2.4 Peptide identification strategy:

There are four commonly used strategies for peptide identification: database searching, spectral library searching, tag-assisted searching and de novo sequencing [17]. Among these, database searching is the most utilized method for peptide identification and characterization. This strategy matches experimental MS/MS data with theoretically possible sequences in a reference database, such as UniProtKB [18] and NCBI RefSeq [19]. By calculating the comparison of experimental spectra with theoretical fragment masses, candidate peptides are scored and ranked based on their match to the experimental data, with the highest-scoring peptide reported. Some search engines for database searching, like Mascot [20], SEQUEST [21], and Andromeda [22, 23], have been developed recently. One of the drawbacks of database searching is its heavy dependence on the quality and availability of the reference database, it may not work well when there is no accurate reference database.

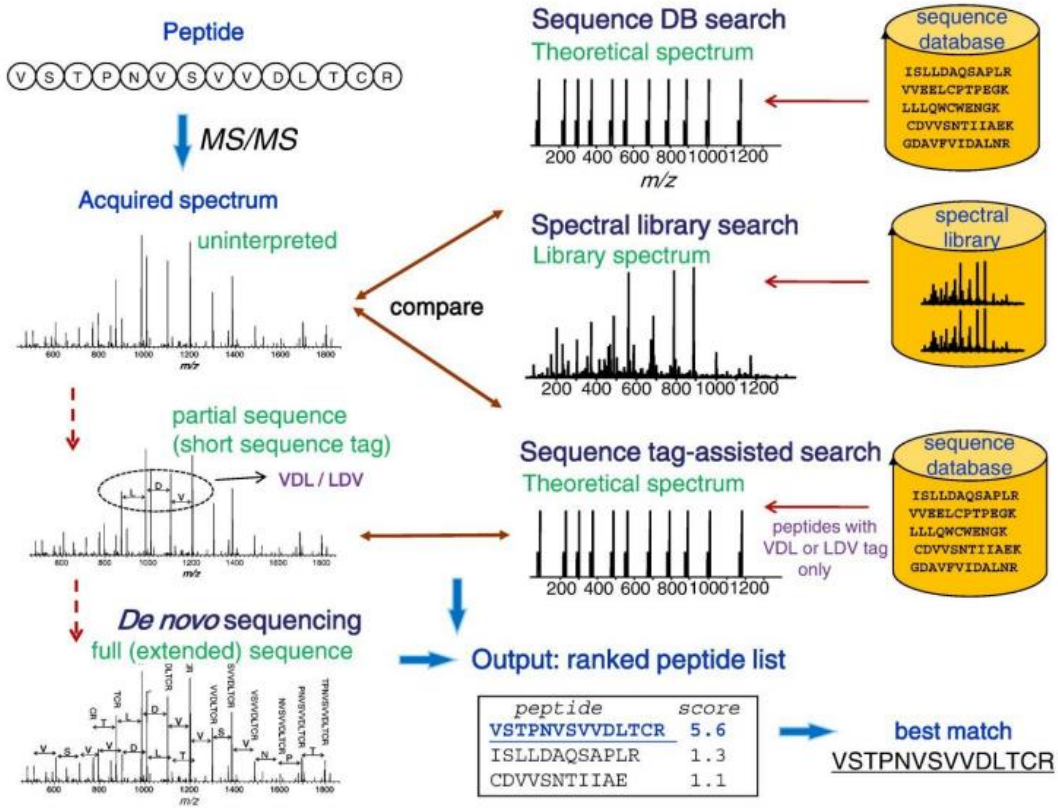


Figure 8 Four commonly used strategies for peptide identification [16]

Spectral library searching is based on MS/MS spectra in a spectral library. Although spectral library searching has a shorter processing time and higher identification rate than database searching, it, like database searching, depends on the availability of a reference database [24]. COSS [25] is one of the most used tools at present.

Tag-assisted protein purification has become a popular approach for academic researchers. The incorporation of purification tags into the protein production process can greatly reduce time and cost. However, the challenges are how to design and implement the tagged fusion proteins [26]. The software JUMP is a Tag-based Database Search Tool [27].

When a suitable database is not available, de novo sequencing is the only option for peptide identification [28]. De novo sequencing can reconstruct the original amino acid sequence from MS/MS spectra, making it possible to identify previously unknown peptide

sequences, peptide homologues, and modifications [29]. There are currently a variety of widely used de novo sequencing software programs, such as PEAKS [30], PepNovo [31], and pNovo3 [32].

1.3 False discovery rates and target-decoy approach

1.3.1 False discovery rate

The database search method is currently one of the most popular and widely used methods. However, the results obtained from this method are not entirely reliable, because not all peptides are present in the reference database, and incorrect candidate peptides may at times be prioritized over the correct sequences. However, high-throughput studies often involve millions of tests which make manual inspection impossible [33]. To address this problem, false discovery rate (FDR) estimation has emerged as an adopted and effective method for controlling error rates in large-scale identification efforts.

1.3.2 Target-decoy search strategy

The target-decoy search strategy has emerged as a standard approach for estimating the FDR [34]. This approach is achieved by constructing decoy peptide sequences that do not exist naturally in the universe. Therefore, if a match is made with the decoy database, it represents a match with an incorrect sequence. This forms the first assumption of the target-decoy strategy, which states that any hits to the decoy database are incorrectly assigned [35]. At the same time, if no match is made with any decoys, then the purpose of using the decoy database is lost, therefore, when constructing the decoy database, it is also important to avoid complete difference from the target database.

Another key assumption of the target-decoy search strategy is that the likelihood of false positive identifications is the same for both the target and decoy databases [36]. The incorrect decoy peptides are designed to closely resemble unknown incorrect peptides that may be present in the target database. Following this, the experimental MS/MS spectra are searched against both the target and decoy databases. Since the peptide sequences in the decoy database are not present in the sample, any PSMs identified against decoy sequences

are deemed incorrect, enabling the estimation of the relative proportion of target and decoy sequences.

1.3.3 Decoy sequence construction

The Target-decoy approach is an effect strategy for estimating the FDR [37]. However, it is important to remember that this strategy is dependent on the quality and completeness of the decoy database, which must be carefully constructed.

In recent years, several methods have been suggested for constructing decoy sequences. One straightforward technique for producing decoy sequences is reversing the target database [36]. This method preserves the features of the original sequence, such as composition and sequence length, and is easy to implement. However, the disadvantage is that the reversing method is less efficient for generating decoy peptides for complex sequences [38]. Another approach for constructing decoy databases is to add random noise to the target spectrum to generate decoy spectra [39].

Although the decoy-target strategy for glycan works similarly to peptides. Unlike the chain-like structure of peptides, the chemical structure of glycans can be abstracted into a tree structure. Saito's [40] stated that the shuffle method is an easy-to-use decoy glycan creating method. This method randomly rearranged the target monosaccharides while preserving the glycan structure and composition. However, we thought that simple shuffling or swapping may not be sufficient for decoy glycan databases. Changing the glycan tree structure should also be considered. Therefore, in this thesis, we will explore a few decoy construction methods by modifying the glycan tree structure.

1.3.4 Separate search and concatenated search

After constructing the decoy database, there are two main types of searching: separate search and concatenated search [41]. In the separate database search method, the target and decoy databases are searched independently, and the best scores from each database are used to identify targets and decoys, respectively. On the other hand, the concatenated search method reports only one target or decoy sequence with the best score for each spectrum. Unlike the separate search method, in concatenated search, there is

competition between the target and decoy sequences in a single search for the highest score. Elias and Gygi [1] recommended using concatenated database searches instead of separate searches because separate searching greatly obstructs the ability to estimate low-scoring correct identifications in the presence of high-scoring incorrect identifications.

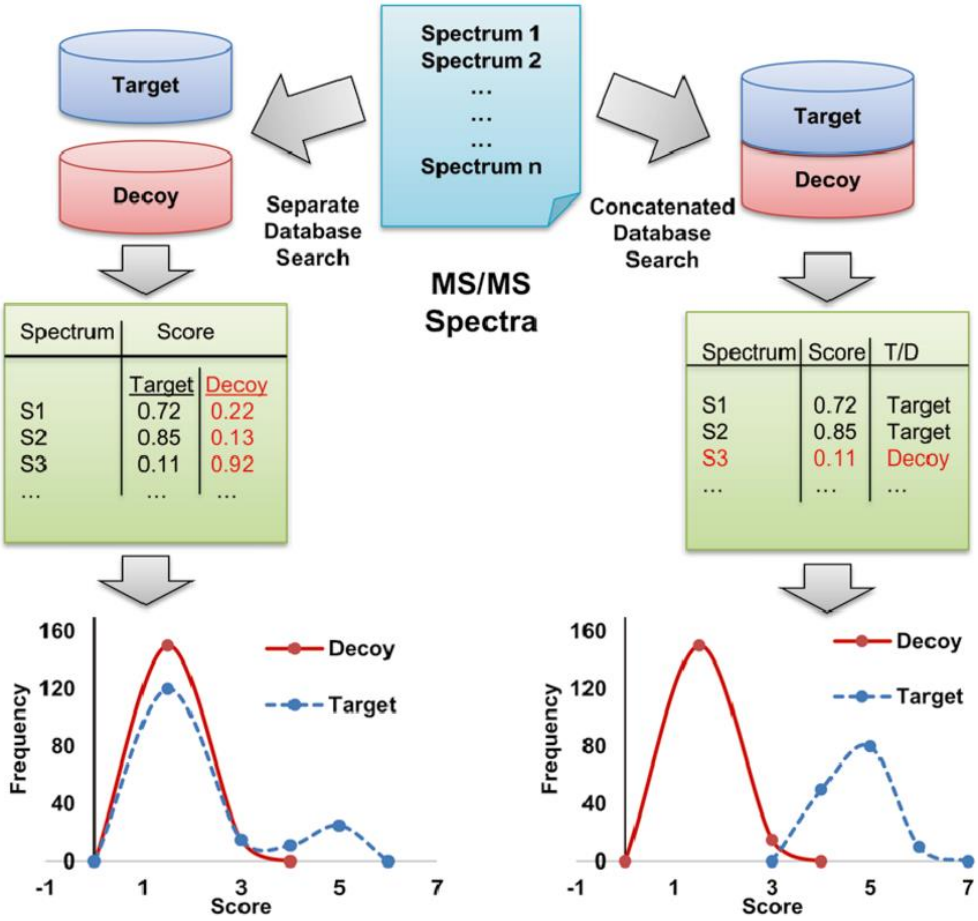


Figure 9 separate database search and concatenated database search [41]

1.3.5 FDR estimation

In the context of target-decoy search strategy, true positives (TP) refer to the number of correct matches above the threshold score, while false positives (FP) indicate the number of incorrect matches above the threshold. True negatives (TN) represent the number of incorrect matches below the threshold, and false negatives (FN) indicate the number of correct matches below the threshold.

In general, the FDR estimates the percentage of incorrect matches above the threshold in the positive matches, and is defined as:

$$FDR = \frac{FP}{TP+FP} \tag{1.1}$$

Since we can't easily get the number of incorrect and correct matches. According to the assumptions of decoy-target strategy, the separate search strategy computes the ratio of the number of decoy matches and target matches above the threshold by the following formula:

$$FDR = \frac{Number\ of\ Decoy\ matches}{Number\ of\ Target\ matches} \tag{1.2}$$

In the concatenated database search, incorrect matches cannot be accurately determined due to the mixture of target and decoy databases. Since we assumed that the likelihood of false positive identifications is the same for both the target and decoy databases, in this situation, true positive matches tend to match targets, while false positive matches are uniformly distributed between targets and decoys. Then, the number of false positives is doubling the number of decoy matches. The FDR in this case is calculated using the formula:

$$FDR = \frac{2*Number\ of\ Decoy\ matches}{Number\ of\ Target\ matches+Number\ of\ Decoy\ matches} \tag{1.3}$$

Since we assumed that the probabilities of incorrect matches were the same in both decoy and target databases. However, in practical applications, this assumption may not always hold true. This can lead to an overestimation or underestimation of FDR. If we can obtain distributions for correct and incorrect matches, then we can calculate the cumulative distribution function (CDF) of the distributions to get a more accurate estimate of the FDR.

$$FDR = \frac{1 - the\ CDF\ of\ incorrect\ glycan\ distribution}{(1 - the\ CDF\ of\ incorrect\ glycan\ distribution) + (1 - the\ CDF\ of\ correct\ glycan\ distribution)} \tag{1.4}$$

1.4 Mixture distribution and EM algorithm

1.4.1 Mixture distribution

In FDR estimation, we aimed to determine the number of incorrect and correct matches that were above the given threshold to calculate the FDR. However, since we could not directly distinguish between correct and incorrect matches, according to the assumptions that decoys represented incorrect matches and that the false positives were the same in both the decoy and target databases, we estimated the FDR by computing the number of target and decoy matches that were above the threshold.

In this article, we proposed a novel algorithm for FDR calculation. We knew that the distribution of glycan scores above the threshold included both incorrect and correct matches, we considered it as a mixture model at least containing two distributions. By employing the EM (Expectation-Maximization) algorithm, we could calculate the parameters of these distributions and directly computed the FDR.

The use of mixture models has gained increasing attention in statistical and machine learning research. A mixture model is a probabilistic model that represents the distribution as a mixture of two or more component distributions, $P_1(x)$, ..., $P_n(x)$, these component distributions are combined in a weighted manner, π_1, …, π_n and $\sum \pi_i = 1$, where the weights represent the proportion of the population that belongs to each component, then the distribution function, F can be represented as:

$$F(x) = \sum_{i=1}^{n} \pi_i P_i(x) \qquad (1.5)$$

One advantage of mixture models is their ability to capture heterogeneity in the data. This is particularly useful when dealing with data that exhibits multiple clusters. By modeling the data as a mixture of distributions, mixture models can accurately capture the characteristics of each cluster, even if they have different mean, variance, or shape. However, there are also some limitations to the use of mixture models. One limitation is their sensitivity to the initialization of the model parameters. It may lead to multiple local optima, which can result in different solutions depending on the starting point. Another

limitation is the assumption of independence between the components, which may not hold in some cases [42].

1.4.2 EM algorithm (Expectation-maximization)

The EM algorithm (Expectation-maximization) is a widely used approach used to estimate model parameters for probabilistic models with latent variables. It is an iterative algorithm that alternates between two steps: the E-step and the M-step. In the E-step, the algorithm calculates the posterior probabilities of each observation belonging to each component of the mixture distribution, given the current estimate of the parameters [43]. In the M-step, the algorithm updates the model parameters to maximize the expected log-likelihood of the data given the latent variables computed in the E-step [44]. The EM algorithm is guaranteed to increase the likelihood of the data at each iteration and will converge to a local maximum of the likelihood. However, the algorithm may converge to a suboptimal solution if it gets stuck in a local maximum. The advantage of the EM algorithm is its ability to handle missing data. In cases where some of the data is missing, the EM algorithm can be used to estimate the missing values and the model parameters simultaneously [45].

Chapter 2

2 Methods

2.1 Databases

In this work, we used the same mouse brain MS mass spectrometry data as pGlyco3[46]. It was utilized for analysis and testing to improve experimental stringency and accuracy. We analyzed the mouse protein database used by GlycanFinder, which contained 17048 mouse brain sequences [47]. Similarly, the glycan database also employed the same glycan database used by GlycanFinder, which consisted of 7887 structurally distinct basic glycans.

2.2 Notation

It is worth noting that the glycan utilized in this project consisted of only six distinct monosaccharides, including Hex(Galactose, Glucose, Mannose), HexNAc(N-Acetylgalactosamine, N-Acetylglucosamine), NeuAc(N-Acetylneuraminic acid), NeuGc(N-Glycolylneuraminic acid), Fuc(Fucose), and Xyl(Xylose). Based on their chemical formulas, the relative molecular masses of these monosaccharides were determined to be 180.06, 221.09, 309.10, 325.10, 164.06, and 150.05, respectively.

Table 1 Monosaccharide mass

Monosaccharide	Generic term	Mass
Galactose, Glucose, Mannose	Hex	180.06
N-Acetylgalactosamine, N-Acetylglucosamine	HexNAc	221.09
N-Acetylneuraminic acid	NeuAc	309.10
N-Glycolylneuraminic acid	NeuGc	325.10
Fucose	Fuc	164.06
Xylose	Xyl	150.05

Monosaccharide	Symbol
Galactose	
Glucose	
Mannose	
N-Acetylgalactosamine	
N-Acetylglucosamine	
Fucose	
Xylose	
N-Acetylneuraminic acid	
N-Glycolylneuraminic acid	

Figure 10 The symbols of monosaccharide

2.3 Glycan distance score

In the thesis, we decided to use mass list distance instead of tree edit distance, such as Saito [40] or Sun[48], to calculate glycan distance scores. This decision was made because the scoring algorithms in most sequencing software heavily relied on B-ion and Y-ion lists as the final evaluation criterion.

We obtained the mass lists of b-ion and y-ion from the decoy glycan and the original glycan respectively, so that a_y and b_y represent the y-ion lists and a_x, b_x represent the b-ion lists, $|a_y|$ represents the length of y-ion list. Therefore we used the formula where $|a_y - b_y|$for the distance of y-ion which is the number of elements in the difference between two y-ion lists, $|a_x - b_x|$ for the distance of b-ion which is the number of elements in the difference between two b-ion lists, s(a, b) is the similarity of glycan a and b, which equals $\lambda|a_y \cap b_y| + (1-\lambda)|a_x \cap b_x|$and $\lambda = 0.7$.

We chose $f(x) = log_2(x+1)$, the scoring formula is below:

$$\frac{d^s_{L\infty}(\mathrm{a,b})}{f(d^s_{L\infty}(\mathrm{a,b})+\mathrm{s(a,b)})}$$

$$= \frac{\max\{s(a,a)-s(a,b), s(b,b)-s(a,b)\}}{f(\max\{s(a,a), s(b,b)\})}$$

$$= \frac{\max\{|a-b|,|b-a|\}}{f(\max\{|a|,|b|\})}$$

$$= \frac{\max\{\lambda(|a_y-b_y|)+(1-\lambda)(|a_x-b_x|),\lambda(|b_y-a_y|)+(1-\lambda)(|b_x-a_x|)\}}{f(\max\{\lambda(|a_y|)+(1-\lambda)(|a_x|),\lambda(|b_y|)+(1-\lambda)(|b_x|)\})} \quad (2.1)$$

2.4 Reciprocal probability distribution

In this research, we proposed a method for randomly generating a decoy glycan based on a reciprocal probability list. The algorithm is as follows:

Given $P = (p_1, p_2, \dots, p_n)$, we assume that $p_i \neq 0$. If we have $p_i = 0$, then add a very small number Ɛ to each p_i and rescale the summation to 1. In the research, we choose Ɛ=0.5% (0.005).

We want to compute Q = (q_1, q_2, . . . , q_n) such that for any i and j:

$$\frac{p_i}{p_j} = \frac{q_j}{q_i}$$

Then we have for any i and j:

$$p_1 q_1 = p_2 q_2 = \cdots = p_n q_n$$

Let $p_i q_i = c$, then $q_i = \frac{c}{p_i}$

Since we have:

$$\sum_{k=1}^{n} q_k = 1$$

Which is:

$$\sum_{k=1}^{n} \frac{c}{p_k} = 1$$

Therefore, we have:

$$c = \frac{1}{\sum_{k=1}^{n} \frac{1}{p_k}}$$

$$q_i = \frac{1}{p_i \sum_{k=1}^{n} \frac{1}{p_k}}$$

$$Q = \left(\frac{1}{p_1 \sum_{k=1}^{n} \frac{1}{p_k}}, \frac{1}{p_2 \sum_{k=1}^{n} \frac{1}{p_k}}, \ldots, \frac{1}{p_k \sum_{k=1}^{n} \frac{1}{p_k}}\right) \quad (2.2)$$

2.5 Decoy glycan construction

2.5.1 Decoy database generating algorithm

We chose to build a one-to-one decoy database, which means that one target glycan generates one corresponding decoy. First, we applied the decoy glycan generation algorithm to each target glycan (the generation algorithm can be any of the algorithms from the following six methods, specific algorithms will be introduced later). Since we cannot guarantee that the generated decoy does not overlap with any other glycans in the target database or the decoys we have already created, thus, we compared the distance between the decoy we just created with all the target glycans and all the decoys we have generated, by using the distance score algorithm, and selected the minimum score as the distance of decoy to the target database. Finally, we repeated the construction process 30 times and selected the decoy with the highest distance to the target database as the final output decoy.

For each glycan g in target glycan database:

Perform 30 iterations:

New glycan g' = glycan construction method (g)

For every Y-ion and B-ion lists in list l:

Get the minimum distance d according to Equation 2.1

Add Y-ion and B-ion lists of g' into list l

Return the decoy g' which has the maximum value of d among 30 times:

2.5.2 Permutation method

For the first method, the monosaccharides of the target glycan were randomly permuted to generate a new decoy glycan. Since this method only alters the order of the monosaccharides in the original glycan, the topology and monosaccharide composition remain unchanged.

Figure 11 Permutation method

2.5.3 Node transfer method

As a simplification of the chemical structure of glycans, they were represented by a tree structure in which each monosaccharide served as a node, and the glycosidic linkages served as the edges. In the second method, decoys were generated by reconstructing the nodes of the target glycan tree.

In the beginning, a node was randomly selected, and we recorded the number of its children before deleting it. In the case of selecting the root, we reselected a non-root node and swapped two monosaccharides, then we deleted the second selected one. Subsequently, for all the children of this node, we transferred them to the parent node through breadth-first searching until the condition (children number less than 4) was met. Moreover, we randomly selected another node and transferred the corresponding number of children of

the second selected node to the previously deleted node, based on the number of children of the deleted node we recorded before. After, we searched for the second chosen node using breadth-first searching until the condition (children number less than 4) was met, then added the deleted node as its child. Finally, we repeated this transfer process 15 times. This method reconstructs the positions of nodes in the glycan tree structure, resulting in a decoy that differs from the original target's tree structure.

Figure 12 Node transfer method (1 time)

Figure 13 Node transfer method (15 times)

2.5.4 Combination method

This method combined the principles of methods 2.5.2 and 2.5.3. Firstly, we repeated the permutation method 10 times and chose the best one among 10 candidate glycans. Next

the 10 random monosaccharides within the glycan were positionally reconstructed. Finally, we permuted 10 more times again and selected the best one as the final decoy.

Figure 14 Combination method

2.5.5 Random method

In this method, the decoy tree structure is generated entirely at random. Firstly, we obtained the number and types of monosaccharides included in the target glycans and converted them into a list. then we randomly selected the root node from the list of target monosaccharides. Since biological rules dictate that each glycan usually has between 1 and 4 monosaccharide children, thus, for the root node, we randomly selected 1 to 4 children's monosaccharides as the next-level nodes. We removed these children from the list and repeated this process level by level until the list of monosaccharides is empty. The random method preserves only the monosaccharide composition of the original target glycan while completely changing its tree structure and monosaccharide positions.

For i in all glycans:

While monosaccharides list! = null:

Random select the number of children from 1 to 4

For j in children number:

Random select monosaccharide from monosaccharide list

Delete the selected monosaccharide in monosaccharides list

Add the selected monosaccharide in the decoy glycan as a node

Figure 15 Random method

2.5.6 Reciprocal probability based on monosaccharides list method

In the random method, we used a completely random method to generate decoys. For peptides, the most used method to generate decoys was by reversing the sequence. Since glycans have a tree-like structure that cannot be easily reversed, we could analyze and create a table of the compositional distribution of glycans in the target glycan database. By reversing this table, we constructed a decoy database that mirrors the structures in the target database, based on statistical information.

2.5.6.1 Statistical probability results

In the reciprocal probability method, we began by determining the number of children at each level in 7887 glycans, which is the target glycan database. We counted the number of branches at each level and obtained statistical results. The statistical results are as follows:

Table 2 The statistical result of children number

	0 child	1 child	2 children	3 children	4 children
Level 1	0	0.511	0.489	0	0
Level 2	0.328	0.672	0	0	0

Level 3	0	0	0.545	0.455	0
Level 4	0.199	0.171	0.627	0.003	0
Level 5	0.129	0.551	0.319	0	0
Level 6	0.482	0.515	0.003	0	0
Level 7	0.814	0.156	0.003	0	0
Level 8	0.547	0.453	0	0	0
Level 9	0.542	0.446	0.012	0	0
Level 10	0.878	0.122	0	0	0
Level 11	0.552	0.448	0	0	0
Level 12	0.86	0.14	0	0	0
Level 13	0.667	0.333	0	0	0
Level 14	0	1	0	0	0
Level 15	0	1	0	0	0
Level 16	1	0	0	0	0

At each level, we also computed the types of monosaccharides and obtained the following results:

Table 3 The statistical result of monosaccharide types

	Hex	HexNAc	HeuAc	NeuGc	Fuc	Xyl
Level 1	0	1	0	0	0	0
Level 2	0	0.672	0.328	0	0	0
Level 3	1	0	0	0	0	0
Level 4	0.815	0.185	0	0	0	0
Level 5	0.072	0.928	0	0	0	0
Level 6	0.67	0.001	0.324	0	0.003	0.003
Level 7	0.113	0.096	0.015	0	0.388	0.388
Level 8	0.528	0.012	0.137	0	0.162	0.161
Level 9	0.202	0.04	0.012	0	0.373	0.373
Level 10	0.289	0	0.025	0	0.343	0.343
Level 11	0.75	0.062	0	0	0.094	0.094
Level 12	1	0	0	0	0	0
Level 13	0.667	0.333	0	0	0	0
Level 14	1	0	0	0	0	0
Level 15	0	1	0	0	0	0
Level 16	1	0	0	0	0	0

2.5.6.2 Reverse the probability lists

Since not all types of monosaccharides were present in each layer of the experimental data, certain monosaccharide types had a probability of 0 in the results. To

create a smoother probability curve, we modified the probability of these data points from 0 to 0.5%. Subsequently, we applied equation 2.2 for probability to recalculate the statistical results and obtained the final data. We followed the same process to invert the children list.

The reciprocal lists are showed:

Table 4 The reciprocal probability of children number

	0 child	1 child	2 children	3 children	4 children
Level 1	0.33109	0.00329	0.00344	0.33109	0.33109
Level 2	0.00512	0.0025	0.33079	0.33079	0.33079
Level 3	0.33107	0.33107	0.00308	0.00369	0.33107
Level 4	0.00922	0.01073	0.00293	0.61184	0.36527
Level 5	0.01897	0.00444	0.00767	0.48446	0.48446
Level 6	0.00283	0.00265	0.45454	0.26999	0.26999
Level 7	0.00281	0.01467	0.07628	0.45312	0.45312
Level 8	0.00307	0.00307	0.33107	0.33107	0.33107
Level 9	0.00382	0.00464	0.17238	0.40958	0.40958
Level 10	0.0019	0.01365	0.32815	0.32815	0.31815
Level 11	0.00304	0.00375	0.33107	0.33107	0.33107
Level 12	0.00194	0.01192	0.32871	0.32871	0.32871
Level 13	0.00252	0.00504	0.33081	0.33081	0.33081
Level 14	0.24968	0.00127	0.24968	0.24968	0.24968
Level 15	0.24968	0.00127	0.24968	0.24968	0.24968
Level 16	0.00127	0.24968	0.24968	0.24968	0.24968

Table 5 The reciprocal probability of monosaccharide types

	Hex	HexNAc	HeuAc	NeuGc	Fuc	Xyl
Level 1	0.1998	0.00102	0.1998	0.1998	0.1998	0.1998
Level 2	0.24856	0.00189	0.00387	0.24856	0.24856	0.24856
Level 3	0.00102	0.1998	0.1998	0.1998	0.1998	0.1998
Level 4	0.00155	0.00684	0.2479	0.2479	0.2479	0.2479
Level 5	0.01738	0.00135	0.24532	0.24532	0.24532	0.24532
Level 6	0.0008	0.53469	0.00165	0.1064	0.17823	0.17823
Level 7	0.03051	0.03591	0.22982	0.686	0.00888	0.00888
Level 8	0.00623	0.2742	0.02402	0.6548	0.02031	0.02044
Level 9	0.01558	0.0787	0.26235	0.62648	0.00844	0.00844
Level 10	0.00777	0.44465	0.08983	0.44465	0.00655	0.00655
Level 11	0.00307	0.0371	0.45545	0.45545	0.02447	0.02447

Level 12	0.00102	0.1998	0.1998	0.1998	0.1998	0.1998
Level 13	0.0019	0.00381	0.24857	0.24857	0.24857	0.24857
Level 14	0.00102	0.1998	0.1998	0.1998	0.1998	0.1998
Level 15	0.1998	0.00102	0.1998	0.1998	0.1998	0.1998
Level 16	0.00102	0.1998	0.1998	0.1998	0.1998	0.1998

2.5.6.3 Rescale the lists

We were not able to rely directly on the reciprocal probability of the monosaccharide types list and the reciprocal probability of the children number list for glycan reconstruction because certain monosaccharides may be absent from the original glycan monosaccharide list but appear in the reciprocal probability lists. For instance, a glycan may contain 3 Hex, 2 HexNAc, but the root selection probabilities in the reciprocal list may only be 19.98% for Hex, 0.1% for HexNAc, and the remaining 79.92% may not correspond to any monosaccharides in the original list. In such cases, it was necessary to rescale the reciprocal list while preserving the probabilities and selecting only the monosaccharides present in the original list. The specific method used for rescaling involved adjusting the probabilities in the reciprocal list to account for the missing monosaccharides,

The algorithm began by obtaining the total probabilities of all glycans that were not present in the target monosaccharide list. Next, for each monosaccharide in the target glycan list, the probability of missing monosaccharides was allocated to that particular monosaccharide based on their respective ratios. The pseudocode is:

Get the total probabilities that does not in original glycan monosaccharides list

For each monosaccharide m in original glycan list:

Allocate the probability of missing monosaccharides based on the ratio and add to m

After the rescaling process, the selection probabilities for the root were as follows: Hex at 99.502% and HexNAc at 0.498%. The ratio of Hex and HexNAc remained consistent before and after the rescaling process.

2.5.6.4 The algorithm of method

Next, a statistical approach was used to generate randomized glycan tree structures based on the reciprocal children list and reciprocal monosaccharides list, using the following algorithm, first of all, we randomly selected the number of children based on the reciprocal children probability list, then for each child, the reciprocal monosaccharide probability list was rescaled before randomly selecting a monosaccharide from it as a node, and removing it from the original monosaccharides list. This process was repeated until the original monosaccharides list became empty.

For i in all glycans:

While monosaccharides list! = null:

Random select the number of children based on reciprocal children probability

For j in children number (break if monosaccharides list is empty):

Rescaled the reciprocal monosaccharide probability list

Random select monosaccharide based on reciprocal mono probability

Add the selected monosaccharide in the decoy glycan

Delete the selected monosaccharide in monosaccharides list

Figure 16 Reciprocal based on list

2.5.7 Reciprocal probability not based on list method

In comparison to method 2.5.6, this approach is not restricted by the composition of the monosaccharide list, thereby enabling the repetition of any selection of monosaccharides. In this approach, the first step remained the same, where the number of children was randomly selected from the reciprocal children probability list. However, the difference was that, instead of rescaling the monosaccharide list, a node was directly chosen from it and added, without removing the selected monosaccharide from the target monosaccharide list. This process was repeated until the number of monosaccharides in the decoy glycan matched that of the target glycan. This method changed all the composition and position of the monosaccharide list of target glycan as well as the tree structure:

For i in all glycans:

While the number of monos in decoy!= the number of monos in target:

Random select the number of children based on reciprocal children probability

For j in children number:

Random select monosaccharide based on reciprocal mono probability

Add the selected monosaccharide in the decoy glycan

Figure 17 Reciprocal not based on list

2.6 Software

The software platform, GlycanFinder, was developed by Bioinformatics Solutions Inc as an upgraded add-on module to the PEAKS studio. A sophisticated database search and newly developed algorithms were employed to perform in-depth glycoproteomic analyses using LC-MS/MS spectra data. These analyses included protein identification and measurement, glycan and peptide scoring and sequencing, as well as peptide de novo sequencing. The platform has been used to provide comprehensive glycoproteomic analyses using LC-MS/MS spectra data.

The default glycopeptide decoy database used by GlycanFinder was generated by adding random noise masses to the original MS spectrum. A default decoy glycan database was integrated into the software to enable comparison of the performance of decoy glycan databases generated by different algorithms.

2.7 FDR estimation

When we tested the decoy database using GlycanFinder, we calculated the FDR by counting the number of decoys and targets that scored above a specific threshold for the glycan score. We used the formula below to calculate FDR:

$$\text{FDR} = \frac{Number\ of\ decoy\ glycan\ matches\ above\ shreshold}{Number\ of\ target\ glycan\ matches\ above\ shreshold} \qquad (2.3)$$

In this thesis, the equation mentioned above was used to calculate the FDR. The target glycan matches and decoy glycan matches were analyzed at the thresholds of 1%, 3%, and 100% glycan FDR.

2.8 Test method

It's worth noting that the default database of GlycanFinder was not absolutely correct, after obtaining the results, we only used the default glycan database as the reference for comparison. The percentage differences between the results obtained using different algorithms and the default database were calculated.

In the results, we counted the number of target matches for different construction functions and the default decoy database at an FDR of 1%. Then, we calculated the complement of the construction functions and the default database to determine the number of newly discovered target glycans and the number of missed target glycans. Dividing these values by the number of target matches in the default decoy database provided us with the proportion of newly discovered and missed target glycans in the decoy database:

$$The\ ratio\ of\ new\ target\ glycan = \frac{The\ number\ of\ matches\ in\ complement\ of\ new\ decoy\ database\ in\ default\ database}{The\ number\ of\ matches\ in\ default\ glycan\ database} \quad (2.4)$$

$$The\ ratio\ of\ missed\ target\ glycan = \frac{The\ number\ of\ matches\ in\ complement\ of\ default\ database\ in\ new\ decoy\ database}{The\ number\ of\ target\ matches\ in\ default\ glycan\ database} \quad (2.5)$$

2.9 Mixture model

2.9.1 Mixture model for glycopeptide score distribution

In this study, we employed the target-decoy strategy if the target and decoy databases had no overlapping and that all decoy hits were incorrect matches. However, the target hits could contain both correct and incorrect matches. Moreover, the assumptions also included the likelihood of false positive identifications is the same for both the target and decoy databases. As a result, the glycan fraction distribution could be a mixture of the correct matches and incorrect matches. To evaluate the effectiveness of the glycopeptide

identification software, we employed a statistical model to distinguish between correct and incorrect identifications.

The mixture of distributions with K mixture components formulated by:

$$f(x; \theta_1, \dots, \theta_k) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k) \quad (2.6)$$

Where θ_k represents parameters of kth component in the mixture distribution and the weight π_k , the weight of kth component, meet the conditions:

$$\sum_{k=1}^{K} \pi_k = 1$$

and

$$\pi_k \geq 0$$

2.9.2 Expectation–maximization (EM) algorithm

The EM algorithm, initially proposed by Dempster, Laird, and Rubin in 1977 [49], is an iterative method used in statistics to estimate (local) maximum likelihood or maximum a posteriori (MAP) parameter in statistical models that rely on unobserved latent variables. The EM iteration alternates between two steps, known as the expectation (E) step and the maximization (M) step. During the E step, a function for the expectation of the log-likelihood is created, evaluated using the current parameter estimate. Then, during the M step, the parameters that maximize the expected log-likelihood found in the E step are computed. These parameter estimates are subsequently utilized to determine the distribution of the latent variables in the next E step. This process is iteratively repeated until the resulting values converge to fixed points or below the threshold.

The algorithm proceeds as follows:

1. Initialization: Initialize the model parameters μ_k, σ_k, and π_k, and evaluate the log-likelihood with these parameters.

2. E-step: Evaluate the posterior probabilities $\gamma_{Zi}(k)$ using the current values of μ_k and σ_k

3. M-step: Estimate new parameters $\widehat{\mu_k}$, $\widehat{\sigma_k^2}$, and $\widehat{\pi_k}$ using the current values of $\gamma_{Zi}(k)$

4. Evaluate the log-likelihood with the new parameter estimates. If the change in the log-likelihood is less than a predetermined threshold ϵ, terminate the algorithm. Otherwise, repeat the E-step and M-step until convergence.

Chapter 3

3 Result

3.1 Software setting

This project utilized the GlycanFinder software for the querying, scoring, and sorting mass spectrometry data. Specific settings included enrichment data with a precursor mass tolerance of 10 ppm, fragment ion tolerance of 0.02 Da, and glycan fragment ion tolerance of 20 ppm. The digest mode was specified with Carbamidomethylation and Oxidation as PTMs. The mouse brain data sources tested in this study were the same as those analyzed by pGlyco3.

3.2 FDR estimation for different databases

The tables presented in the appendix display the outcomes acquired via GlycanFinder for the default database and six methods when subjected to FDR of 100%, 3%, and 1%.

Figures 18 to 24 display the glycan score distribution of each decoy database construction method when the FDR is set to 100%. Appendix A to G display the distributions when FDR=1%, appendix H to N display the distribution when FDR=3%.

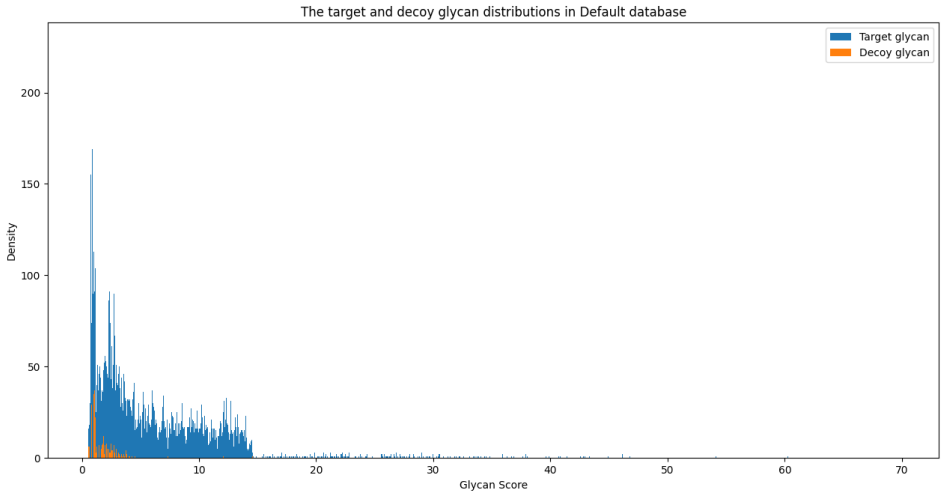


Figure 18 The Target and Decoy distributions in Default database (FDR=100%)

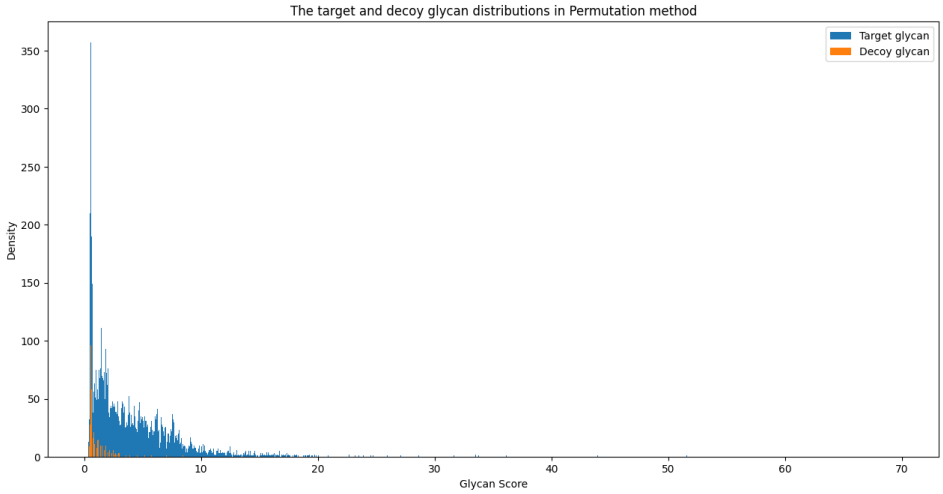


Figure 19 The Target and Decoy distributions in Permutation method (FDR=100%)

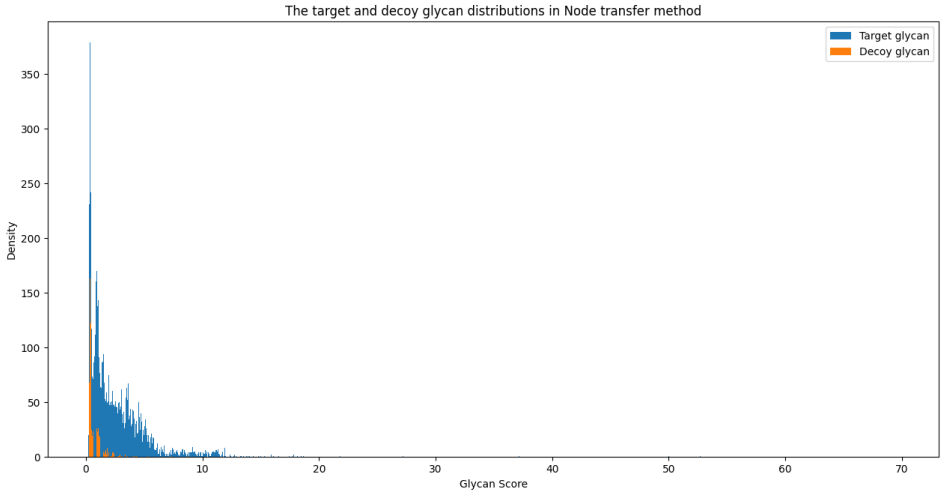


Figure 20 The Target and Decoy distributions in Node transfer method (FDR=100%)

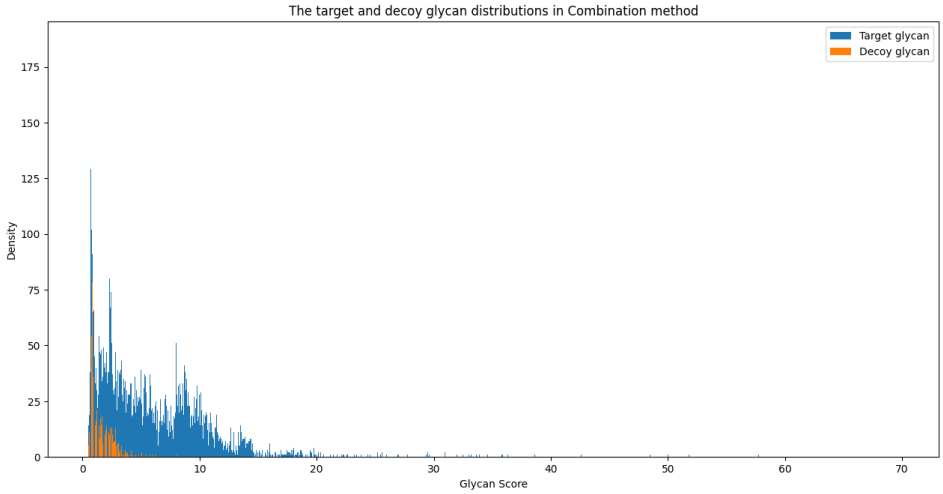


Figure 21 The Target and Decoy distributions in Combination method (FDR=100%)

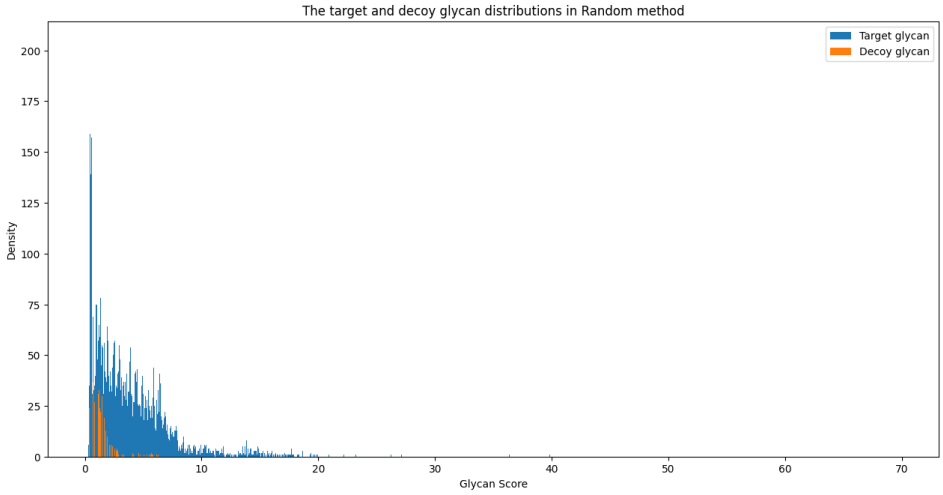


Figure 22 The Target and Decoy distributions in Random method (FDR=100%)

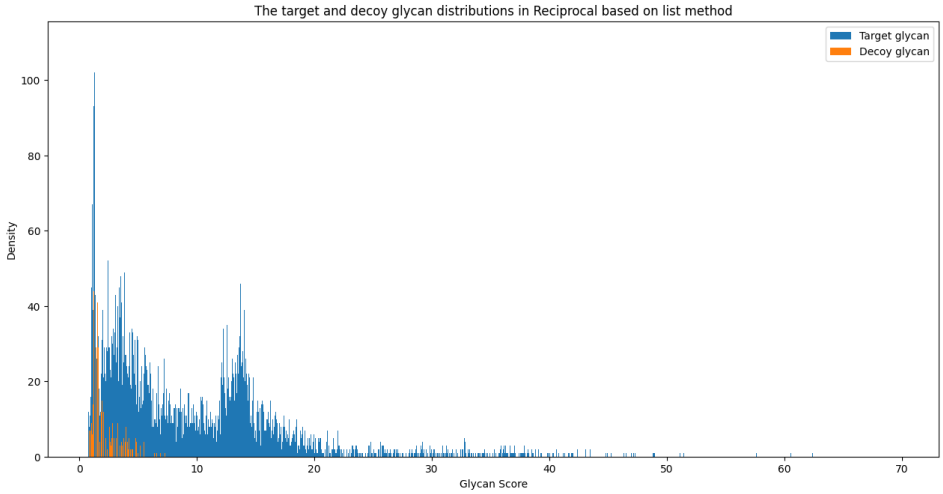


Figure 23 The Target and Decoy distributions in Reciprocal based on list method (FDR=100%)

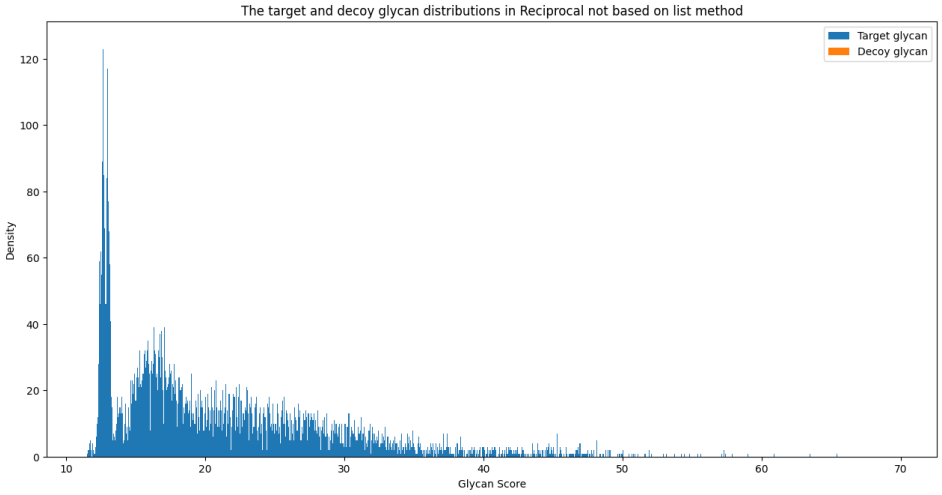


Figure 24 The Target and Decoy distributions in Reciprocal not based on list method (FDR=100%)

Table 6 The results (FDR=1%, 30 iterations)

Decoy Construction	Number of glycan	Threshold glycan score
Default	8548	4.27
Permutation	3013	6.68
Node transfer	1154	5.98
Combination	7699	4.48
Random	1666	7.02
Reciprocal based on list	8648	5.84
Reciprocal not based on list	18387	11.5

We evaluated the performance of each decoy glycan construction method using the GlycanFinder software, with data collected at an FDR of 1% (table 6). And the threshold and the number of glycans results for FDR=100% and 3% were displayed in appendix O, P. By labeling the GlycanID and based on table 6 data, we obtained the performance of each method compared to the default database using equations 2.4 and 2.5, which is shown as follows:

Compare with default database, permutation method misses 390 glycans, difference is 4.61%, finds 0 new glycans, new finding is 0.00%.

Compare with default database, node transfer method misses 553 glycans, difference is 6.53%, finds 0 new glycans, new finding is 0.00%.

Compare with default database, permutation and node transfer combination method misses 75 glycans, difference is 0.89%, finds 1 new glycans, new finding is 0.01%.

Compare with default database, random method misses 505 glycans, difference is 5.96%, finds 0 new glycans, new finding is 0.00%.

Compare with default database, reciprocal based on list method misses 3 glycans, difference is 0.04%, finds 16 new glycans, new finding is 0.19%

Compare with default database, reciprocal not based on list method misses 0 glycans, difference is 0%, finds 491 new glycans, new finding is 5.80%.

We observed that when the FDR was set to 1%, the reciprocal based on list method identified 16 new glycans while missing 3 glycans. In other words, the performance of detecting candidate glycopeptides was better than the algorithm used by GlycanFinder.

Furthermore, we observed that the data obtained using the reciprocal not based on list for FDR of 1%, 3%, and 100% were the same. This suggests that the decoy glycans generated by this algorithm did not match at all, indicating significant differences between the generated decoys and the targets. Consequently, the decoy glycan database generated by the reciprocal not based on list method is not suitable for reference comparison and cannot be used further.

On the other hand, the node transfer and random methods contained only 1000 glycans for FDR 1%. This was caused by the decoys generated by these two methods being too similar to the targets, resulting in higher scores and a lower number of detected glycans. Thus, we used these two methods only as control groups for the combination and composite based on list methods, which means they did not participate in the performance comparison.

Table 7 The results (FDR=1%, 50 iterations)

Decoy Construction	Number of glycan	Threshold glycan score
Permutation	3253	6.18
Node transfer	1216	5.84
Combination	7946	4.07
Random	1600	5.64
Reciprocal based on list	8631	4.17

We also tested the case of 50 iterations. In this situation, the results are shown in table 7. We can conclude that the performance of the reciprocal method was almost same between the decoy databases generated from 30 and 50 iterations. This was caused by the reciprocal method constructed decoys based on the statistical distribution of the target database, which was more stable. On the other hand, for these four methods (combination, permutation, node transfer, and random), 50 iterations showed slight differences compared to 30 iterations, resulting in an increased number of glycans. This phenomenon occurred due to these methods exhibiting higher randomness compared to the reciprocal methods. However, the performances of all methods from 30 and 50 iterations were similar.

As a result, by comparing the results of 50 iterations with 30 iterations, we observed that a higher number of iterations implied a greater distance between the decoy and the target. Additionally, decoy databases created by different construction methods generated varying distances to the target database. The smaller the distance, the fewer glycans were detected. Among all methods, the reciprocal-based list method exhibited the largest distance, thus detecting the highest number of glycans. On the other hand, if the distance between the decoy and target glycans was too far, it indicated greater dissimilarity, and the number of glycans increased. However, if the decoy glycans were too dissimilar from the target glycans, it might have led to a situation where there were no matches in the decoy database. Therefore, we needed to find a method to generate a decoy database that ensured the distance to the target glycans was neither too close nor too far. To be noticed, in the following tests, we used the data of 30 iterations for each decoy construction methods.

3.3 2 components mixture model

To determinate FDR value, we decomposed the total matches into correct matches and incorrect matches, then fitted them to a normal distribution for further comparison.

The normal distribution parameters are shown in appendix R, and are presented in Figures 25 to 30, which display the distribution of correct matches, incorrect matches and glycan scores. In these figures, incorrect matches are represented in black, correct matches are represented in red, and the glycan scores are blue.

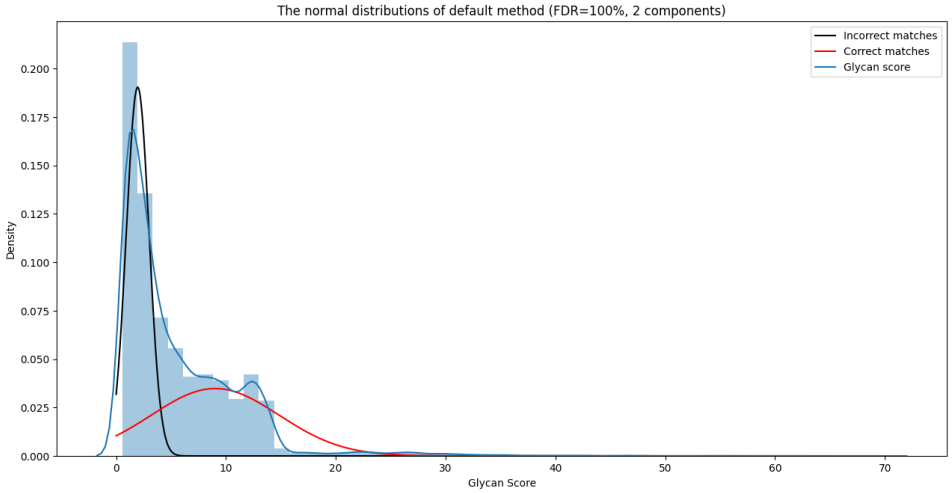


Figure 25 Incorrect and correct matches distributions in Default database (2 components)

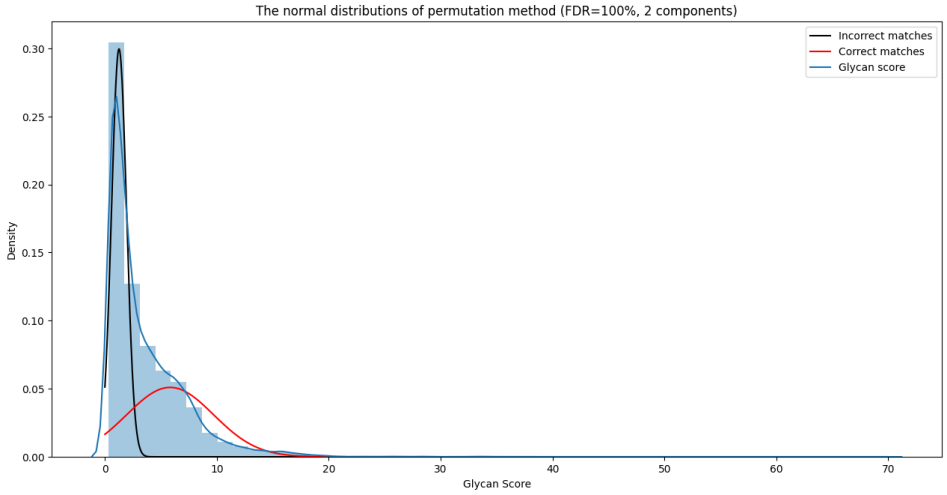


Figure 26 Incorrect and correct matches distributions in Permutation method (2 components)

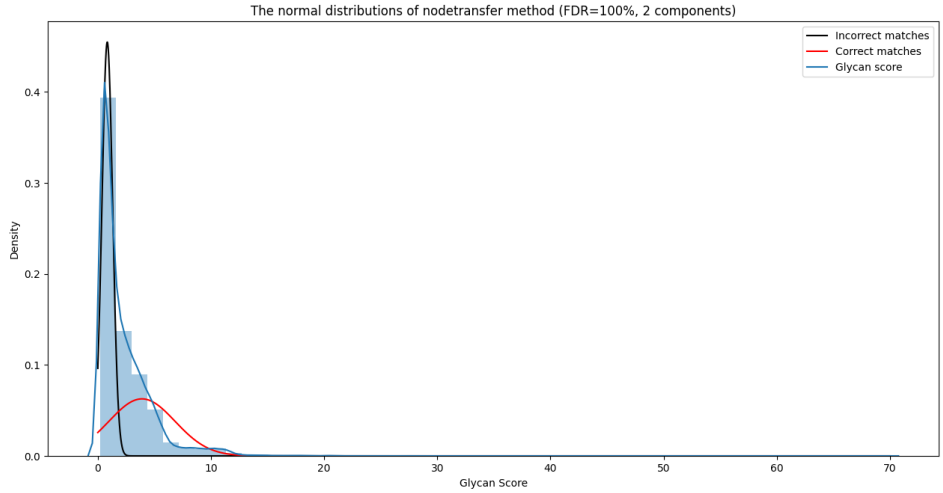


Figure 27 Incorrect and correct matches distributions in Node transfer method (2 components)

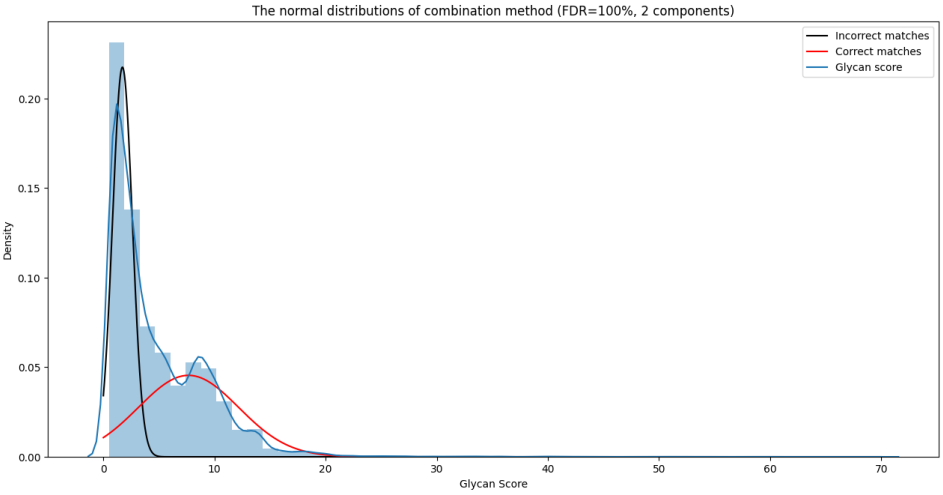


Figure 28 Incorrect and correct matches distributions in Combination method (2 components)

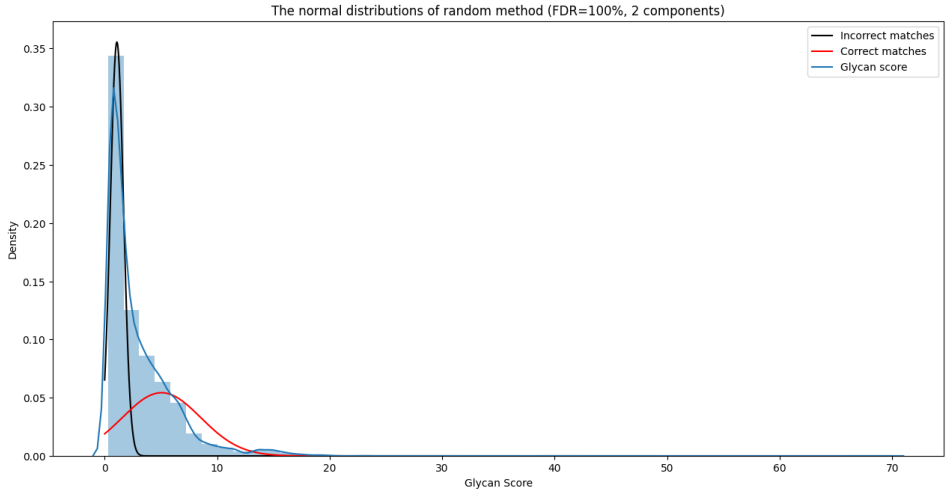


Figure 29 Incorrect and correct matches distributions in Random method (2 components)

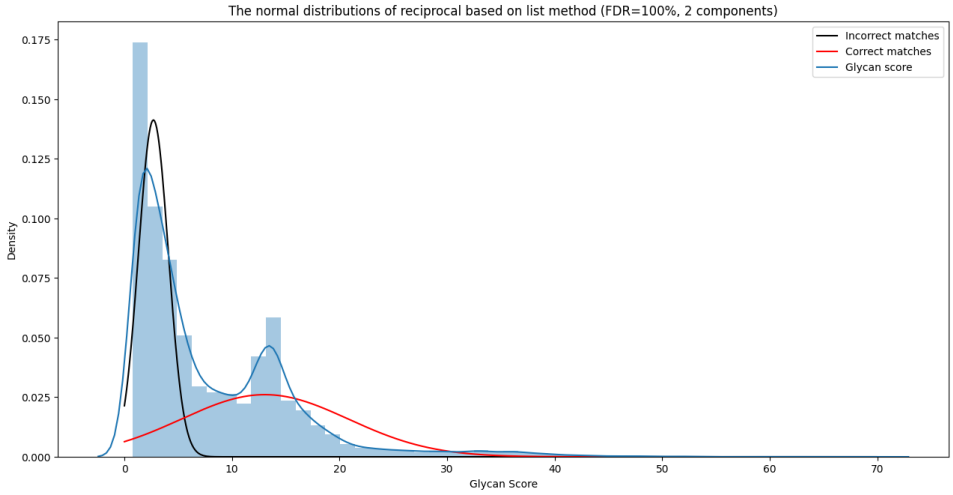


Figure 30 Incorrect and correct matches distributions in Reciprocal based on list method (2 components)

The GlycanFinder applied equation 2.3 to estimate the FDR of the results. Next, we rescaled the integral of the correct and incorrect distributions of π and calculated the FDR using the equation 1.4. The results are shown in the following figure and table:

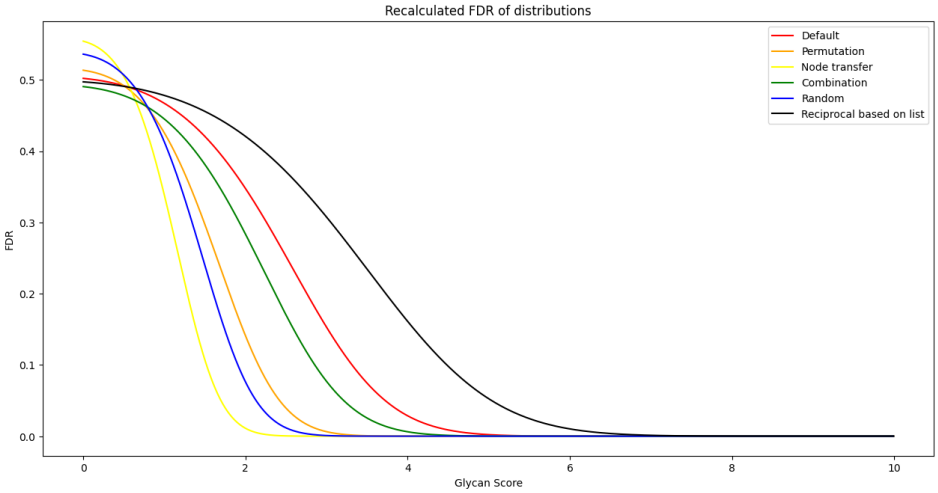


Figure 31The recalculated FDR (2 components)

Table 8 The results (FDR=1%, 2 components)

	Threshold score	Number of glycans	Number of decoys
Default database	4.45	8322	71
Permutation method	2.89	8256	275
Combination method	3.85	8494	123
Reciprocal based on list	6.01	8479	68

One method was found to have lower glycan scores than those obtained from GlycanFinder, while all other methods had differences. This indicates that the use of the 2-component Normal-based EM algorithm was not appropriate.

In image 32, an obvious "gap" was observed, which separated the distribution into two distinct parts. It was expected that the left and right sides of the gap would correspond to different distributions, but the 2-component EM algorithm did not match them accordingly. Therefore, it was necessary to consider the possibility of using more components.

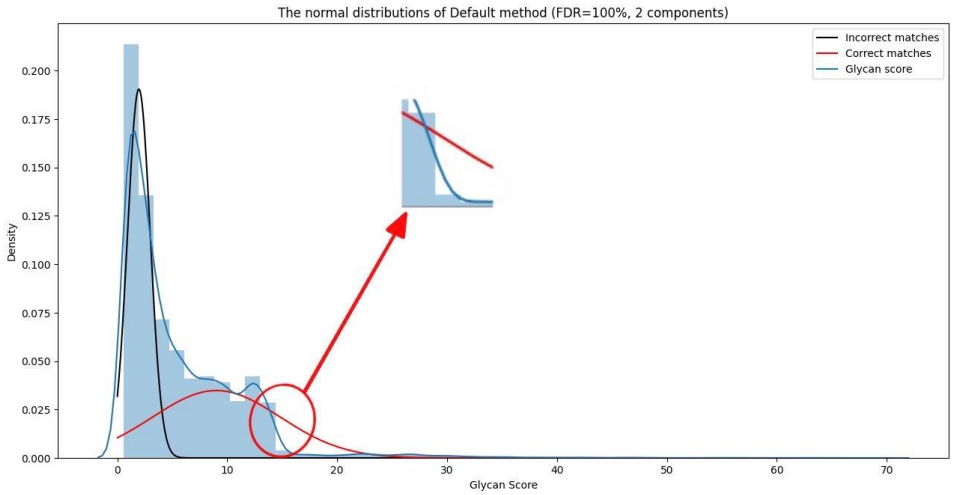


Figure 32 The “gap”

3.4 3 components

For the 3-component case, we used an EM algorithm based on a 3 components normal mixture for the entire dataset. The left distribution corresponded to incorrect matches, while the right side corresponded to two correct matches.

The parameters are showed in appendix S. In figures 33 to 38, incorrect matches are represented in black, correct matches 1 are represented in red, correct matches 2 are represented in deep blue and the glycan scores are shown in blue.

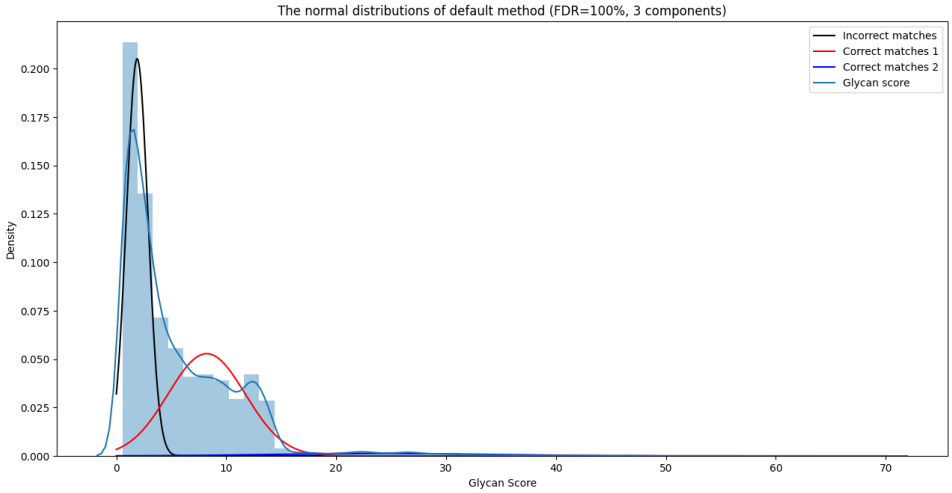


Figure 33 Incorrect and correct matches distributions in Default database (3 components)

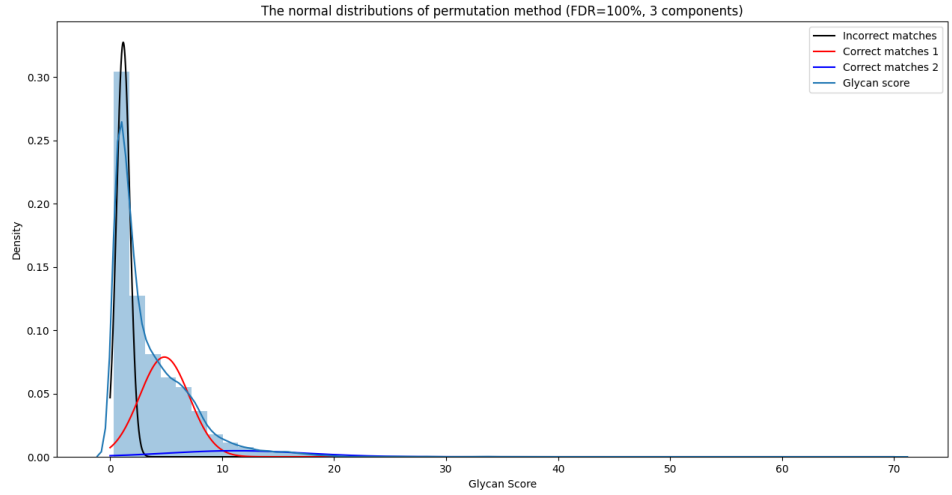


Figure 34 Incorrect and correct matches distributions in Permutation method (3 components)

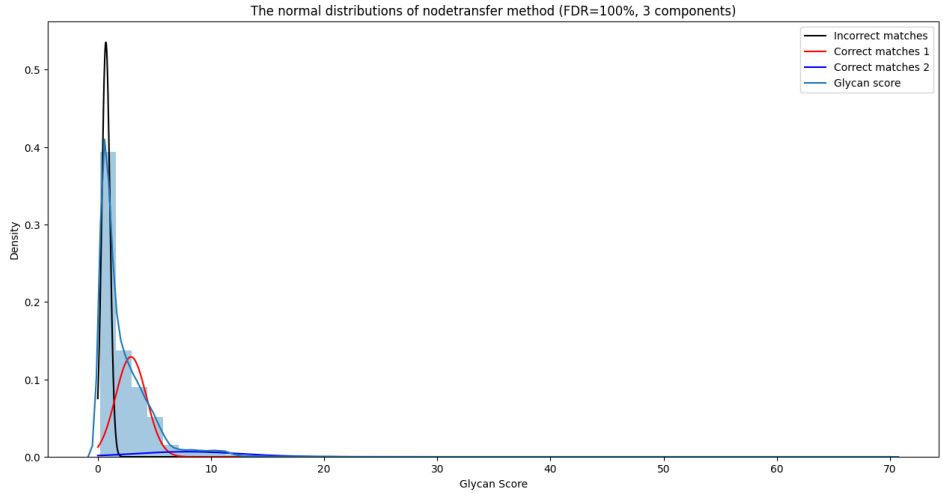


Figure 35 Incorrect and correct matches distributions in Node transfer method (3 components)

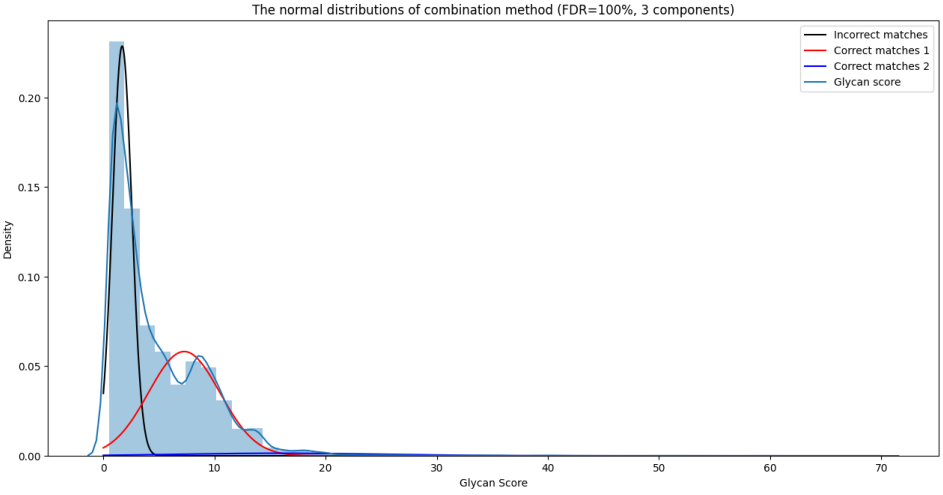


Figure 36 Incorrect and correct matches distributions in Combination method (3 components)

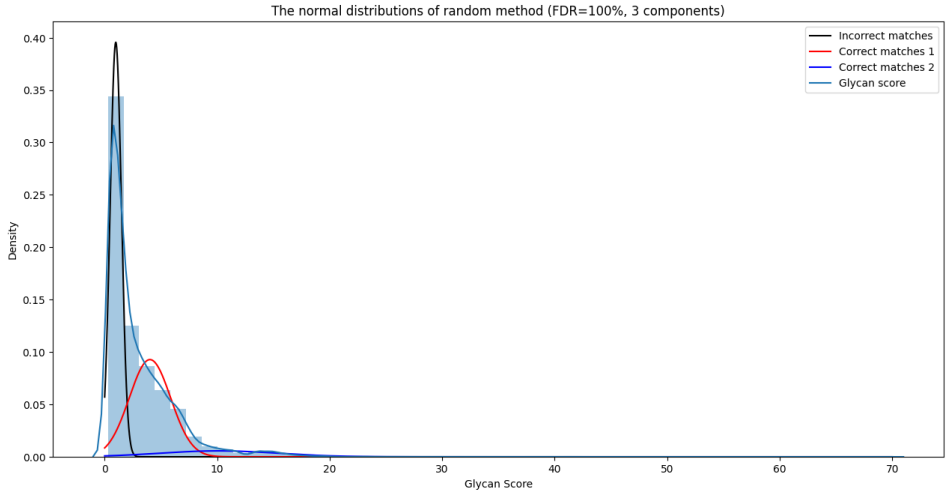


Figure 37 Incorrect and correct matches distributions in Random method (3 components)

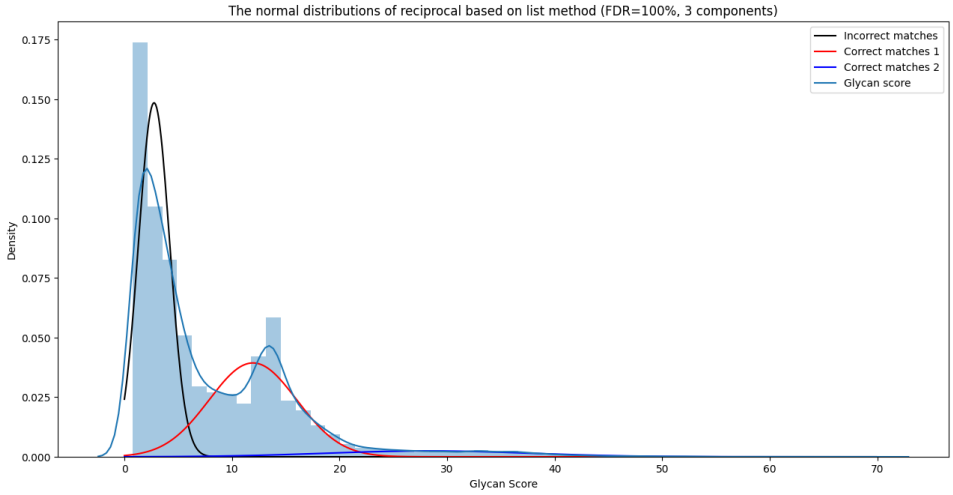


Figure 38 Incorrect and correct matches distributions in Reciprocal based on list method (3 components)

By using equation 1.4, we calculated the FDR:

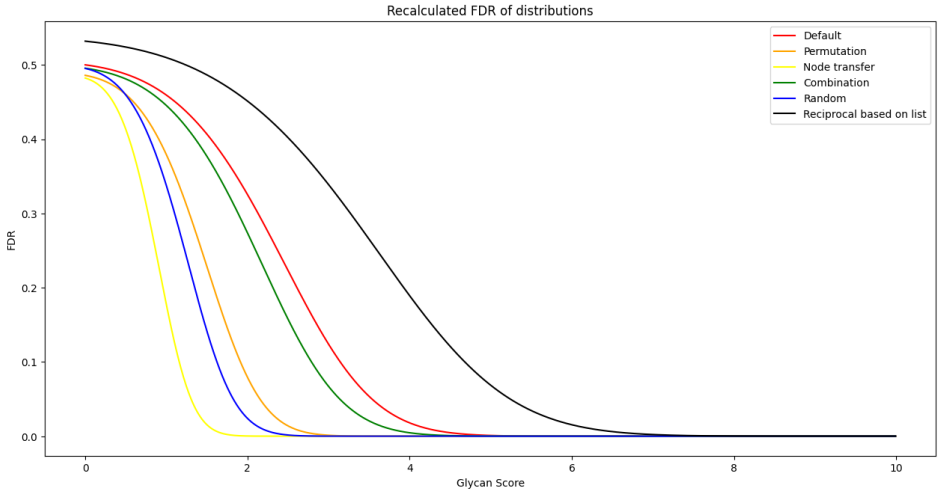


Figure 39 The recalculated FDR (3 components)

Table 9 The results (FDR=1%, 3 components)

	Threshold score	Number of glycans	Number of decoys
Default database	4.23	8601	86
Permutation method	2.58	8887	348
Combination method	3.76	8631	136
Reciprocal based on list	6.25	8299	59

The results indicated that the number of decoys recalculated by the default database was the same as those obtained directly from the GlycanFinder software. Additionally, the numbers of decoys from the two methods were higher than the theoretical value. We believe the reason for this phenomenon is that there were abnormally high peaks on the left in the dataset (figure 18 to 23) where the score is very small, and these peaks may cause overfitting by the EM algorithm. In the next section, we will subsequently consider dropping some parts of the data to ensure a better match.

3.5 4 components and gamma-normal distributions

In the Last section, we tried using 3 components, however, the results showed that this method did not perform well. Therefore, we considered using a cutoff and 4 components. At the same time, we used a mixture model of gamma and normal distributions to fit the curve.

To reduce the impact of small glycan scores on the final FDR evaluation, we first dropped off the bottom 2.5% of the lowest scores, and then we used an EM algorithm based on a 4 components gamma-normal mixture for the remaining dataset. The left distribution, corresponding to incorrect matches, was modeled as a gamma distribution, while the right side was modeled using one gumbel distribution and two normal distributions for the correct matches.

Overall, we obtained four distributions: the gamma distribution represented the first incorrect matches, located on the far left of the plot; the second gamma distribution, located on the left side of the figure, represented the second incorrect match; and the third distribution, located in the middle, represented the first correct match, and the fourth distribution represented the second correct match, which is located on the right.

The parameters are shown in appendix T. In figures 40 to 45, incorrect matches are represented in black, correct matches 1 are represented in red, correct matches 2 are represented in deep blue, correct matches 3 are represented in green and the glycan scores are shown in blue.

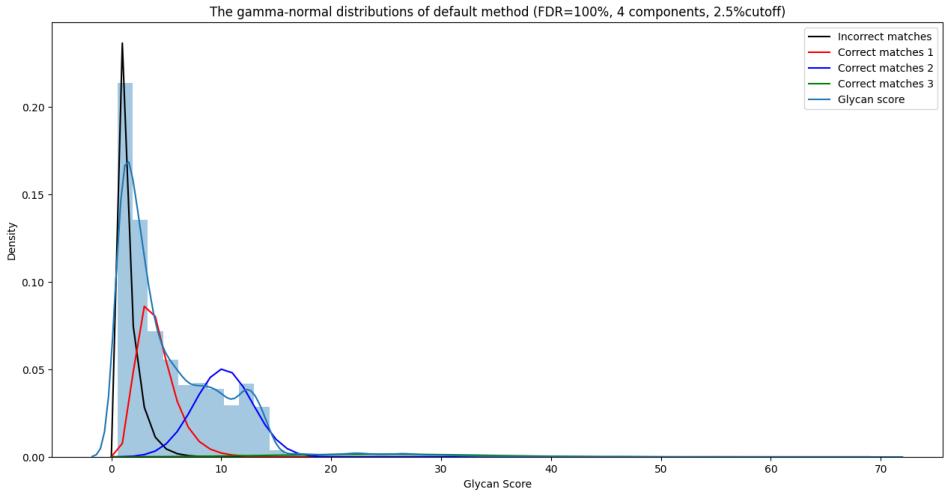


Figure 40 Incorrect and correct matches distributions in Default database (4 components, 2.5% cutoff)

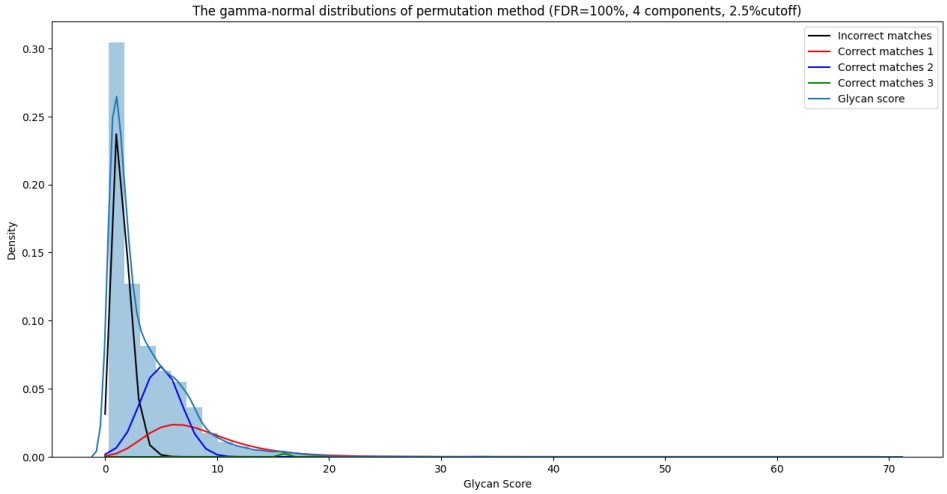


Figure 41 Incorrect and correct matches distributions in Permutation method (4 components, 2.5% cutoff)

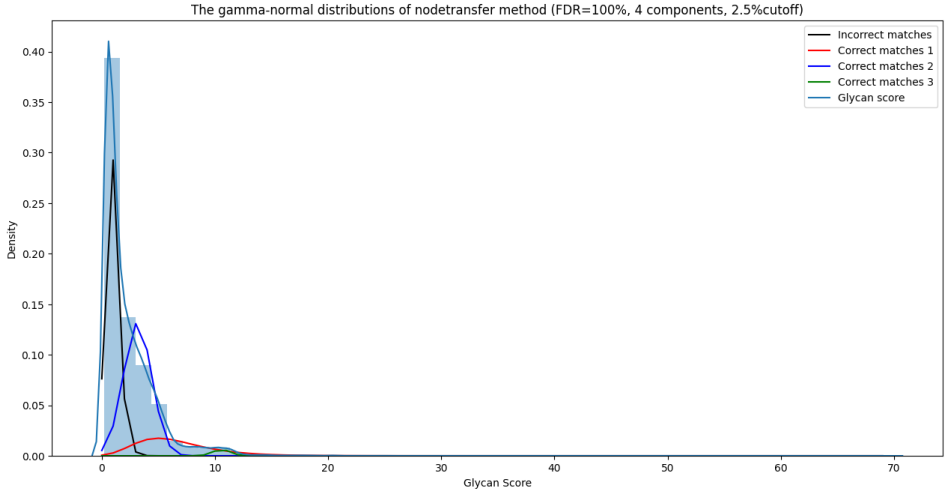


Figure 42 Incorrect and correct matches distributions in Node transfer (4 components, 2.5% cutoff)

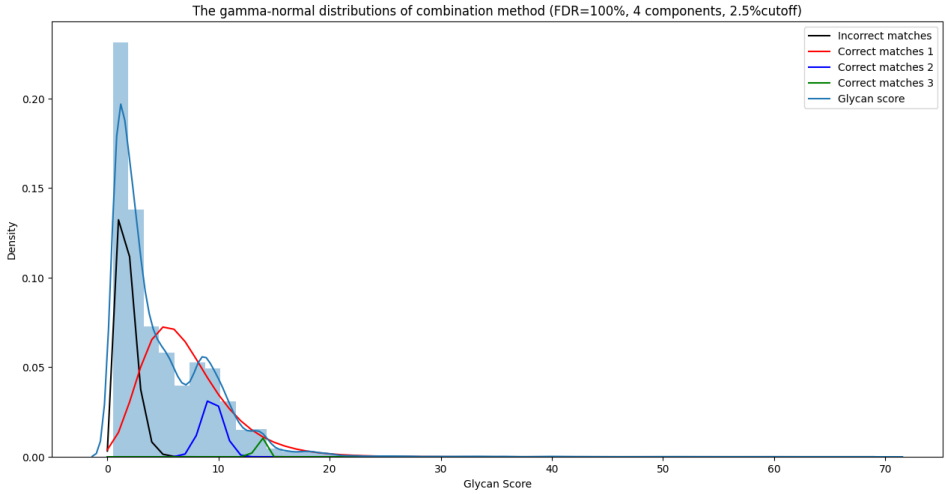


Figure 43 Incorrect and correct matches distributions in Combination method (4 components, 2.5% cutoff)

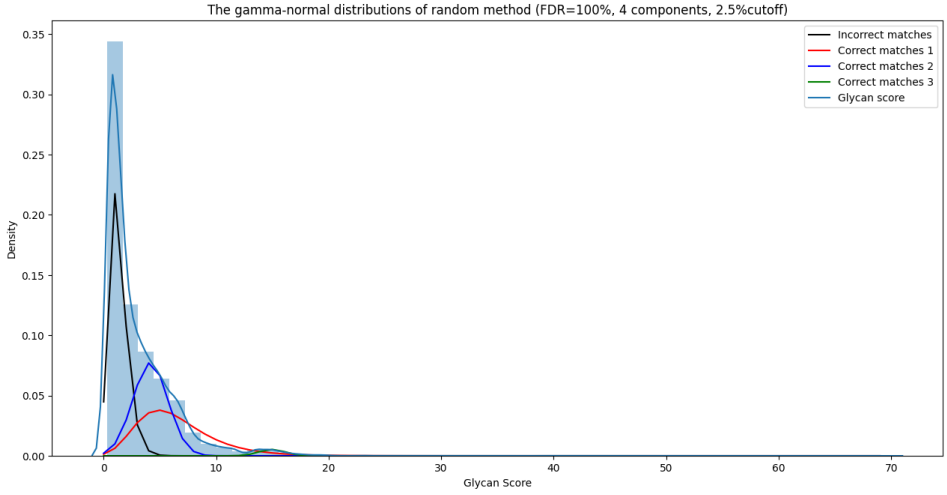


Figure 44 Incorrect and correct matches distributions in Random method (4 components, 2.5% cutoff)

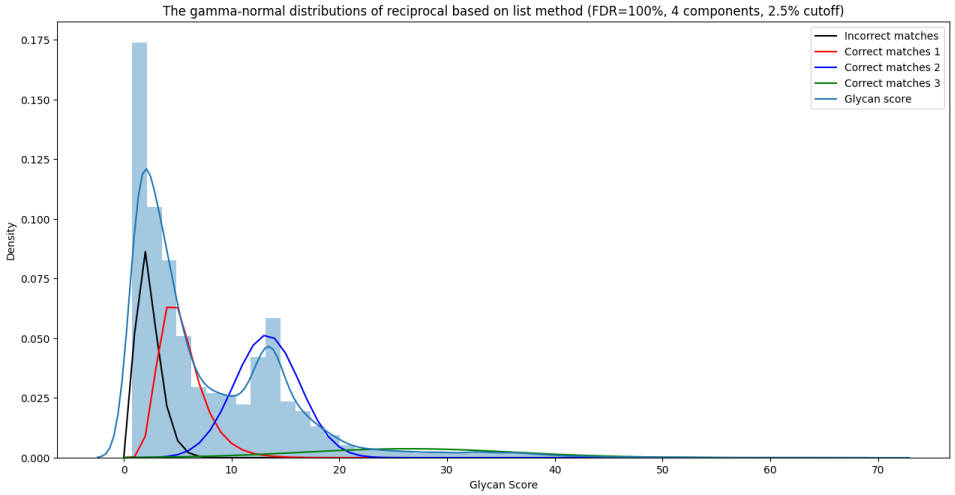


Figure 45 Incorrect and correct matches distributions in Reciprocal based on list (4 components, 2.5% cutoff)

With the same approach, we recalculated the FDR.

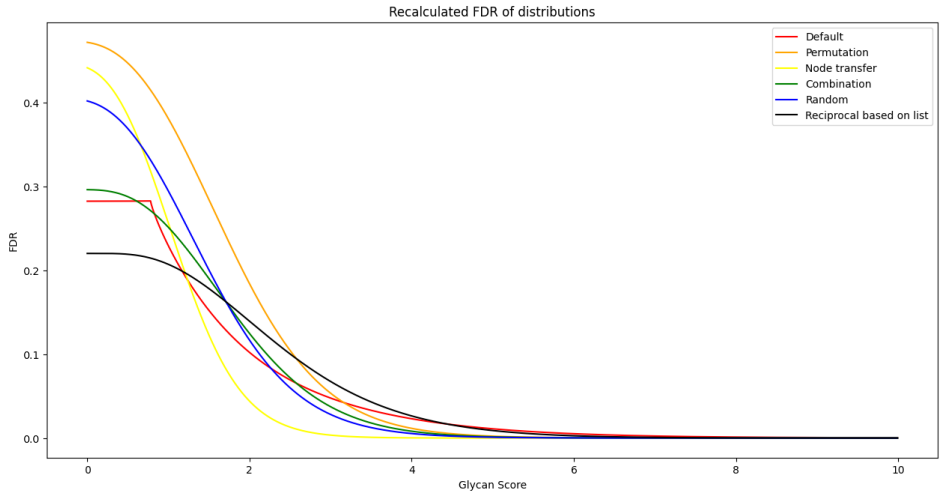


Figure 46 The recalculated FDR (4 components, 2.5% cutoff)

Table 10 The results (FDR=1%, 4 components, 2.5% cutoff)

	Threshold score	Number of glycans	Number of decoys
Default database	5.13	7591	46
Permutation method	4.09	6252	146
Combination method	3.88	8450	122
Reciprocal based on list	4.91	9610	207

The results showed that among the four methods (not including node transfer and random), one method had fewer decoys than theoretically expected at an FDR of 1%. Although the other three methods performed poorly with higher-than-expected decoy numbers, two of them had a better performance than the 3 components condition.

3.6 5 components

Previously we dropped a portion of the data to make sure it was a better match. However, we were not sure how much data we had to cut to avoid the abnormal peak, and sometimes we were required to use the entire dataset. Therefore, in this section, we propose a new approach to circumvent this overfitting problem. We will add a new normal distribution to match this peak without any data drops, and then apply the previous 4 components method. The overall distributions will contain a normal distribution on the far left for the incorrect match of the abnormal peak, a gamma distribution on the right for the second incorrect match, and a gumbel distribution on the right for the first correct match, the two normal distributions on the far right represent the two correct matches.

The parameters are shown in appendix U. In figures 47 to 52, incorrect matches 1 are represented in yellow, incorrect matches 2 are represented in black, incorrect matches 2 are represented in red, correct matches 1 are represented in deep blue, correct matches 2

are represented in green and the glycan scores are in blue.

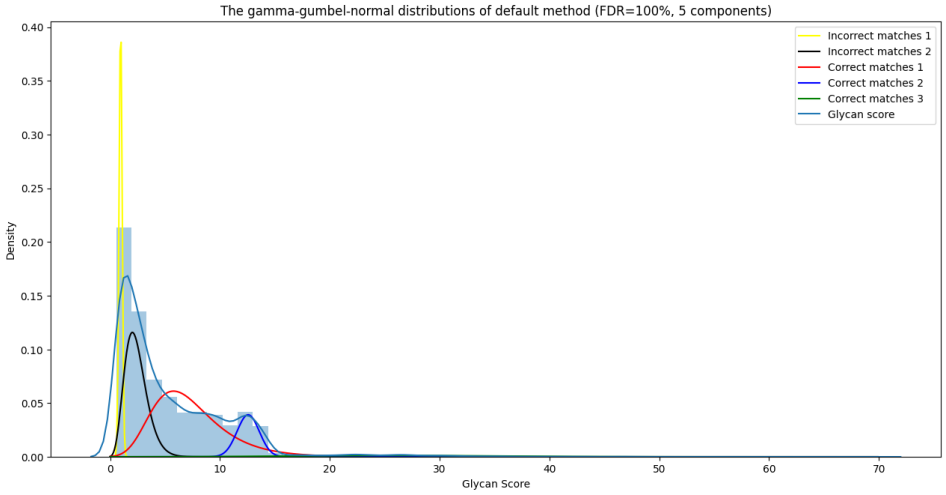


Figure 47 Incorrect and correct matches distributions in Default (5 components)

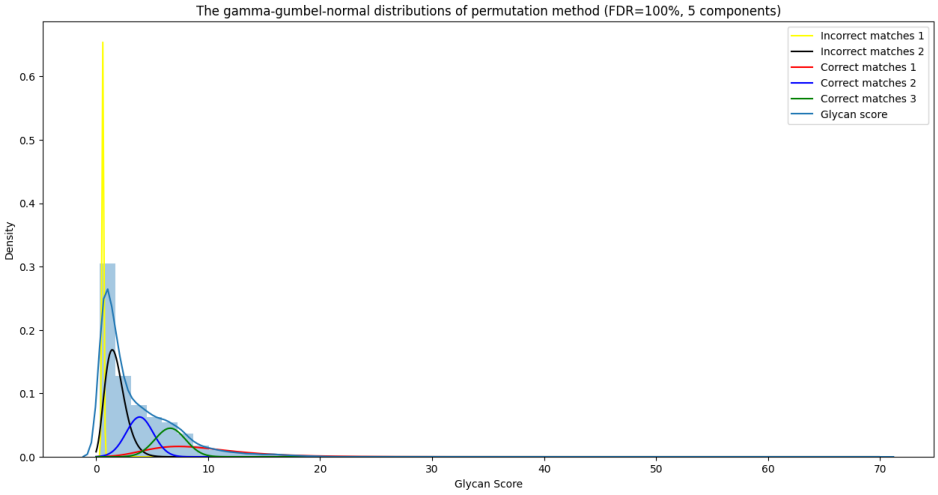


Figure 48 Incorrect and correct matches distributions in Permutation (5 components)

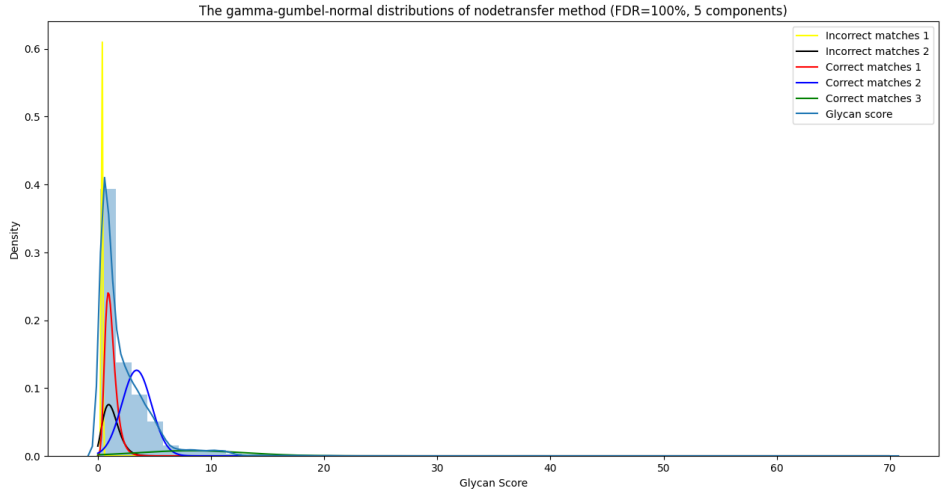


Figure 49 Incorrect and correct matches distributions in Node transfer (5 components)

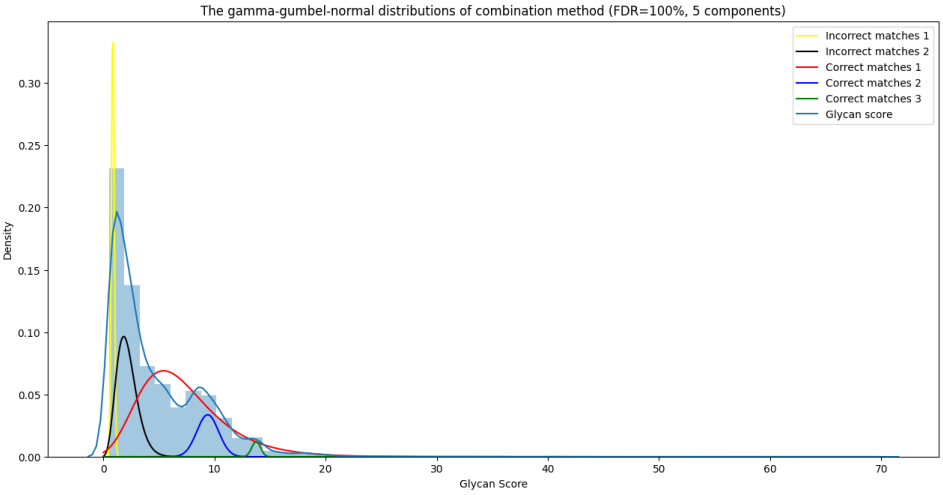


Figure 50 Incorrect and correct matches distributions in Combination (5 components)

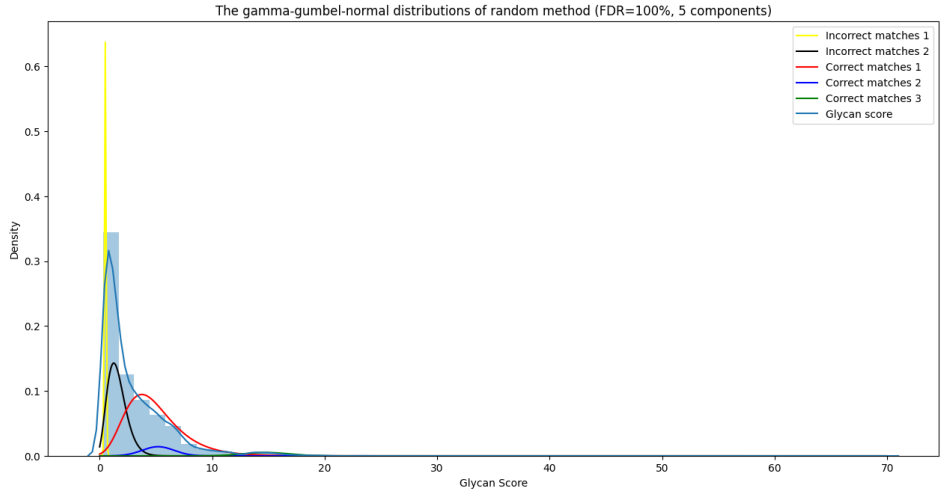


Figure 51 Incorrect and correct matches distributions in Random (5 components)

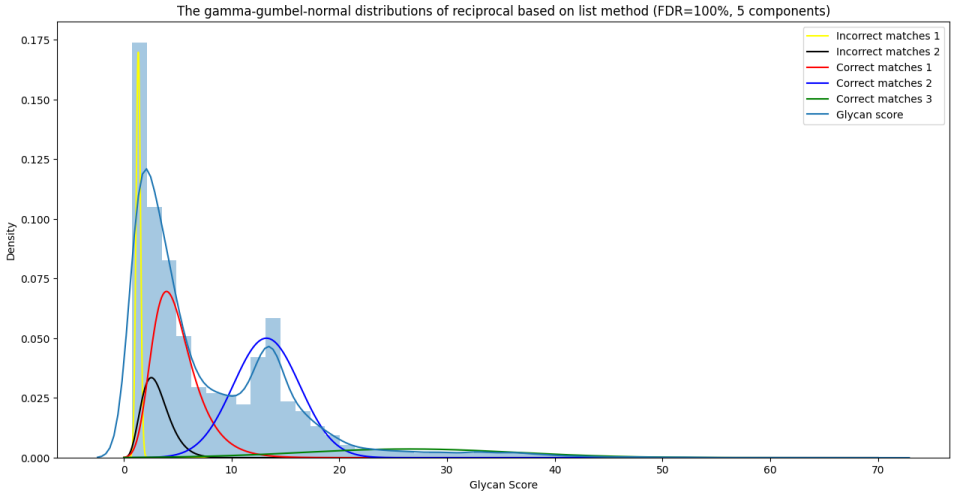


Figure 52 Incorrect and correct matches distributions in Reciprocal based on list (5 components)

With the same approach, we recalculated FDR.

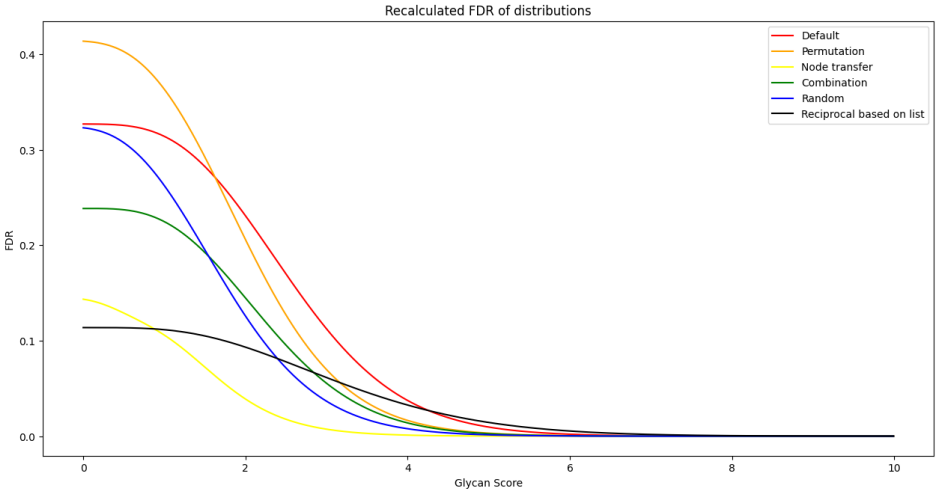


Figure 53 The recalculated FDR (5 components)

Table 11 The results (5 components)

	Threshold score	Number of glycans	Number of decoys
Default database	4.97	7757	52
Permutation method	4.32	5915	130
Combination method	4.21	8147	107
Reciprocal based on list	5.39	9100	130

The results showed that among the four methods (not including node transfer and random), the default database detected fewer decoys than theoretically expected at an FDR of 1%, and the number of glycans detected was close to that obtained from the software. Although, the other three methods had higher decoy numbers, their threshold score and the number of decoys were the best among all other components condition, and they were close to the software outputs (table 6).

Overall, after testing different components, although the default database of GlycanFinder was not absolutely correct, as a comparison, we can conclude that using an EM algorithm based on a mixture distribution to calculate the distribution of incorrect and correct matches and to estimate FDR through CDF is a feasible method to consider.

Chapter 4

4 Conclusion and discussion

In this study, we constructed seven glycan databases, including permutation, node transfer, and reciprocal probability methods. To evaluate the performance of each method, we first used the GlycanFinder software for glycan scoring, with test data from mouse brain spectra used in testing Pglyco3. Additionally, we evaluated and compared the distribution of the obtained results. Moreover, we used the EM algorithm based on mixture distributions to separate correct and incorrect parameters and visualize them. Finally, by using the CDF of the distributions to recalculate the threshold at FDR=1% and compare it with the data from software.

As for the results, we found that combining the node transfer and permutation method, which alters the structure of the glycan tree, was more efficient than using only permutation algorithm, which only rearranges monosaccharides, it caused by more reported number of glycans when FDR=1%. Furthermore, constructing a reverse probability list and using the original glycan monosaccharides list resulted in the discovering of more potential candidate glycans compared to the default database used by the GlycanFinder software. However, the decoys of the fully random construction method were too close to the targets since it only reported thousand around glycans when FDR=1%. On the other side, the database constructed using the method of constructing a reverse list not based on the original glycan monosaccharides list had a largest distance and differed greatly from the original database and could not match effective decoys.

Additionally, we tested 2 and 3 components EM algorithm based on normal distribution. However, the results were not satisfactory. Finally, we tested the 4 components EM algorithm with a 2.5% cutoff and the 5 components approach for recalculating the FDR, and the results showed that 3 out of 4 algorithms generated decoy databases that were close to the conclusion obtained from the software. Therefore, we concluded that using the EM algorithm to estimate FDR is a feasible approach.

However, there may be some limitations to this study. In this work, we employed a method of randomly generating decoy glycan databases, but there were still many possibilities that were not considered, even though we created 30 candidate glycans and compared their distances. According to Zhikai Zhu [50], the efficiency of a 1-to-many decoy database is higher than that 1-to-1 decoy database. Therefore, in the future, we may consider using algorithms to generate a 1-to-many decoy glycan database to improve performance.

Lastly, in this study, we only tested gamma, gumbel and normal distributions in the mixture model. In future work, it is possible to test more potential mixture distributions, such as an extreme value distribution, and different combinations of distributions to better match the correct and incorrect distribution. More importantly, determining the number of components is crucial, we must find the optimal number of distributions. Additionally, determining the initial parameters can be important too.

5 References or Bibliography

[1] Bagdonaite, I. M. (2022). Glycoproteomics. Nat Rev Methods Primers.

[2] Yuan Tian, H. Z. (2010). Glycoproteomics and clinical applications. Proteomics Clinical Apps, 124-132

[3] Suttipong Suttapitugsakul, F. S. (2020). Recent Advances in Glycoproteomic Analysis by Mass Spectrometry. Anal Chem.

[4] BeMiller, J. N. (2019). Carbohydrate Chemistry for Food Scientists (Third Edition).

[5] Caballero, B. (2023). Encyclopedia of Human Nutrition (Fourth Edition).

[6] Elhadi Yahia, A. C.-L. (2018). Postharvest Physiology and Biochemistry of Fruits and Vegetables (1st Edition).

[7] Shuang Yang, N. H. (2017). Simultaneous analyses of N-linked and O-linked glycans of ovarian cancer cells using solid-phase chemoenzymatic method. Clin Proteom.

[8] L. Renee Ruhaak, G. X. (2018). Mass Spectrometry Approaches to Glycomic and Glycoproteomic Analyses. Chem. Rev.

[9] Inka Brockhausen, H. S. (n.d.). O-GalNAc Glycans. In Essentials of Glycobiology.

[10] Li Cao, Y. Q. (2016). Intact glycopeptide characterization using mass spectrometry.

[11] Xin You, H. Q. (2018). Recent advances in methods for the analysis of protein o-glycosylation at proteome level.

[12] S. Herget, R. R.-W. (2008). GlycoCT—a unifying sequence format for carbohydrates. Carbohydrate Research, 2162-2171.

[13] Haag, A. M. (2016). Mass Analyzers and Mass Spectrometers.

[14] Bandeira, A. G. (2012). Peptide Identification by Tandem Mass Spectrometry with Alternate Fragmentation Modes. Mol Cell Proteomics.

[15] Orbitrap Fusion Mass Spectrometer for Glycan and Glycopeptide Analysis. (n.d.). Thermo Scientific.

[16] Hong Yang, C. Y. (2018). Characterization of glycopeptides using a stepped higher-energy C-trap dissociation approach on a hybrid quadrupole orbitrap. Rapid Communications in MS.

[17] Yao-Yi Chen, S. D.-Q.-M. (2012). Refining comparative proteomics by spectral counting to account for shared peptides and multiple search engines. Analytical and Bioanalytical Chemistry.

[18] UniProt: a hub for protein information. (2015). Nucleic Acids Research.

[19] Kim D. Pruitt, T. T. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research.

[20] Mascot Science. (n.d.). Retrieved from https://www.matrixscience.com/

[21] SEQUEST. (n.d.). Retrieved from UWPR: https://proteomicsresource.washington.edu/protocols06/sequest.php

[22] Jürgen Cox, N. N. (2011). Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. Journal of Proteome Research.

[23] Paulo, J. A. (2014). Practical and Efficient Searching in Proteomics: A Cross Engine Comparison.

[24] Griss, J. (2016). Spectral library searching in proteomics. Proteomics.

[25] Genet Abay Shiferaw, E. V.-J. (2020). COSS: A Fast and User-Friendly Tool for Spectral Library Searching. Journal of Proteome Research.

[26] Mohammad Mahmoudi Gomari a, N. S.-A. (2020). Opportunities and challenges of the tag-assisted protein purification techniques: Applications in the pharmaceutical industry. Biotechnology Advances.

[27] Xusheng Wang, Y. L. (2014). JUMP: A Tag-based Database Search Tool for Peptide Identification with High Sensitivity and Accuracy. MCP

[28] Fusong Jua, J. Z. (2019). De novo glycan structural identification from mass spectra using tree merging strategy. Computational Biology and Chemistry.

[29] Jiechen Shen, L. J. (2021). StrucGP: de novo structural sequencing of site-specific N-glycan on glycoproteins using a modularization strategy. nature.

[30] Bin Ma, K. Z.-K. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Communications in MS.

[31] Pevzner, A. F. (2005). PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. Analytical Chemistry.

[32] Hao Yang, H. C.-F.-J.-M. (2019). pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework . Bioinformatics.

[33] Guanghui Wang, W. W.-F. (2009). Decoy Methods for Assessing False Positives and False Discovery Rates in Shotgun Proteomics. PMC.

[34] Lukas Käll, J. D. (2008). Assigning Significance to Peptides Identified by Tandem Mass. Journal of Protemome research.

[35] Gygi, J. E. (2007). Target-decoy search strategy for increased confidence. Nature Methods.

[36] Gygi, J. E. (2010). Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods Mol Biol.

[37] Keegan Korthauer, P. K. (2019). A practical guide to methods controlling false discoveries in computational biology. Genome Biology.

[38] Hyunwoo Kim, S. L. (2019). Target-small decoy search strategy for false discovery rate estimation. BMC Bioinformatics.

[39] Fergus Imrie, A. R. (2021). Generating property-matched decoy molecules using deep learning. Bioinformatics.

[40] Shun, S. (2022). False Discovery Rate Analysis for Glycopeptide Identification.

[41] Suruchi Aggarwal, A. K. (2015). False Discovery Rate Estimation in Proteomics. In Statistical Analysis in Proteomics (pp. 119-128).

[42] Ji Zexuan, H. Y. (2017). A robust modified Gaussian mixture model with rough set for image segmentation. Neurocomputing.

[43] Siyu Liu, X. Z. (2022). Expectation–maximization algorithm for bilinear systems by using the Rauch–Tung–Striebel smoother. Automatica.

[44] Shu Kay Ng, T. K. (2011). The EM Algorithm. Handbook of Computational Statistics.

[45] Mohammad Nadjafi, P. G. (2022). Expectation-maximization algorithm to develop a normal distribution NHPP reliability growth model. Engineering Failure Analysis.

[46] Wen-Feng Zeng, W.-Q. C.-Q.-M.-Y. (2021). Precise, fast and comprehensive analysis of intact glycopeptides and modified glycans with pGlyco3.

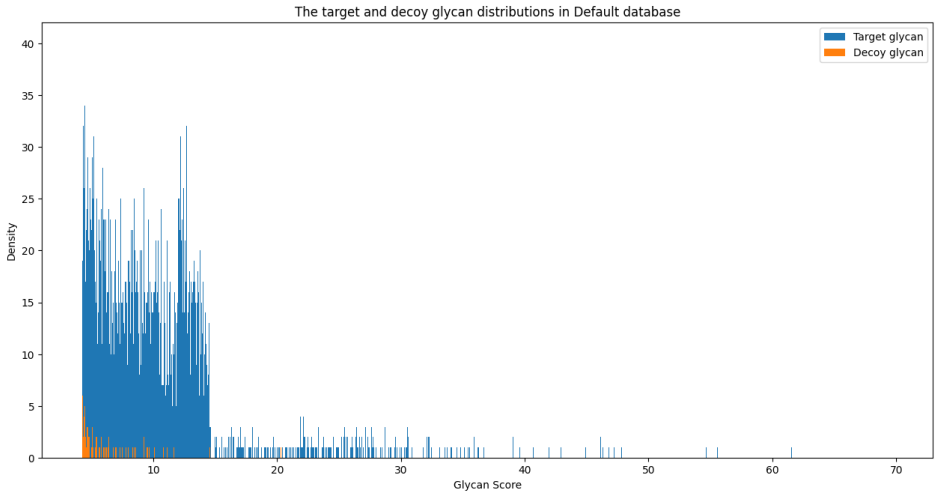
[47] Jing Zhang, L. X. (2012). PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Molecular & Cellular Proteomics.

[48] Sun, W. (2016). ALGORITHMS FOR GLYCAN STRUCTURE

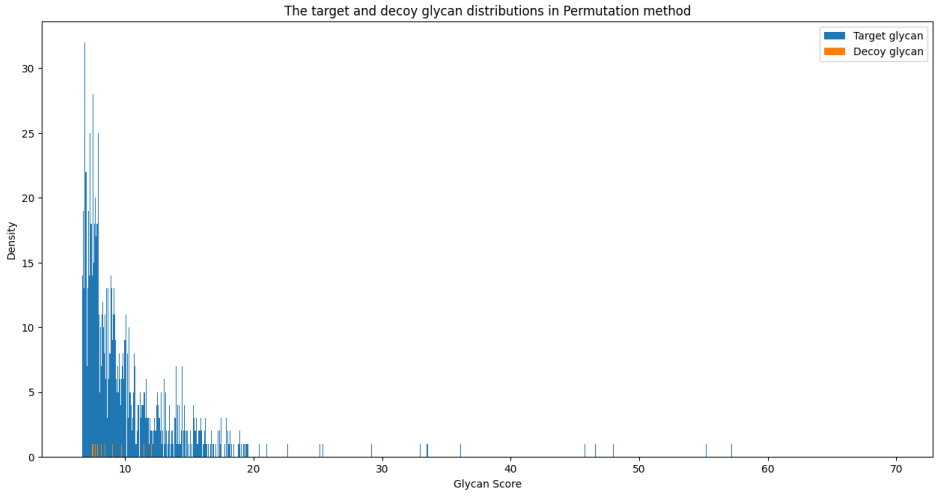
[49] Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society.

[50] Zhikai Zhu, X. S. (2014). New Glycoproteomics Software, GlycoPep Evaluator, Generates Decoy Glycopeptides de Novo and Enables Accurate False Discovery Rate Analysis for Small Data Sets. analytical chemistry.

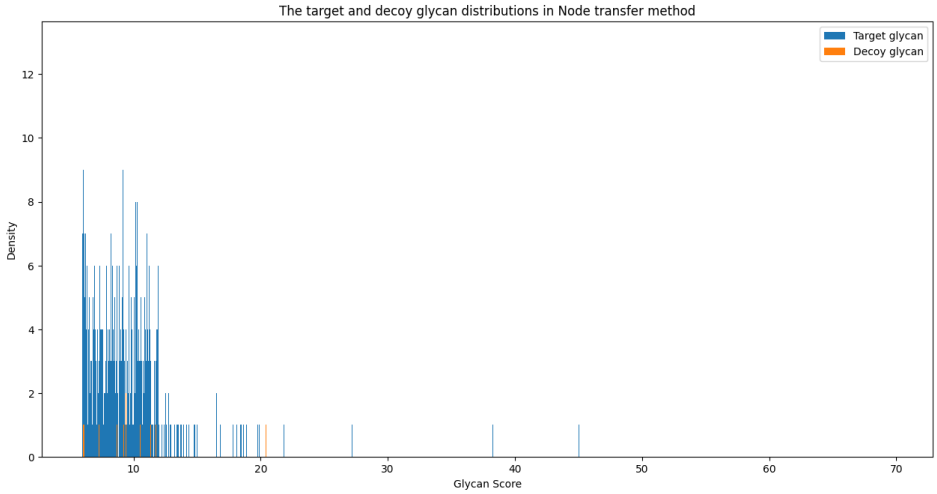
Appendices



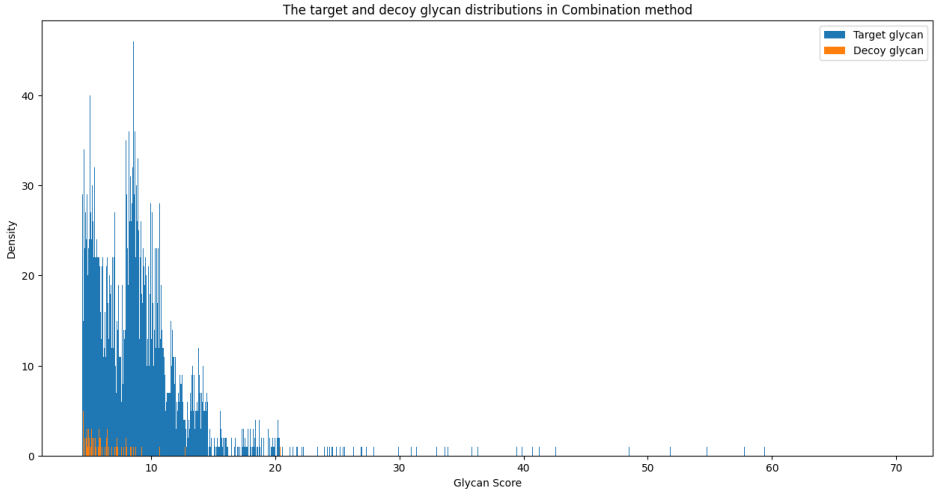
Appendix A The Target and Decoy glycan distributions in Default database (FDR=1%)



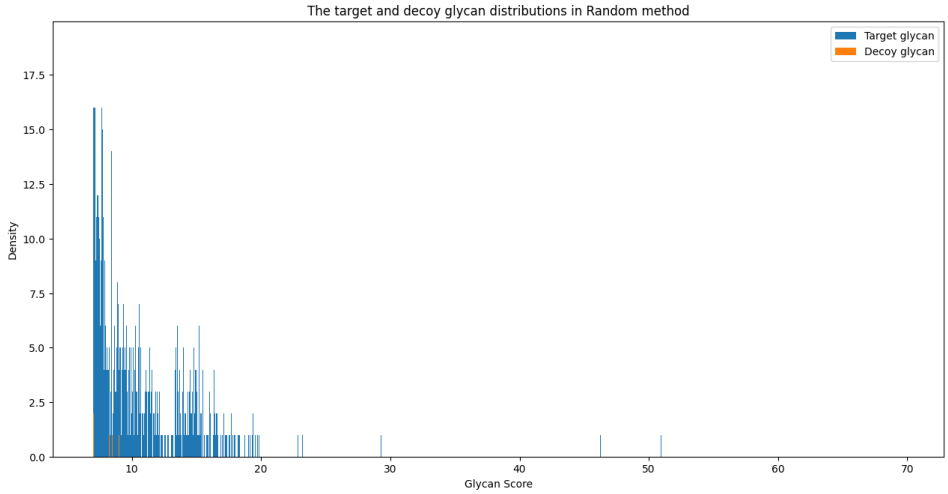
Appendix B The Target and Decoy glycan distributions in Permutation method (FDR=1%)



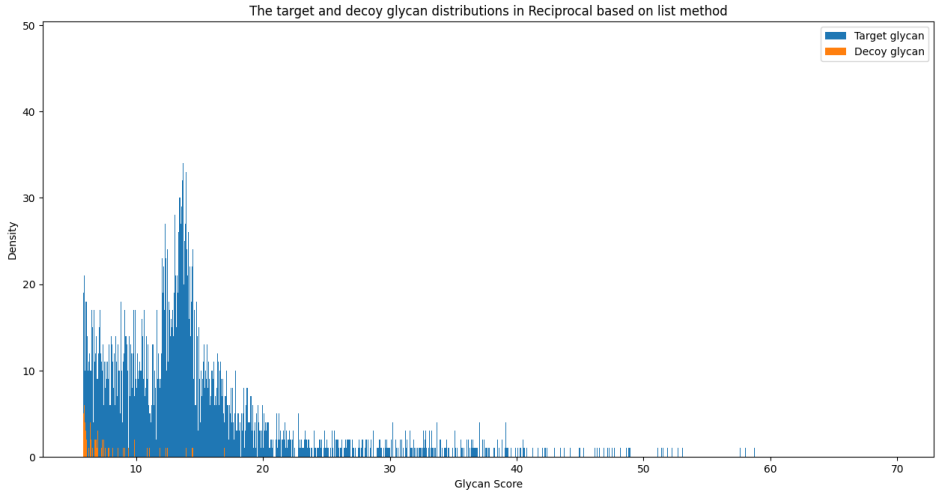
Appendix C The Target and Decoy glycan distributions in Node transfer method (FDR=1%)



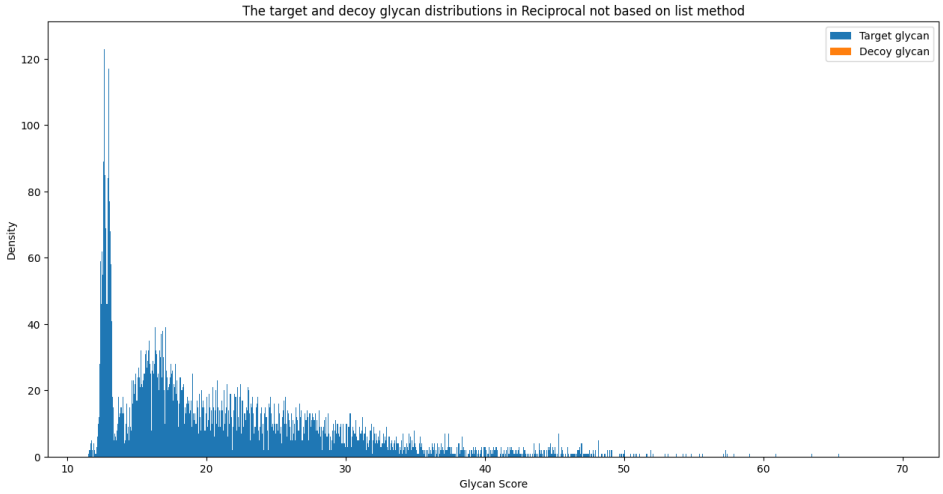
Appendix D The Target and Decoy glycan distributions in Combination method (FDR=1%)



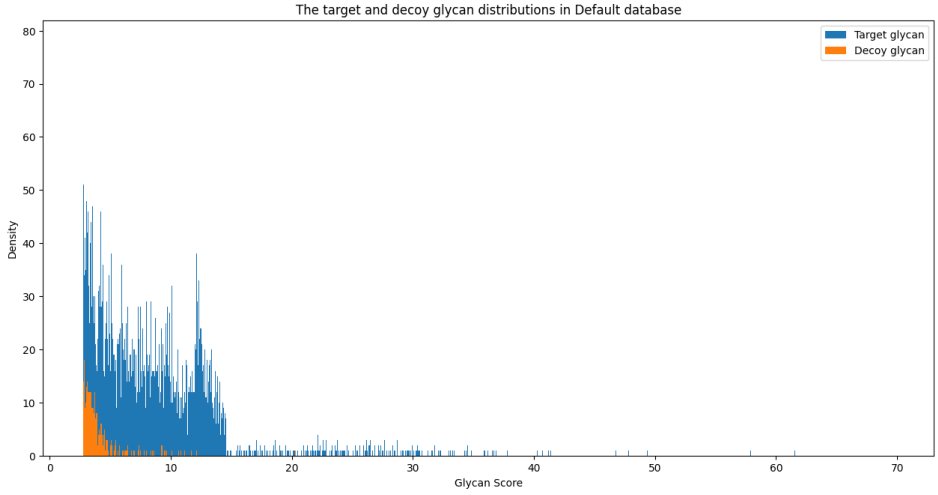
Appendix E The Target and Decoy glycan distributions in Random method (FDR=1%)



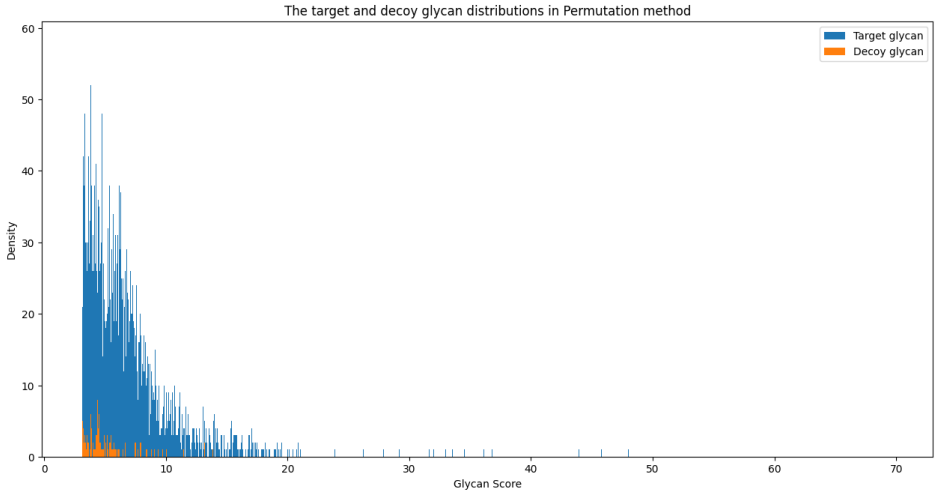
Appendix F The Target and Decoy glycan distributions in Reciprocal based on list method (FDR=1%)



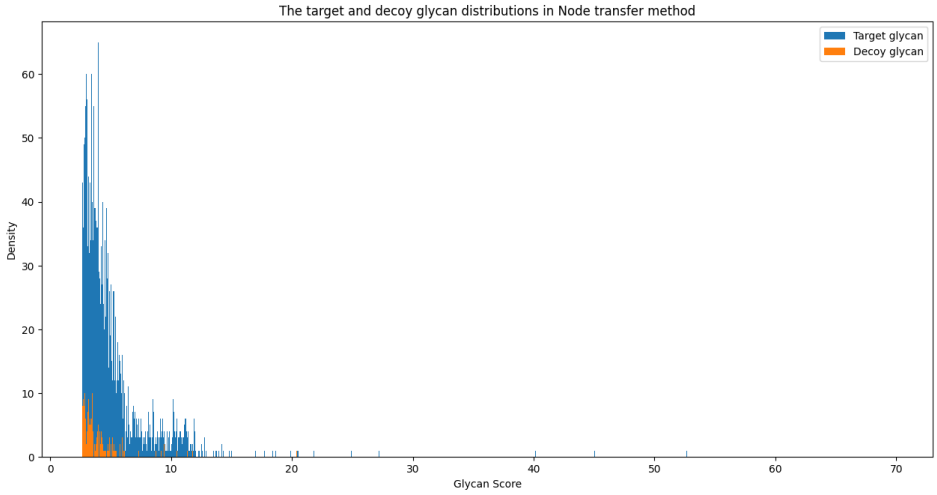
Appendix G The Target and Decoy glycan distributions in Reciprocal not based on list (FDR=1%)



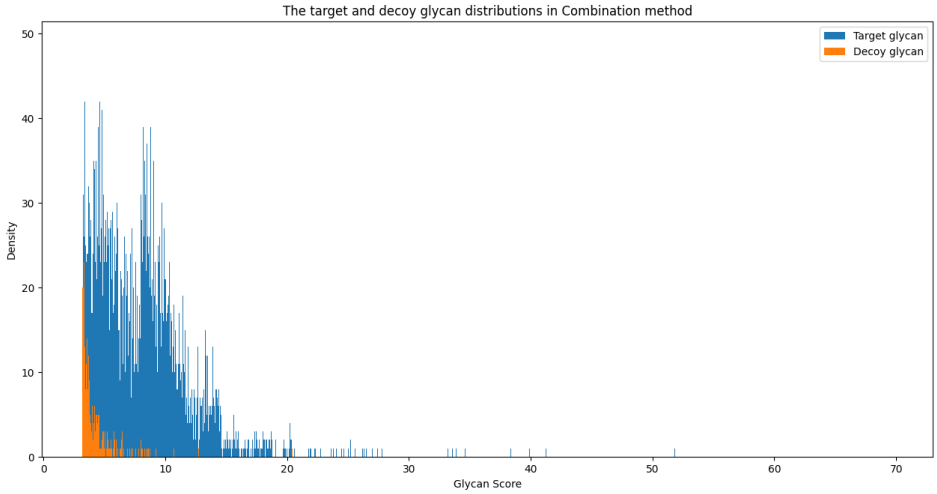
Appendix H The Target and Decoy glycan distributions in Default database (FDR=3%)



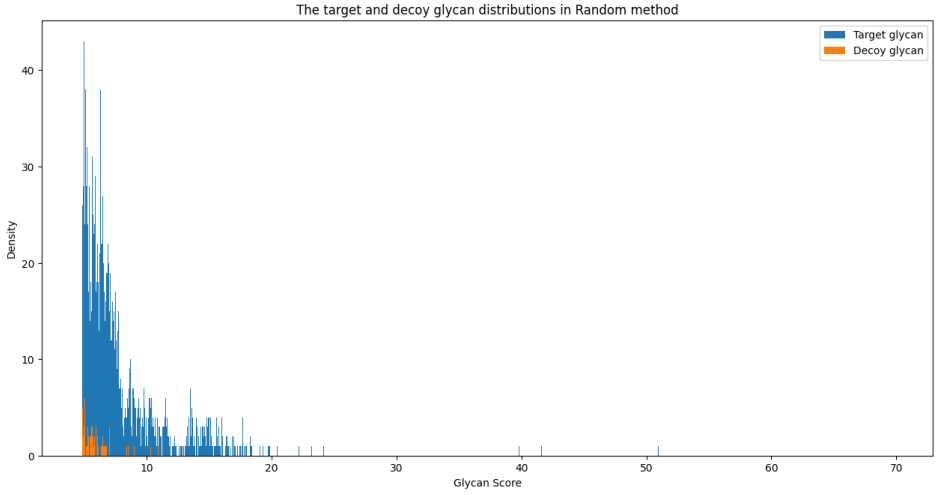
Appendix I The Target and Decoy glycan distributions in Permutation method (FDR=3%)



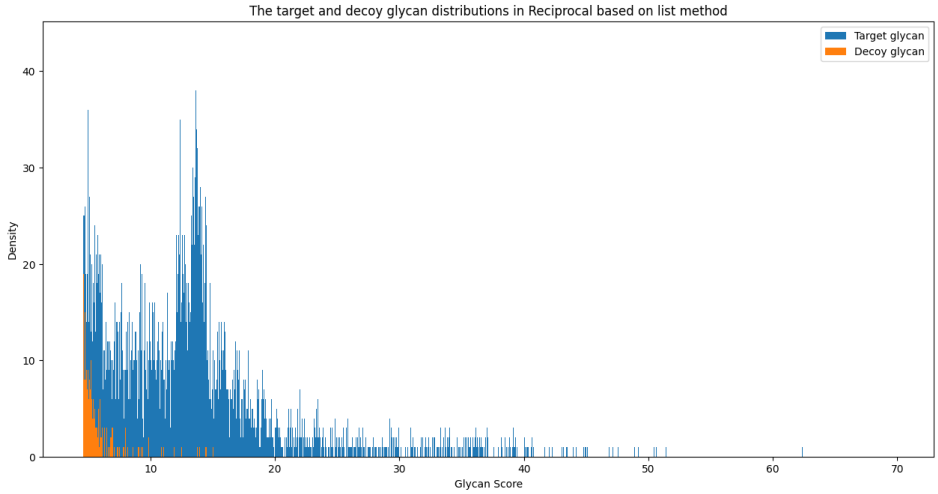
Appendix J The Target and Decoy glycan distributions in Node transfer method (FDR=3%)



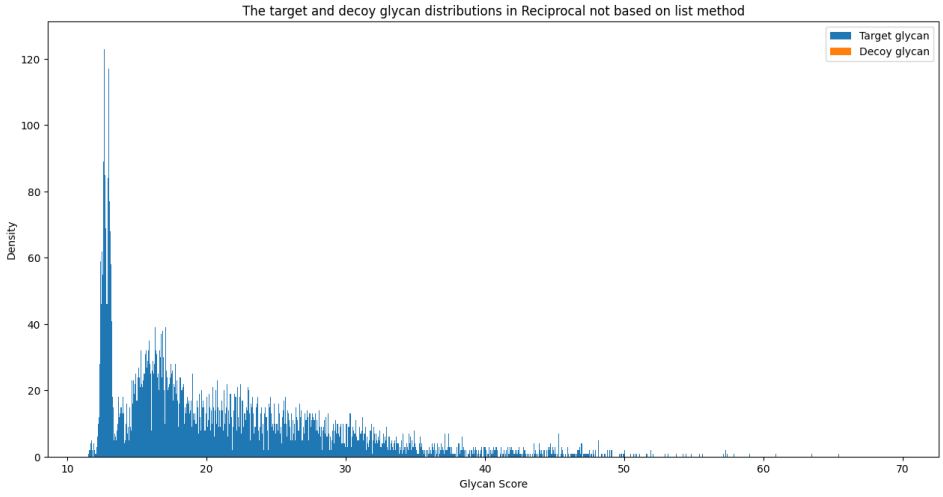
Appendix K The Target and Decoy glycan distributions in Combination method (FDR=3%)



Appendix L The Target and Decoy glycan distributions in Random method (FDR=3%)



Appendix M The Target and Decoy glycan distributions in Reciprocal based on list method (FDR=3%)



Appendix N The Target and Decoy glycan distributions in Reciprocal not based on list (FDR=3%)

Decoy Construction	Number of glycan	Threshold glycan score

Default	19251	0.55
Permutation	19503	0.36
Node transfer	19371	0.21
Combination	19299	0.49
Random	19799	0.23
Reciprocal based on list	19185	0.73
Reciprocal not based on list	18387	11.5

Appendix O The results (FDR=100%)

Decoy Construction	Number of glycan	Threshold glycan score
Default	11130	2.76
Permutation	7829	3.13
Node transfer	5820	2.66
Combination	9545	3.19
Random	3998	4.84
Reciprocal based on list	10006	4.59
Reciprocal not based on list	18387	11.5

Appendix P The results (FDR=3%)

Decoy Construction	Number of glycan	Threshold glycan score
Default	8548	4.27
Permutation	3013	6.68
Node transfer	1154	5.98
Combination	7699	4.48
Random	1666	7.02
Reciprocal based on list	8648	5.84
Reciprocal not based on list	18387	11.5

Appendix Q The results (FDR=1%)

Method name	Default	Permutation	Node transfer	Combination	Random	Reciprocal based on list
π_1	0.493	0.503	0.540	0.485	0.525	0.491
μ_1	1.958	1.260	0.835	1.715	1.085	2.697
σ_1^2	1.071	0.450	0.225	0.794	0.347	1.928
π_2	0.506	0.496	0.459	0.514	0.474	0.508
μ_2	9.014	5.826	3.902	7.661	5.054	13.055

σ_2^2	33.701	15.068	8.556	20.330	12.164	60.409

Appendix R The parameters (2 components)

Method name	Default	Permutation	Node transfer	Combination	Random	Reciprocal based on list
π_1	0.504	0.487	0.483	0.498	0.497	0.538
μ_1	1.893	1.172	0.713	1.687	0.987	2.756
σ_1^2	0.964	0.353	0.129	0.756	0.251	2.093
π_2	0.463	0.438	0.439	0.468	0.428	0.402
μ_2	8.221	4.850	2.931	7.273	4.024	11.949
σ_2^2	12.289	4.935	1.841	10.317	3.388	16.595
π_3	0.031	0.073	0.077	0.032	0.074	0.059
μ_3	24.044	11.235	7.904	16.017	10.138	28.226
σ_3^2	93.774	37.246	21.819	81.747	27.488	90.915

Appendix S The parameters (3 components)

Method name	Default	Permutation	Node transfer	Combination	Random	Reciprocal based on list
π_1	0.282	0.473	0.446	0.296	0.405	0.220
α_1	0.828	5.337	8.045	5.436	5.538	4.128
Θ_1	0.779	-0.510	-0.700	-0.310	-0.590	0.090
μ_1	1.160	0.372	0.203	0.372	0.342	0.551
π_2	0.343	0.217	0.130	0.608	0.277	0.294
μ_2	3.320	6.347	4.989	5.321	4.896	4.469
β_2	1.431	3.373	2.747	3.073	2.693	1.645
π_3	0.336	0.306	0.411	0.082	0.301	0.392
μ_3	10.201	4.954	3.155	9.409	4.149	13.282
σ_3	2.668	1.839	1.244	0.973	1.553	3.045
π_4	0.037	0.001	0.012	0.012	0.015	0.092
μ_4	24.308	15.855	10.578	13.748	14.970	26.453
σ_4	9.427	0.280	0.765	0.394	1.177	9.884

Appendix T The parameters (4 components, cutoff=2.5%)

Method name	Default	Permutation	Node transfer	Combination	Random	Reciprocal based on list
π_1	0.144	0.138	0.099	0.112	0.112	0.087

μ_1	0.954	0.605	0.374	0.853	0.503	1.344
σ_1^2	0.141	0.084	0.058	0.125	0.070	0.205
π_2	0.279	0.356	0.130	0.211	0.287	0.103
α_2	7.758	7.348	7.673	8.369	7.146	6.300
Θ_2	-0.450	-0.650	-0.790	-0.510	-0.710	-0.240
μ_2	0.365	0.330	0.262	0.317	0.319	0.527
π_3	0.447	0.159	0.265	0.581	0521	0.330
μ_3	5.776	7.358	0.932	5.413	3.746	3.963
β_3	2.691	3.591	0.405	3.098	2.028	1.746
π_4	0.099	0.189	0.413	0.082	0.050	0.387
μ_4	12.567	3.887	3.426	9.390	5.192	13.242
σ_4^2	1.010	1.205	1.305	0.976	1.419	3.092
π_5	0.029	0.154	0.091	0.012	0.027	0.090
μ_5	25.992	6.616	8.238	13.748	14.767	26.514
σ_5^2	9.351	1.371	4.723	0.391	2.088	9.893

Appendix U The parameters (5 components)

Curriculum Vitae

Name:	Xiaoou Li
Post-secondary Education and Degrees:	University of Western Ontario London, Ontario, Canada 2016-2020 B.Sc. The University of Western Ontario London, Ontario, Canada 2021-2023 M.Sc.
Related Work Experience	Teaching Assistant The University of Western Ontario 2021-2023