© American Society for Mass Spectrometry, 2017

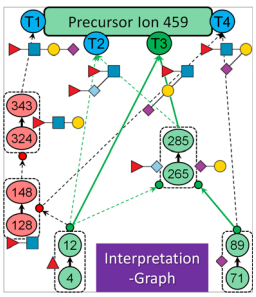
J. Am. Soc. Mass Spectrom. (2017) 28:2288–2301
DOI: 10.1007/s13361-017-1760-6

RESEARCH ARTICLE

GlycoDeNovo – an Efficient Algorithm for Accurate de novo Glycan Topology Reconstruction from Tandem Mass Spectra

Pengyu Hong,[1] Hui Sun,[1] Long Sha,[1] Yi Pu,[2] Kshitij Khatri,[3] Xiang Yu,[3] Yang Tang,[2] Cheng Lin[3]

[1]Department of Computer Science, Brandeis University, Waltham, MA 02453, USA
[2]Department of Chemistry, Boston University, Boston, MA 02215, USA
[3]Department of Biochemistry, Boston University School of Medicine, Boston, MA 02118, USA



Abstract. A major challenge in glycomics is the characterization of complex glycan structures that are essential for understanding their diverse roles in many biological processes. We present a novel efficient computational approach, named GlycoDeNovo, for accurate elucidation of the glycan topologies from their tandem mass spectra. Given a spectrum, GlycoDeNovo first builds an interpretation-graph specifying how to interpret each peak using preceding interpreted peaks. It then reconstructs the topologies of peaks that contribute to interpreting the precursor ion. We theoretically prove that GlycoDeNovo is highly efficient. A major innovative feature added to GlycoDeNovo is a data-driven IonClassifier which can be used to effectively rank candidate topologies. IonClassifier is automatically learned from experimental spectra of known glycans to distinguish B- and C-type ions from all other ion types. Our results showed that GlycoDeNovo is robust and accurate for topology reconstruction of glycans from their tandem mass spectra.

Keywords: De novo glycan sequencing, Machine learning, Electronic excitation dissociation, Fourier-transform ion cyclotron resonance mass spectrometry

Received: 7 April 2017/Revised: 3 July 2017/Accepted: 9 July 2017/Published Online: 7 August 2017

Introduction

Glycosylation is a common modification by which a glycan (or oligosaccharide) is covalently attached to a target biomolecule such as proteins and lipids. It serves important purposes in many biological processes, including protein folding and clearance, cell adhesion, and immunological responses, among others [1, 2]. Glycosylation is one of the key factors that determine the solubility, stability, and efficacy of many biopharmaceuticals [3, 4]. Change in glycosylation pattern is often observed under different disease conditions, such as tumorigenesis [5, 6]. Glycan structural analysis is essential for understanding their diverse roles in biological systems, yet it remains a challenging task, in part due to the vast number of topologies that they may assume even for a moderate-sized glycan. Glycans are tree ensembles of monosaccharides linked via glycosidic bonds. A glycosidic bond is formed via condensation reaction between the hemiacetal group of one monosaccharide (the non-reducing-end residue) and a hydroxyl group of another (the reducing-end residue). Theoretically, there could be up to four branches at any branching point in an oligosaccharide though these seldom occur naturally because of steric hindrance.

Recently, tandem mass spectrometry (MS/MS) has become one of the most powerful tools for elucidating glycan structures [7, 8]. In a tandem MS experiment, a single glycosidic cleavage produces B, C, Y, and Z ions, whereas cross-ring cleavages generate A and X ions (Figure 1a) [9]. Internal fragment ions, or fragment ions with loss of multiple branches may also be formed by two or more glycosidic and/or cross-ring cleavages. Here, we group A and X ions and internal fragment ions into a category termed O ions (i.e., other ions). The glycosidic

Electronic supplementary material The online version of this article (doi:10.1007/s13361-017-1760-6) contains supplementary material, which is available to authorized users.

Correspondence to: Pengyu Hong; *e-mail:* hongpeng@brandeis.edu, Cheng Lin; *e-mail:* chenglin@bu.edu

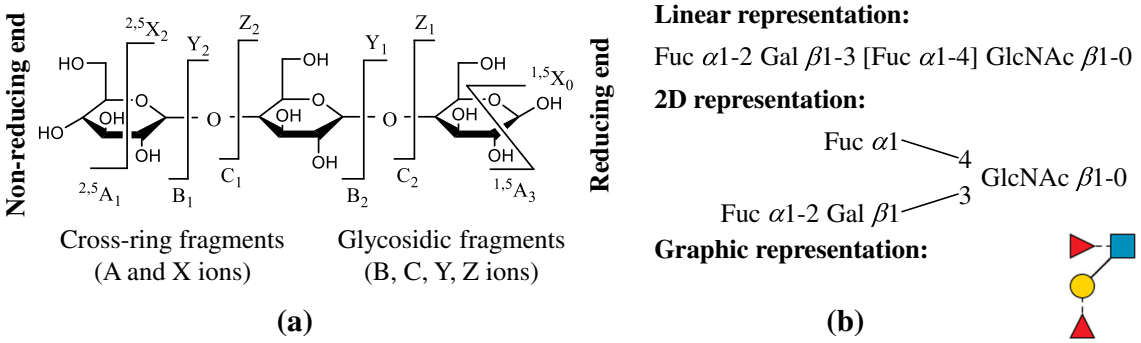


Figure 1. (a) The glycan fragmentation nomenclature system as proposed by Domon and Costello. (b) The linear, 2D, and graphic representations of a glycan (Lewis B)

fragments are important for topology deduction. Since a Y ion differs in mass from its related Z ion by that of a water molecule, as does a B ion from its related C ion, C and Z ions provide redundant information to B and Y ions. A and X ions are useful for deciphering the branching pattern and linkages, as well as for ranking the candidate topologies. The topology of a glycan can be represented as a tree with nodes representing monosaccharide residues and edges representing glycosidic linkages (Figure 1b).

Several tools exist for determining the topologies of glycans by searching their experimental spectra against prebuilt glycan databases [10–13]. The accuracy of the search results depends not only on the quality of the query (i.e., the tandem MS data) but also on the quality and completeness of the databases. To date, glycan databases are often populated with lower-quality spectral data obtained on ion trap and time-of-flight instruments, typically generated by collision-induced dissociation (CID). This can adversely affect the performance of database searching algorithms that identify and score candidate structures based on the similarity of the query to spectra in the database, especially for experimental data generated by radical-induced fragmentation methods, and/or on higher-performance MS instruments. More importantly, because glycan databases are generally incomplete [14], it is necessary to develop a de novo method for determination of glycan structures from their experimental spectra. Given enough information (e.g., precursor ion mass, possible monosaccharide components, charge carrier, and product ion masses), brute-force search methods, such as STAT [15], may be used to exhaustively compare an experimental tandem mass spectrum to those of all possible theoretical structures. However, the number of possible structures increases exponentially as the number of monosaccharides in a glycan increases, and the search space quickly becomes too big to explore for large glycans. Thus, the brute-force approach is feasible only for relatively small glycans. Mizuno et al. proposed to reconstruct glycan topologies by building a relationship tree trying to interpret peaks as Y ions [16], but it is not clear how their method deals with branching except for those within the *N*-glycan core. Ethier et al. improved the relationship tree approach mainly by including more biosynthetic rules [17, 18]. However, our knowledge of biosynthetic rules is incomplete for many organisms, and this limits the general applicability of the above relationship-tree based methods.

Tang et al. proposed an approach termed GLYCH that constructs a set of prefix residue masses (PRMs) for each peak and uses a dynamic programming algorithm to find a series of PRMs for inferring glycan structures from tandem MS spectra [19]. However, its topology scoring method may repeatedly use peaks in scoring a structure, which should be avoided because it favors linear structures over branched ones [20]. Shan et al. showed in theory that generating glycan topology candidates without repeatedly counting peaks (i.e., the Peak Assignment Problem) was an NP-hard problem, and proposed a heuristic algorithm that saves time and space by keeping a fixed number of high-score forests for each peak [21]. Bocker et al. developed an algorithm for solving the Peak Assignment Problem that uses the fixed-parameter tractability concept to restrict the running time, and showed that the complexity of counting the number of rooted trees is polynomial in time and space with respect to the number of monosaccharides and the maximal out-degree [20]. When the number of peaks in a spectrum became too large, they deployed some heuristics to make computation tractable, for example, by restricting the k (e.g., $k = 10$) most intense peaks to be used at most once in scoring candidates, whereas allowing all other peaks to be used multiple times. Sun et al. proposed to reconstruct topologies from the root to leaves by adding a monosaccharide at a time [22], while keeping only a fixed number of topologies, the theoretical spectra of whichbest match the data in each iteration. Dong et al. represented a glycan structure as a directed acyclic graph and developed an algorithm to reconstruct a glycan iteratively by storing all confirmed substructures and using them to build larger substructures [23]. To make computation manageable, they kept a limited number of top-scored substructures (20 in their pseudo codes) in each iteration. They also proposed a data preprocessing method to filter out noisy peaks and a probability-based cleavage method to produce theoretical tandem mass spectra for scoring candidate structures. To circumvent the NP difficulty in the Peak Assignment Problem, Kumozaki et al. [24] applied *Lagrangian* relaxation [25] to turn the Peak Assignment Problem into a relaxed Integer Programming problem, which can then be optimized by dynamic programming and subgradient optimization. They

also proposed to learn how to score structural elements (e.g., branching at a residue, connection between two residues, and cleavage at a residue) from the experimental data.

We present in this paper a novel method, named GlycoDeNovo, for de novo glycan topology reconstruction using tandem MS data. Different from the catalog-library approaches [10–13], GlycoDeNovo does not rely on any database of known glycans and can be used to discover new structures. Given a tandem MS spectrum, it reconstructs the possible glycan topologies in a bottom-up way by building an interpretation-graph that interprets some non-precursor peaks as B or C ions and specifies how to interpret each B or C ion by appending one or more preceding B and/or C ions to a monosaccharide. The computational complexity of the above peak interpretation procedure is $O(N^{H+1})$, where N is the number of peaks in the spectrum and H is the highest number of branching allowed. Hence GlycoDeNovo has significant advantages over other recent de novo glycan sequencing algorithms [20, 23], computational complexities of which are $O(3^N \cdot M^2)$, where M is the precursor ion mass. GlycoDeNovo has the same computational complexity as that of GLYCH [19], but it does not suffer from the problem of double peak counting in scoring candidates. In addition, GlycoDeNovo avoids unnecessary reconstruction of sub-topologies that do not lead to interpretation of the precursor ion. Hence, the constant factor in the computational complexity of GlycoDeNovo is actually much lower than that of GLYCH. This also allows GlycoDeNovo to avoid solving the NP-hard Peak Assignment Problem. It is possible that GlycoDeNovo may misinterpret a peak as a B or C ion when it belongs to a different type. To tackle this problem, GlycoDeNovo learns IonClassifier from experimental data to distinguish B and C ions from other types of ions. IonClassifier greatly improves the accuracy of GlycoDeNovo in ranking candidate topologies. GlycoDeNovo is capable of handling missing cleavages, which happens occasionally in experimental data. In its current setting, GlycoDeNovo can handle missing ions corresponding to gaps of two monosaccharides (i.e., two monosaccharides are needed to link several substructures into a bigger one).

Computational Approach

The pipeline of GlycoDeNovo (Figure 2) works as the following. It first enriches the peak list by adding artificial peaks complementary to those observed. This is necessary because although each glycosidic cleavage could in theory generate a pair of complementary ions, not all fragments are observed in the experimental data due to the lack of charge carrier, secondary fragmentation, or other reasons. Since GlycoDeNovo only attempts to interpret non-reducing-end glycosidic fragments, complementary peaks are computationally added to facilitate topology reconstruction. The second component of GlycoDeNovo reconstructs glycan topologies using the peaks in the enriched list. It calls ***PeakInterpreter*** (see pseudo codes in Algorithm I) to build an interpretation-graph consisting of nodes and edges to respectively represent peaks and how a peak can be interpreted as a B or C ion by using the interpretations of preceding peaks. In each iteration, ***PeakInterpreter*** tries to interpret each peak as a B or C ion by attaching up to four branches to a monosaccharide. The branches are the interpretations of peaks lighter than the one being interpreted. Figure 2b shows an example of the interpretation-graph. GlycoDeNovo then calls ***CandidateSetReconstructor*** (pseudo codes in Algorithm II), which is guided by the interpretation-graph, to recursively reconstruct all candidate topologies of the precursor ion. The detailed calculation process for peak interpretation and topology reconstruction that produced the interpretation-graph in Figure 2 can be found in the Supporting Materials. Finally, GlycoDeNovo learns an ion classifier from a collection of experimental data to score all candidate topologies of the precursor ion. Hereafter, we use the term "topology" loosely to denote also the "partial topologies" or "sub-topologies" of non-precursor ions.

Reconstructing Candidate Topologies

Given a spectrum and a user-defined mass accuracy constraint, ***PeakInterpreter*** builds an interpretation-graph that specifies how to interpret each peak using the topologies of other peaks with lighter masses. ***CandidateSetReconstructor*** takes the interpretation-graph and reconstructs all candidate topologies of the precursor ion that satisfy the user-defined mass accuracy constraint. We first explain the symbols and data structures used in ***PeakInterpreter*** and ***CandidateSetReconstructor***:

- Let $\boldsymbol{G}$ be the set containing all monosaccharide classes of interest. No attempt is made to differentiate isomeric monosaccharides.
- The enriched peak list contains a set of peaks sorted ascendingly by their masses $\{m_1, m_2, \ldots, m_N\}$, where m_N is the observed mass of the precursor ion.
- Let τ be the user-defined mass accuracy.
- Each peak, say the n-th peak, has a candidate set s_n, which is represented as <*peakID*, *cmass*, *lmass*, *hmass*, *topoReconstructionSet*, *topologySet*>, where *peakID* = n, *cmass* = m_n, *lmass* and *hmass*, respectively, are the low- and high-mass bounds of the topologies that can be used to interpret this peak and are stored in *topologySet*, and *topoReconstructionSet* is a set containing information for deriving *topologySet*.
- Each member in s_n.*topoReconstructionSet* is an object *topoReconstruction* = <*root*, *branchSet*, *topologySet*> representing a set of topologies that use the same *root* (a monosaccharide class $\in \boldsymbol{G}$) and choose their branches from *branchSet* (each member in *branchSet* contributes one branch). Each member in *branchSet* is a candidate set of a peak preceding the n-th peak. Basically, each topology in *topoReconstruction.topologySet* chooses one branch from the *topologySet* of each member in *topoReconstruction.branchSet*.
- A topology is represented by a structure <*mass*, *representation*, *supports*>, where *mass* is its theoretical mass, *representation* is a

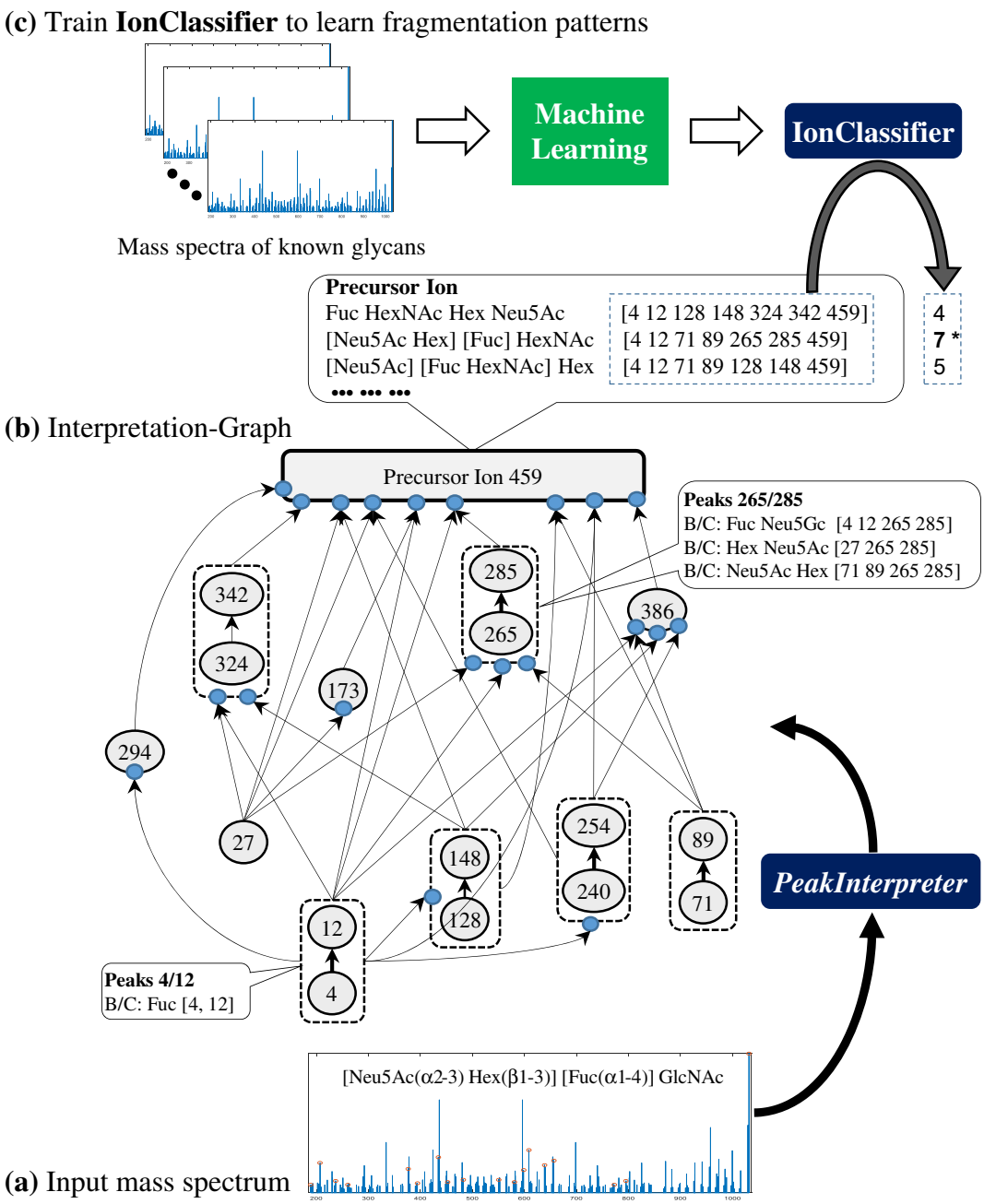


Figure 2. GlycoDeNovo pipeline. **(a)** An example EED spectrum (SLA). Peaks marked by red circles are interpretable by ***PeakInterpreter*** to derive the interpretation-graph shown in **(b)**. The numbers inside the interpretation-graph nodes are peak IDs. Each blue dot associated with a node is a *topoReconstruction* object that specifies how the corresponding peak can be interpreted by other peaks pointing to it. The node-pairs in dashed rounded rectangles are interpreted as B/C-ion pairs. The reconstructed topologies of three peaks (4/12, 265/285, precursor ion) are shown as examples in the rounded rectangle callouts along with their supporting peaks in brackets. This precursor ion has 14 candidate topologies. Based on the supporting peak count, three of them tie for the best and are listed in the call out. **(c)** Machine learning is applied to automatically learn an **IonClassifier** to distinguish B and C ions from other ion types. **IonClassifier** is then used to score topology candidates (see main text for details). In this example, it gives the highest score of 7 to the correct topology [Neu5Ac Hex] [Fuc] HexNAc

text string following the modified IUPAC condensed text nomenclature without linkage information, and *supports* contains peaks in the enriched peak list that can be interpreted as B- or C-type ions and be generated from this topology.

- Let $\boldsymbol{S}$ be the candidate pool containing all non-empty candidate sets.

The current design of ***PeakInterpreter*** allows candidate topologies to have up to four branches at each branching

Algorithm I: $\boldsymbol{S}$ = ***PeakInterpreter***($\{m_1, m_2, \ldots, m_N\}$)

(1) Initialize the candidate pool $\boldsymbol{S} = \{\varnothing\}$.
(2) **for** $n = 1$ **to** N
(3) Initialize the candidate set s_n of the n-th peak: $s_n.cmass = m_n$, $s_n.lmass = m_n - \tau$, $s_n.hmass = m_n + \tau$, $s_n.topoReconstructionSet = \varnothing$, $s_n.topologySet = \varnothing$.
(4) **for** all possible combinations of up to 4 candidate sets $s_a, s_b, s_c, s_d \in \boldsymbol{S}$
(5) Calculate $lm = s_a.lmass + s_b.lmass + s_c.lmass + s_d.lmass$
$hm = s_a.hmass + s_b.hmass + s_c.hmass + s_d.hmass$
δ = mass difference caused by creating a B-ion (or the precursor ion if $n = N$) by linking s_a, s_b, s_c, s_d to a monosaccharide.
(6) **if** $\exists$ g $\in \boldsymbol{G}$ *s.t.* $(lm, hm) = (m_n - \tau, m_n + \tau) \cap (g.mass + lm + \delta, g.mass + hm + \delta) \neq \varnothing$
(7) Create a *topoReconstruction* object $r = <g, \{s_a, s_b, s_c, s_d\}, \varnothing>$, and add r to $s_n.topoReconstructionSet$.
Set $s_n.lmass = \min(s_n.lmass, lm)$ and $s_n.hmass = \max(s_n.hmass, hm)$.
(8) **end**
(9) **end**
(11) **if** $s_n.topoReconstructionSet \neq \varnothing$, add s_n to $\boldsymbol{S}$, **end**
(12) **end**

point, but this constraint can be tightened to allow a lower degree of branching if needed. ***PeakInterpreter*** maintains a candidate pool $\boldsymbol{S}$. Each candidate serves as a potential building block for interpreting a heavier peak. ***PeakInterpreter*** starts from the lightest peak and tries to interpret every peak as a B ion, C ion, or the precursor ion by searching for all allowable combinations of building blocks in the candidate pool $\boldsymbol{S}$ (steps 4–9) that can be appended to a monosaccharide g to obtain a candidate set with mass within the accuracy range specified by τ. The mass difference δ in step 5 depends on the ion type and the glycan derivatization method employed (e.g., permethylation). We use the intensities of non-precursor peaks interpretable by ***PeakInterpreter*** to normalize the intensities of all peaks into z-scores [26].

We do not need to reconstruct the topologies (i.e., $s_n.topologySet$) at this step. Topology reconstruction will be done later by calling ***CandidateSetReconstructor*** after ***PeakInterpreter*** terminates. Although ***PeakInterpreter*** does not have the accurate mass of each candidate topology that is yet to be reconstructed, the test performed at step 6 gives an estimate of the mass range tight enough to include all true positives, but it may also include a small number of false positives (i.e., topologies with masses outside of the accuracy range). Because each interpreted peak is still represented as one yet-to-be-reconstructed candidate set, the false positives will not increase the computational complexity, and they will be removed later by ***CandidateSetReconstructor***.

Theorem The complexity of building an interpretation-graph is $O(|G| \times N^{H+1})$, where **G** is the monosaccharide set, N is the number of peaks in the given spectrum, and H ≤ 4 is the maximal branching number permitted.

Proof The computation of ***PeakInterpreter*** mainly resides in the *for-loop* between steps 4 and 9 complexity of which is $O(|G| \times |S^{(n)}|^H)$, where $\boldsymbol{S}^{(n)}$ is the value of the candidate pool $\boldsymbol{S}$ at the n-th loop and $|\boldsymbol{S}^{(n)}|$ is the size of $\boldsymbol{S}^{(n)}$ (i.e., the number of interpretable peaks up to the n-th loop). The overall complexity of ***PeakInterpreter*** is $O(|G| \times \sum_{n=1}^{N} |S^{(n)}|^H)$. Since $|S^{(n)}| \leq n$, $O(|G| \times \sum_{n=1}^{N} |S^{(n)}|^H) = O(|G| \times \sum_{n=1}^{N} n^H) = O(|G| \times N^{H+1})$.

Comment In practice, we found that most peaks cannot be interpreted so that $|\boldsymbol{S}^{(n)}|$ is often much smaller than n. Therefore, the empirical complexity of ***PeakInterpreter*** has a small constant in $O(|G| \times N^{H+1})$.

After obtaining the interpretation-graph, GlycoDeNovo passes the candidate set object of the precursor ion into ***CandidateSetReconstructor*** to reconstruct all legal candidate topologies. ***CandidateSetReconstructor*** first checks if each *topoReconstruction* object r in the input candidate set s has been reconstructed. If not, it recursively calls itself to reconstruct all branches of r. Then ***CandidateSetReconstructor*** creates all legal topologies of r (steps 11–19), which are rooted at *r.root* and satisfy the mass accuracy constraint. At step 14, the branches are linked by their alphabetic order to *r.root* so that isomorphic topologies can be effectively detected and removed at step 16. The union operation at step 15 effectively and efficiently solves the problem of repeated counting of supporting peaks, which was a problem in GLYCH [19]. Finally, at step 19, the candidate topology set of r is added to that of s. ***CandidateSetReconstructor*** runs extremely fast, and its running time is negligible compared with that of ***PeakInterpreter***.

Algorithm II: *CandidateSetReconstructor*(*s*)

```
(1)  if s.topologySet ≠ ∅
(2)      return // s has been reconstructed.
(3)  end
(4)  for each r ∈ s.topoReconstructionSet
(5)      if r.topologySet ≠ ∅
(6)          continue // r has been reconstructed
(7)      end
(8)      for each branch ∈ r.branchSet
(9)          CandidateSetReconstructor( branch )
(10)     end
(11)     for each of all possible branch combinations (a combination is formed by choosing one
                 topology from the topologySet of each s ∈ r.branchSet)
(12)         Calculate tmass = total mass of the topology with the chosen branches linked to r.root.
(13)         if tmass ∈ (massLow, massHigh)
(14)             Create a topology t by linking the chosen branches to r.root, let t.mass = tmass.
(15)             t.supports = {peakID} ∪ {peak supports of t's branches}.
(16)             Add t to r.topologySet.
(17)         end
(18)     end
(19)     Add r.topologySet to s.topologySet.
(20) end
```

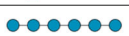
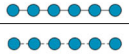
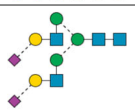
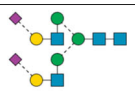
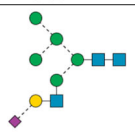
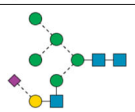
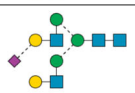
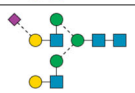
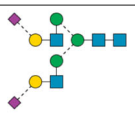
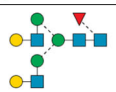
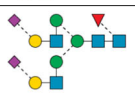
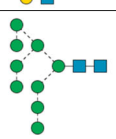
One of the major differences between GlycoDeNovo and previous de novo approaches [19–21, 23, 27] is that it uses the mass range to confine the search space within the experimental mass accuracy window without reconstructing any topology during the peak interpretation process. GlycoDeNovo delays topology reconstruction until it finishes deriving the interpretation group of the precursor ion, and hence it only needs to reconstruct topologies that are required to interpret the precursor ion. In our experiments, since most of the partial topologies did not lead to precursor ions, this simple strategy dramatically reduced the computational time and space. GlycoDeNovo starts from the non-reducing end to incrementally build up interpretations of B and C ions because (1) glycosidic fragments are in general substantially more likely to be observed than cross-ring fragments; and (2) Y and Z ions provide redundant mass information to B and C ions, and even in cases where only Y and/or Z ions are observed at a cleavage site, their information is recaptured in the enriched peak list. This strategy is different from the one used by Mizuno et al. [16] and by Sun et al. [22] that start the reconstruction procedure from the reducing end. Growing topologies from the reducing end may run into difficulties when dealing with branching points where each of the branches contain more than one monosaccharide residue. In such a scenario, some of the reconstructed topologies can correspond to internal fragments, which are more likely to be missing in data, thus making it difficult to evaluate those topologies.

To handle the problem of missing peaks, we made one modification to ***PeakInterpreter*** so that it will consider monosaccharide pairs in addition to individual monosaccharides at step 6. Basically, for each possible ordered pair of monosaccharides $[g_1, g_2]$ satisfying the mass accuracy constraint, we can expand the interpretation graph by (1) creating a *topoReconstruction* object r_1 that links s_a, s_b, s_c, and s_d to g_2 and then another *topoReconstruction* object r_2 that links r_1 to g_1 or (2) for each s in $\{s_a, s_b, s_c, s_d\}$, creating a *topoReconstruction* object r_1 that links s to g_2 and then another *topoReconstruction* object r_2 that link $r_1 \cup (\{s_a, s_b, s_c, s_d\} - s)$ to g_1. Obviously, allowing missing peaks greatly increases the search space. Therefore, we suggest turning this option on only when no topology can be found without considering missing cleavages. Biosynthetic rules (e.g., the chitobiose *N*-glycan core) can also be incorporated to constrain the search space of ***PeakInterpreter***.

Scoring Topologies via Machine Learning

Mass spectrometry data can be noisy. In addition, the presence of internal fragments can greatly complicate the de novo topology reconstruction process. ***PeakInterpreter*** may misinterpret some Y, Z, or O ions as B or C ions and generate ambiguities. Misinterpretation may lead to false topologies being ranked as high as or better than the correct topology based on the supporting peak count alone. To tackle this problem, we applied machine learning to build an IonClassifier for distinguishing B and C ions from other ion types (Figure 2c). IonClassifier takes a peak and its context, currently defined as the neighboring peaks within a predetermined mass-difference

Table 1. Glycan Standards Used in This Study

Short Name	Formula	Structure (CFG with linkage placement notation)
SLA	[Neu5Ac(α2-3) Gal(β1-3)] [Fuc(α1-4)] GlcNAc	
SLX	[Neu5Ac(α2-3) Gal(β1-4)] [Fuc(α1-3)] GlcNAc	
Lewis B	[Fuc(α1-2) Gal(β1-3)] [Fuc(α1-4)] GlcNAc	
Lewis Y	[Fuc(α1-2) Gal(β1-4)] [Fuc(α1-3)] GlcNAc	
LNT	Gal(β1-3) GlcNAc(β1-3) Gal(β1-4) Glc	
LNnT	Gal(β1-4) GlcNAc(β1-3) Gal(β1-4) Glc	
LNFP I	Fuc(α1-2) Gal(β1-3) GlcNAc(β1-3) Gal(β1-4) Glc	
LNFP II	[Gal(β1-3)] [Fuc(α1-4)] GlcNAc(β1-3) Gal(β1-4) Glc	
LNFP III	[Gal(β1-4)] [Fuc(α1-3)] GlcNAc(β1-3) Gal(β1-4) Glc	
CelHex	Glc(β1-4) Glc(β1-4) Glc(β1-4) Glc(β1-4) Glc(β1-4) Glc	
MalHex	Glc(α1-4) Glc(α1-4) Glc(α1-4) Glc(α1-4) Glc(α1-4) Glc	
N002	[Neu5Ac(α2-3) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [Neu5Ac(α2-3) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N003	[Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N012	[Neu5Ac(α2-3) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [[Man(α1-3)] [Man(α1-6)] Man(α1-6)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N013	[Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [[Man(α1-3)] [Man(α1-6)] Man(α1-6)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N222	[Neu5Ac(α2-3) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] [Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N223	[Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] [Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
N233	[Neu5Ac(α2-3) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] Man(β1-4) GlcNAc(β1-4) GlcNAc	
NA2F	[Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] [Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] Man(β1-4) GlcNAc(β1-4) [Fuc(α1-6)] GlcNAc	
A2F	[Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-3)] [Neu5Ac(α2-6) Gal(β1-4) GlcNAc(β1-2) Man(α1-6)] Man(β1-4) GlcNAc(β1-4) [Fuc(α1-6)] GlcNAc	
Man9	[[Man(α1-2) Man(α1-6)] [Man(α1-2) Man(α1-3)] Man(α1-6)] [Man(α1-2) Man(α1-2) Man(α1-3)] Man(β1-4) GlcNAc(β1-4) GlcNAc	

window (e.g., 105 Da), and classifies the peak as +1 (i.e., a B or C ion) or –1 (i.e., a non-B or C ion). The neighboring peaks can be expressed as an array of contextual features (i.e., mass shifts from the peak of interest). The final score of a candidate topology is calculated by summing up the IonClassifier values of its supporting peaks. IonClassifier is trained by boosting [28] the decision tree classifier [29] on the experimental tandem mass spectra of a set of known glycans. For each glycan standard, we can match its theoretical spectrum to the experimental spectrum to collect the observed context of each theoretical peak found in the experimental spectrum. We grouped the supporting peaks of candidates into true B ions, true C ions, true Y ions, true Z ions, and O ions, and trained IonClassifier to distinguish true B,ions and true C ions from Y, Z, and O ions. If a supporting peak is interpreted by ***PeakInterpreter*** as a B ion, it will be validated by the B-ion classifier of IonClassifier. Similarly, if a supporting peak is interpreted by ***PeakInterpreter*** as a C ion, it will be validated by the C-ion classifier of IonClassifier.

Experimental

Although GlycoDeNovo can handle glycans containing residue(s) with up to four branches, its performance was tested only on bifurcated structures due to the availability of glycan standards. The structures of glycans used in our study are listed in Table 1.

Materials

Sialyl lewis A (SLA), sialyl lewis X (SLX), lewis B, lewis Y, lacto-*N*-tetraose (LNT), and lacto-*N*-neotetraose (LNnT) were purchased from Dextra Laboratories (Reading, UK). Lacto-*N*-fucopentaose (LNFP) I, II, and III were acquired from V-LABS, Inc. (Covington, LA, USA). Cellohexaose (CelHex), maltohexaose (MalHex), A2F, and NA2F glycans were purchased from Carbosynth Ltd. (Berkshire, UK). Synthetic *N*-linked glycan standards (N002 to N233) were obtained from Chemily Glycoscience (Atlanta, GA, USA). Man9 *N*-glycan, $H_2{}^{18}O$ (97%) water, 2-aminopyridine, acetic acid, dimethyl sulfoxide (DMSO), sodium hydroxide, methyl iodide, chloroform, sodium borodeuteride, and cesium acetate were purchased from Sigma-Aldrich (St. Louis, MO, USA). Pierce PepClean C18 spin columns were acquired from ThermoFisher Scientific.

Sample Preparation

For reducing-end ^{18}O-isotope labeling, each dry native glycan (5 μg) was dissolved in 20 μL of $H_2{}^{18}O$ to which 2 μL of catalyst solution (2.7 mg/mL 2-aminopyridine in anhydrous methanol) and 1 μL of acetic acid were added. The reaction mixture was incubated at 65 °C for 16 h. Solvent was removed by a SpeedVac concentrator before permethylation. For deutero reduction, approximately 10 μg each of glycan standards were incubated with 0.5 M sodium borodeuteride in 0.2 M ammonium hydroxide solution for 2 h at room temperature while mixing, followed by drop-by-drop addition of acetic acid (10%) until bubbling stopped. The reaction mixture was dried down in a centrifugal evaporator. Excess borates were removed by repeated resuspension and drying of the samples in methanol. Permethylation was performed according to the method described previously [30, 31]. Briefly, the underivatized, ^{18}O-labeled, or deutero-reduced glycan was suspended in 100 μL of DMSO/NaOH solution and gently vortexed for 1 h at room temperature. Methyl iodide (50 μL) was added to the reaction mixture and the reaction was allowed to proceed for another 1 h at room temperature in the dark. Additional NaOH/DMSO (100 μL) and methyl iodide (50 μL) were added together followed by 1 h of vortexing. This process was repeated up to five times to ensure complete methylation before the reaction was terminated by addition of 200 μL of chloroform and 200 μL of water. Permethylated glycans were extracted by liquid–liquid fractionation in water and chloroform, and desalted using PepClean C18 spin columns.

Mass Spectrometry Analysis

Permethylated glycans were dissolved to a concentration of 2–5 μM in 50/50 (v/v) methanol/water solution that also contains 20–50 μM of sodium hydroxide or cesium acetate to produce sodium or cesium adducts of permethylated glycans. For electronic excitation dissociation (EED) analysis, each glycan sample was loaded onto a pulled glass capillary tip with a 1-μm orifice diameter and directly infused into a solariX hybrid Qh-Fourier transform ion cyclotron resonance (FTICR) mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with a hollow cathode dispenser. Sodiated or cesiated precursor ions were isolated by the quadrupole mass filter, externally accumulated in the collision cell, and fragmented in the ICR cell by irradiation of electrons for up to 1 s, with the cathode bias voltage set at –14 V and the ECD lens voltage at –13.95 V. Each transient was recorded at a 0.55-s length, and up to 40 transients were summed for improved *S/N* ratio. Peak picking and deconvolution were achieved with the DataAnalysis software (Bruker Daltonics), using the SNAP algorithm [32] with the quality factor threshold set at 0.01, *S/N* threshold set at 2. All tandem MS spectra were internally calibrated with several fragment ions assigned with high confidence to give a typical mass accuracy of <2 ppm.

Results and Discussions

Experimental Considerations

The output accuracy of a computer analysis is intimately tied to the quality of the input data. For the task at hand, the quality of the glycan tandem mass spectral data is characterized by its

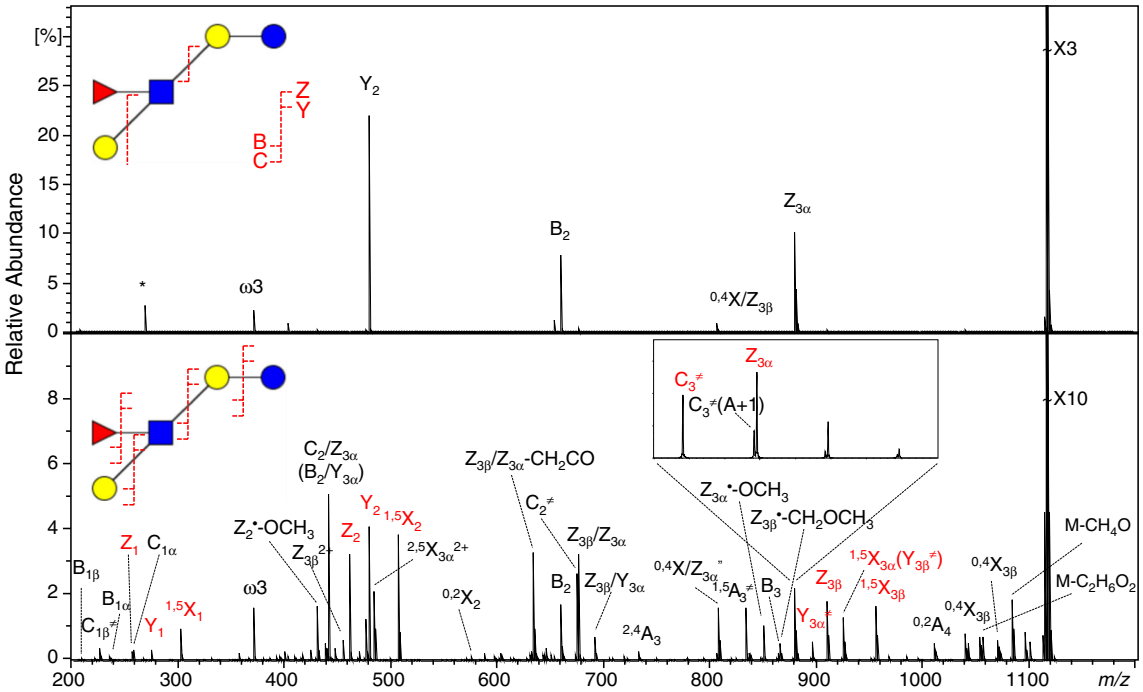


Figure 3. CID (38 eV, top panel) and EED (16 eV, bottom panel) tandem mass spectra of deutero-reduced and permethylated lacto-*N*-fucopentaose II (LNFP II, $[M + Na]^+$). ≠ indicates double hydrogen losses, and " indicates double hydrogen gains. Cleavage maps are shown on the top left corner of each panel. The inset shows the zoomed-in region where C_3-2H and $Z_{3\alpha}$ ions are present, highlighting the importance of the mass resolving power for accurate peak picking. Peaks labeled in red illustrate the contextual feature of Z/Y/^{1,5}X triplets. A complete list of all assigned peaks can be found in Supporting Table S1

cleavage coverage and the data ambiguity. Although GlycoDeNovo can analyze spectral data with missing cleavage(s) by considering addition of two monosaccharide residues at a time during the peak interpretation and topology reconstruction steps, such a practice inevitably increases the computational cost by effectively making $|\boldsymbol{G}|$ larger, while leaving part of the glycan sequence undetermined. Thus, complete sequence determination requires glycosidic cleavage at every linkage site. However, the prevailing glycan fragmentation method to date, CID, often fails to produce a complete series of glycosidic cleavages. Lately, a number of radical-induced dissociation methods have been applied to structural analysis of glycans, many of which were capable of producing more extensive sequence information than CID [33–43]. Among them, the recently developed EED is a particularly powerful method, as it can generate rich structural information for glycan characterization, including linkage differentiation, for a wide variety of glycans, with or without derivatization [40, 41, 44, 45]. Figure 3 shows the CID and EED spectra and cleavage maps of deutero-reduced and permethylated LNFP II, $[M + Na]^+$, with all assigned peaks listed in Supporting Table S1. Whereas CID failed to cleave between the Fuc and GlcNAc residues, and between the reducing-end Gal and Glc residues, EED generated complete sets of B, C, Y, and Z ions. Since complete elucidation of the glycan topology requires cleavages of all glycosidic bonds, the performance of GlycoDeNovo was initially evaluated on EED spectra of glycan standards.

Data ambiguity can arise from several origins. A common confounding factor in de novo glycan sequencing is the presence of internal fragments that may be misinterpreted as a terminal glycosidic fragment with the same saccharide composition. Permethylation is a useful strategy for differentiating terminal and internal fragments based on the number of unmethylated "scars" generated by each glycosidic cleavage. Therefore, all glycans analyzed here were permethylated before tandem MS analysis. Another challenge is that B and Z ions, as well as C and Y ions, are isomeric if they contain the same set of monosaccharide residues. This symmetry may be broken by ^{18}O-stable isotope labeling, leading to a mass shift of 2.004 Da for all reducing-end fragments. However, because typical ^{18}O-labeling conditions can lead to facile loss of sialic acid residues, deutero-reduction was performed as an alternative for glycans containing sialic acid residues, which introduced a 17.038-Da mass shift to all reducing-end fragments. A third complicating factor is that glycans are typically analyzed as metal adducts to minimize proton-mediated gas-phase structural rearrangement [46, 47], yet the number of metal cations in a fragment ion does not always equal to its charge state. Whereas it is possible to expand the peak list by assigning a fragment ion in n+ charge state with either n-1, n, or n+1 (if n is less than the precursor ion charge state) metal cations, this practice not only dramatically increases the computational time by increasing N, but also increases the chance of spurious matches. Since analysis of glycans adducted with a metal cation having a large mass

defect can facilitate metal counting [41], the performance of GlycoDeNovo on EED spectra of both sodiated and cesiated glycans was evaluated here. Finally, glycan tandem mass spectra, especially those generated by EED, can be extremely complex. All experimental data here were acquired on an FTICR instrument, as the high mass accuracy measurement it affords is essential for reducing the chance of fortuitous matches due to the presence of isobaric (but not isomeric) fragments.

Topology Reconstruction

The test results for reducing-end modified glycans are summarized in Table 2. The number of peaks in the enriched spectrum ranged from 216 to 2683. The percentage of interpretable peaks ranged from ~4.4% to ~23.2%, but the percentage of reconstructed peaks was substantially lower, ranging from ~1% to ~5.7%, because GlycoDeNovo only needed to build small interpretation-graphs and reconstruct the topologies of a small number of peaks. These numbers confirmed the computational advantage of the strategy used by GlycoDeNovo to first build the interpretation-graph and delay topology reconstruction after interpreting the precursor ion. For example, the largest peak list (from the EED spectrum of a synthetic *N*-glycan standard of the hybrid type, N012) contained 2683 peaks with 273 interpretable as non-reducing end glycosidic fragments, only 50 of which needed to be reconstructed.

As the masses used in the GlycoDeNovo algorithm were those of the singly protonated species, the *m/z* values of peaks found in the experimental spectrum, typically those of metal-adducts, needed to be converted first. To reduce the run time and to minimize spurious matches, we assumed that the number of metal cations in a given fragment is the same as its charge state. Although this may not be the case for all fragment ions, we asserted that the presence of nonconforming fragments would not prevent reconstruction of the correct topology so long as at least one fragment ion produced by each glycosidic cleavage carried the same number of metal cations as its charge state. This appeared to be a reasonable assumption, since the correct topologies were recovered in all cases studied. The nature of the metal charge carriers did not seem to have a major impact on the accuracy of topology reconstruction.

Ultimately, the performance of a de novo glycan sequencing algorithm should be judged by not only whether it is capable of deducing the correct topology, but also how the correct topology is ranked among all candidate structures. Although

Table 2. Experimental Results

Glycan	REM	Metal	#Peaks	#Interpretable	#Reconstructed	#Candidates	Rank by SPN	Rank by IonClassifier
Lewis B	O18	Cs	329 (133)	18	6	2	1 (0)	1 (0)
Lewis B	O18	Na	216 (76)	24	8	4	1 (0)	1 (0)
Lewis Y	O18	Cs	461 (193)	28	8	4	1 (0)	1 (0)
Lewis Y	O18	Na	283 (105)	26	6	2	1 (0)	1 (0)
LNFP I	O18	Cs	469 (209)	45	19	16	1 (1)	**1 (0)**
LNFP I	O18	Na	516 (224)	23	11	13	1 (4)	**1 (0)**
LNFP II	O18	Cs	390 (178)	26	14	16	5 (0)	**1 (0)**
LNFP II	O18	Na	534 (245)	32	12	1	1 (4)	**1 (0)**
LNFP III	O18	Cs	471 (212)	24	11	10	5 (3)	**1 (0)**
LNFP III	O18	Na	477 (210)	21	13	17	3 (2)	**1 (0)**
LNFP II	D-R	Na	546 (232)	50	16	13	1 (2)	**1 (0)**
NA2F	O18	Na	2389 (1109)	395	24	22	5 (5)	**1 (1)**
Man9	O18	Na	2532 (1182)	588	101	1870	205 (563)	**1 (4)**
A2F	Red	Na	2646 (1222)	597	151	990750	207829 (201169)	**1 (1)**
A2F	D-R	Na	914 (435)	71	25	37	5 (5)	**1 (1)**
N002	D-R	Na	2320 (1063)	262	52	116290	26628 (19903)	**1 (0)**
N003	D-R	Na	1571 (731)	175	49	834	599 (80)	**1 (0)**
N012	D-R	Na	2683 (1229)	273	50	4619	25 (79)	**1 (0)**
N013	D-R	Na	2544 (1179)	351	48	2385	7 (5)	**2 (0)**
N222	D-R	Na	953 (411)	78	18	34	1 (0)	1 (0)
N223	D-R	Na	2674 (1189)	226	30	1577	1 (0)	1 (0)
N233	D-R	Na	2326 (1078)	234	33	1920	568 (420)	**1 (0)**
Lewis B	None	Na	218 (91)	30	9	4	1 (1)	**1 (0)**
LNT	None	Na	317 (126)	21	7	5	1 (1)	**1 (0)**
LNnT	None	Na	270 (105)	23	9	5	1 (1)	**1 (0)**
SLA	None	Na	459 (195)	48	17	14	1 (2)	**1 (0)**
SLX	None	Na	333 (125)	55	18	22	1 (2)	**1 (0)**
CelHex	None	Na	412 (166)	47	11	11	1 (0)	1 (0)
MalHex	None	Na	468 (207)	58	18	22	1 (0)	1 (0)

All glycans are permethylated. The "**REM**" column indicates the type of reducing end modifications (O18 = ^{18}O-labeled, D-R = deutero-reduced, Red = reduced). The "**#Peaks**" column lists the number of peaks in each enriched spectrum with the number of complementary peaks inside the parentheses. The "**#Interpretable**" column lists the number of peaks that can be interpreted as B or C ions by ***PeakInterpreter***. The "**#Reconstructed**" column lists the number of peaks reconstructed by ***CandidateSetReconstructor***. The "**#Candidates**" column lists the number of reconstructed topology candidates. The "**Rank by SPN**" and "**Rank by IonClassifier**" columns list the rank of the true topology among all inferred candidates using their supporting peaks and IonClassifier, respectively. The number inside the parenthesis is the number of other candidates that were ranked the same as the true topology. Cells containing **bold** text in the last column indicate improved ranking by IonClassifier

experimental measures, such as permethylation, reducing-end isotope labeling, and high-mass-accuracy measurement, may be taken to improve the accuracy of ranking by reducing the data ambiguity, it is not always feasible to perform all these procedures experimentally. For example, reducing-end isotope labeling is only applicable towards glycans with a free reducing end, and not suitable for *O*-linked glycans released via reductive β-elimination that result in a reduced reducing end. The experimental strategies and necessary modifications to the GlycoDeNovo algorithm to allow its effective application to analysis of native (as in not permethylated) glycans are beyond the scope of the current study, and will be addressed in a later report. Here, we focus our discussion on the influence of mass accuracy and reducing-end modification on the performance of GlycoDeNovo.

The results presented in Table 2 were obtained with the mass tolerance set to 5 ppm, which was considerably higher than the typical mass accuracy (<1–2 ppm) achieved here (see, for example, Supporting Table S1). Nonetheless, the 5 ppm mass tolerance was chosen because it is easily attainable, even without internal calibration, thus allowing realistic performance evaluation since internal calibration is not always possible, especially for unknown structures and/or for LC-MS/MS data. We note that the 5 ppm mass tolerance was sufficiently tight for differentiating the most common isobar in glycan tandem mass spectra, with Δm of 0.036 Da (CH_4 versus O), for fragment ions with a mass of up to ~4000 Da. As demonstrated by the numbers in the "Rank by SPN (number of supporting peaks)" column in Table 2, our algorithm performed fairly well for small glycans, including Lewis antigens, human milk oligosaccharides (LNFP's, LNT, and LNnT) and linear hexasaccharides. In most cases, the correct topology was ranked the highest, either by itself or with a small number (≤2) of other structures. For larger synthetic *N*-linked glycan standards, the accuracy of SPN ranking is very inconsistent, with the rank of the true topology ranging from 1 (0) out of 1577 candidate structures (N223, deutero-reduced) to 207829 (201169) out of 990750 candidates (A2F, reduced), where the number inside the parenthesis following the candidate rank indicates the number of other candidates that were ranked the same as the true topology. One way to improve the ranking accuracy is to enforce the biosynthetic rules. For *N*-glycans, when only candidate structures containing the pentasaccharide core ($Man_3GlcNAc_2$) were considered, the rank of true topologies greatly improved. For example, the number of candidates dropped to 52 from 4619 for N012, with the true topology now ranked at third with four other structures; for Man9, the rank of true topology was promoted to 1 (4) out of 6 from 205 (563) out of 1870. However, sequencing with biosynthetic rules enforced is no longer truly de novo, and incapable of discovering unusual structures. Furthermore, even with this option turned on, the SPN ranking for some *N*-glycans remains unsatisfactory. For instance, the rank of the true topology for N233 was 29 (2) out of 32, which is the worst in the shrunk candidate pool. Clearly, there is a need to develop a better scoring method for ranking candidate structures. In the next section, we will demonstrate that IonClassifier gives much better performance by utilizing the peak context information.

Candidate Ranking by IonClassifier

The analysis result of A2F (reduced, Na^+-adduct) offers a perfect example to showcase the utility of IonClassifier in candidate ranking. It should come as no surprise that a large number of candidate topologies (990,750) were derived by GlycoDeNovo for this 12-residue complex *N*-glycan (the largest studied here) without a reducing-end label, the enriched peak list of which contains 2646 peaks. When ranked by SPN alone, the true topology was placed at the 207,829th along with 201,169 other candidates. This is because ***PeakInterpreter*** misinterpreted 97 peaks as B or C ions. For example, the peak at *m/z* 406.2071 was misinterpreted as a B ion, "Neu5Gc", which was used to support 34,741 candidates ranked higher than the true topology; the peak at *m/z* 464.249 was misinterpreted as a B ion with two possible topologies, "Hex HexNAc" and "HexNAc Hex", which supported 139,971 candidates ranked higher than the true topology. IonClassifier was able to recognize these peaks as non-B or C ions, and rank the true topology at 1st based on the cumulative IonClassifier values of all its supporting peaks. The use of IonClassifier can also boost the ranking of the true topology for glycans with a reducing-end isotope label. For example, ranking by IonClassifier promoted the correct topology of ^{18}O-labeled Man9 *N*-glycan (Na^+-adduct) from the 205th to the 1st with four other structures; it also ranked the true topology of every ^{18}O-labeled LNFP glycan as the top candidate by itself. Notably, this superior performance of IonClassifier was achieved without enforcing the biosynthetic rules.

Importantly, IonClassifier can be very useful for ranking topologies for glycans without any reducing-end modification (including reduction), where misinterpretation of a Y ion as a C ion or a Z ion as a B ion cannot be avoided based on the accurate mass measurement alone. We recognized that the context for a C ion and that for a Z ion can be very different. For example, a C ion may be accompanied by a ^{1,5}A ion that is 46.005 Da lighter, whereas a Z ion may be accompanied by a ^{1,5}X ion that is 27.995 Da heavier. The topology reconstruction results for glycans without any reducing-end modification are shown in the last seven rows of Table 2. For symmetric linear structures, such as cellohexaose and maltohexaose, the peak lists for C and Y ion series are identical, so are those for B and Z ion series, thus there is no need to differentiate C and Y or B and Z ion pairs. Consequently, ranking by SPN was sufficient to place the correct topology as the top-ranked candidate by itself. For asymmetric linear structures (e.g., LNT) and for branched structures (e.g., SLA), ranking by SPN often resulted in several structures (including the correct one) sharing the top rank because of its inability to differentiate C and Y, or B and Z ion pairs. When ranked by IonClassifier, however, the correct topology was always ranked the highest by itself. This result is

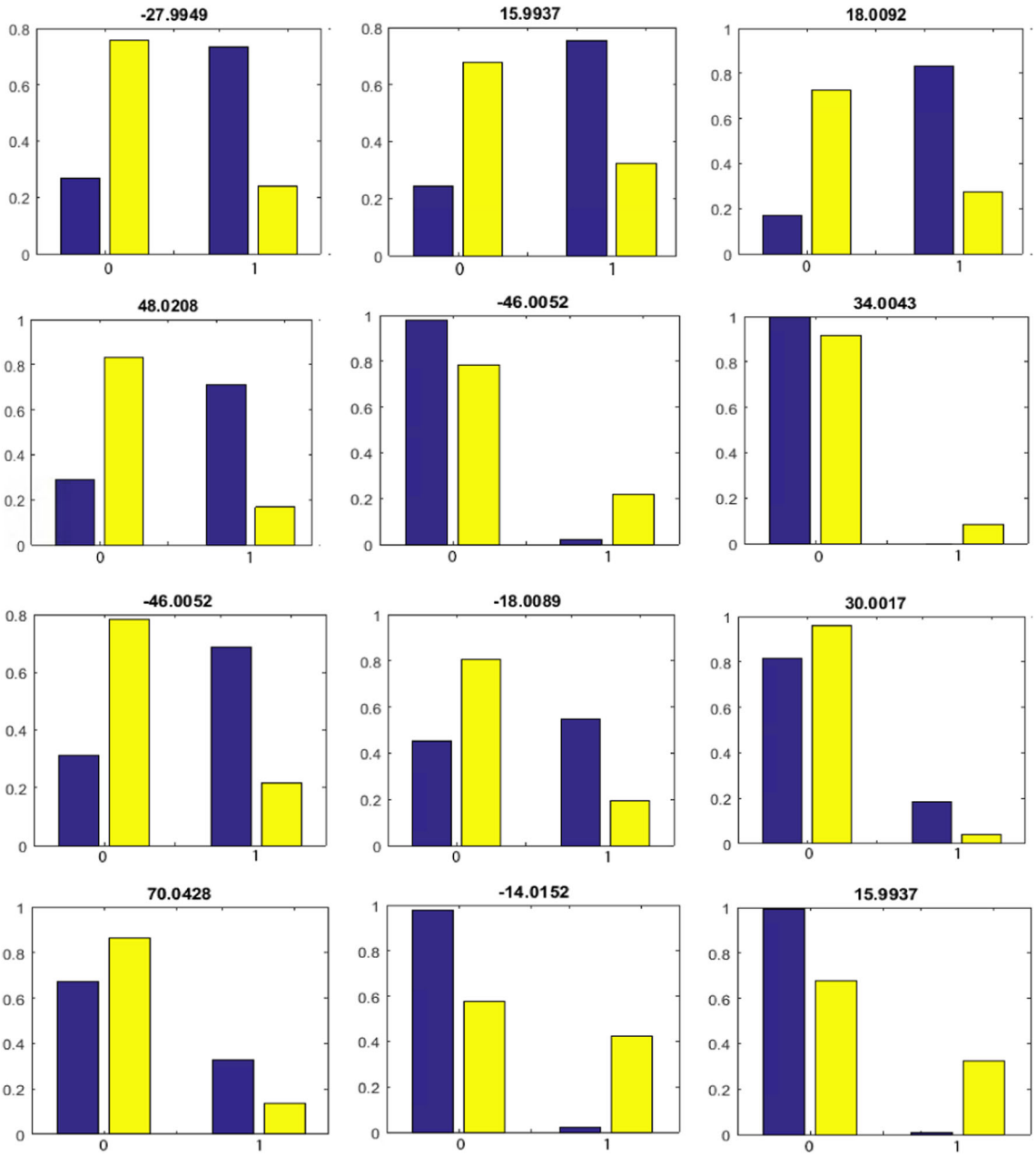


Figure 4. Distributions of example contextual features that are useful for differentiating B (top two rows) and C (bottom two rows) ions (blue bars) from Y, Z, or O ions (yellow bars). Horizontal axes indicate if a feature exists in a spectrum: 0 – not present; 1 – present. Vertical axes indicate the percentage of a certain type (or types) of ions displaying or missing a given feature

significant, as it demonstrates that GlycoDeNovo can be effectively applied to analysis of non-reducing glycans.

Close inspection showed that IonClassifier could detect meaningful contextual features that were useful for differentiating ion types and identifying fragmentation patterns. Some of these features can be easily assigned, e.g., $B_n - 27.9949$ ($^{1,5}A_n$), $B_n + 18.0089$ (C_n), $B_n + 15.9937$ ($C_n - 2H$), $C_n - 46.0052$ ($^{1,5}A_n$), and $C_n + 70.0428$ ($^{2,4}A_{n+1}$), whereas others may have resulted from fragmentation processes that are not yet understood, e.g., $B_n + 48.0208$ ($B_n + CH_4O_2$). IonClassifier also captured some contextual features that were significantly more likely to appear in the context of Y, Z, or O ions than in the context of B or C ions. For example, –46.0052 and +34.0043 were barely observed in the context of B ions, and –14.0152 and +15.9937 appeared scarcely in the context of C ions. The distributions of these contextual features are shown in Figure 4. Fragmentation patterns such as these can be difficult for human eyes to capture because of the volume of data and noises. It is important to note that the IonClassifier is not perfect and needs further improvements. In some cases, it was not able to distinguish the true topology from a few other candidates because they shared the same set of supporting (glycosidic) peaks, and had identical cumulative IonClassifier score. For example, the canonical Man9 topology shared the top rank with four other

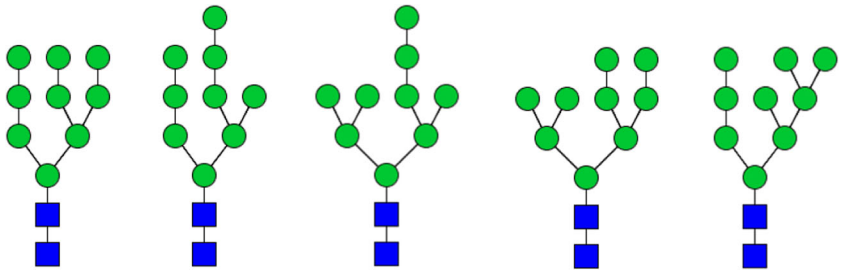


Figure 5. Top-ranked candidate topologies for the Man9 *N*-glycan, with its canonical structure shown on the left. IonClassifier could not differentiate these candidates because they shared the same set of supporting peaks, including the saccharide compositions of Hex, Hex_2, Hex_3, Hex_5, and $Hex_9HexNAc$ for non-reducing-end glycosidic fragments

structures (Figure 5) even when ranked by IonClassifier. Differentiation of structures sharing the same set of supporting peaks would require consideration of other types of ions, such as cross-ring fragments, but this cannot be achieved until the linkage configuration is established, and will be the subject of future studies.

We adopted the leave-one-out approach for IonClassifier training, that is, for any given glycan spectrum that was being tested by IonClassifier, it was excluded from being used to train IonClassifier. In addition, only the spectral data of reducing-end modified glycans were used to train IonClassifier. The rationale is that without any reducing-end modification, many B (or C) ions would have the same mass as Z (or Y) ions, even for asymmetric structures, such as B_1 and Z_1 ions, as well as B_3 and Z_3 ions, of LNT and LNnT. Because the contexts of isomeric B and Z, or C and Y ions are essentially the same, inclusion of these spectral data for training would only serve to misguide the training of IonClassifier. Nonetheless, the IonClassifier learned from the spectral data of modified glycans appeared to work very well for unmodified glycans (see, for example, the last seven rows of Table 2). This is perhaps not surprising as the reducing-end isotope-labeling is not expected to significantly alter the glycan fragmentation pattern. Naturally, presence of similar structural motifs in the training dataset can boost the performance of IonClassifier. Thus, the accuracy and robustness of IonClassifier can be further improved as more experimental data become available for training.

Conclusions

GlycoDeNovo is an efficient and robust algorithm for accurate reconstruction of glycan topologies from their tandem mass spectra. It uses an efficient strategy with a polynomial time complexity to reconstruct candidate topologies. In addition, GlycoDeNovo is equipped with a machine learning-based IonClassifier for candidate topology scoring. The experimental results clearly demonstrated the power of GlycoDeNovo and IonClassifier for de novo glycan sequencing. The present study showed that it is possible to automatically learn fragmentation patterns from real-world tandem MS data. We expect that the availability of more experimental data will allow us to develop better machine learning techniques for building a more powerful and accurate IonClassifier. In the future, we will improve IonClassifier to further take advantage of local structural information in decision making. The IonClassifier can be trained to be specific to different derivatization schemes and fragmentation modes, thus allowing a broader application of GlycoDeNovo. Currently, GlycoDeNovo is implemented in MATLAB, and will be converted into Java for faster computation. Presently, GlycoDeNovo considers eight common monosaccharide classes (Xyl, Fuc, Hex, HexA, HexNAc, Kdo, NeuAc, and NeuGc). Other types of monosaccharide residues (e.g., HexN, Kdn) can be easily incorporated as needed to expand the capability of GlycoDeNovo to analyze a wide variety of glycans (e.g., glycans from lower organisms, and modified glycans).

Acknowledgements

This work is supported by the NIH grants P41 GM104603, S10 RR025082, and R21 GM122635, and by a Brandeis University research fund. The authors thank Dr. Lei Li and Dr. Peng Wang at Chemily Glycoscience for their generous supply of the synthetic *N*-linked glycan standards.

References

1. Helenius, A., Aebi, M.: Intracellular functions of N-linked glycans. Science **291**, 2364–2369 (2001)
2. Ohtsubo, K., Marth, J.D.: Glycosylation in cellular mechanisms of health and disease. Cell **126**, 855–867 (2006)
3. Jefferis, R.: Glycosylation as a strategy to improve antibody-based therapeutics. Nat. Rev. Drug Discov. **8**, 226–234 (2009)
4. Solá, R.J., Griebenow, K.: Glycosylation of therapeutic proteins. BioDrugs **24**, 9–21 (2010)
5. Dennis, J.W., Granovsky, M., Warren, C.E.: Glycoprotein glycosylation and cancer progression. Biochimica et Biophysica Acta (BBA)-Gen. Subj. **1473**, 21–34 (1999)
6. Dube, D.H., Bertozzi, C.R.: Glycans in cancer and inflammation—potential for therapeutics and diagnostics. Nat. Rev. Drug Discov. **4**, 477–488 (2005)
7. Dell, A., Morris, H.R.: Glycoprotein structure determination by mass spectrometry. Science **291**, 2351–6 (2001)
8. Zaia, J.: Mass spectrometry of oligosaccharides. Mass Spectrom. Rev. **23**, 161–227 (2004)
9. Domon, B., Costello, C.E.: A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconjugate J. **5**, 397–409 (1988)

10. Tseng, K., Hedrick, J.L., Lebrilla, C.B.: Catalog-library approach for the rapid and sensitive structural elucidation of oligosaccharides. Anal. Chem. **71**, 3747–54 (1999)
11. Joshi, H.J., Harrison, M.J., Schulz, B.L., Cooper, C.A., Packer, N.H., Karlsson, N.G.: Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics **4**, 1650–64 (2004)
12. Lohmann, K.K., von der Lieth, C.W.: GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates. Nucleic Acids Res. **32**, W261–6 (2004)
13. Cooper, C.A., Gasteiger, E., Packer, N.H.: GlycoMod–a software tool for determining glycosylation compositions from mass spectrometric data. Proteomics **1**, 340–9 (2001)
14. Packer, N. H.; von der Lieth, C. W.; Aoki-Kinoshita, K. F.; Lebrilla, C. B.; Paulson, J. C.; Raman, R.; Rudd, P.; Sasisekharan, R.; Taniguchi, N.; York, W. S. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11-13, 2006). *Proteomics* 8, 8-20 (2008)
15. Gaucher, S.P., Morrow, J., Leary, J.A.: STAT: a saccharide topology analysis tool used in combination with tandem mass spectrometry. Anal. Chem. **72**, 2331–6 (2000)
16. Mizuno, Y., Sasagawa, T., Dohmae, N., Takio, K.: An automated interpretation of MALDI/TOF postsource decay spectra of oligosaccharides. 1. automated peak assignment. Anal. Chem. **71**, 4764–71 (1999)
17. Ethier, M., Saba, J.A., Ens, W., Standing, K.G., Perreault, H.: Automated structural assignment of derivatized complex N-linked oligosaccharides from tandem mass spectra. Rapid Commun. Mass Spectrom. **16**, 1743–54 (2002)
18. Ethier, M., Saba, J.A., Spearman, M., Krokhin, O., Butler, M., Ens, W., Standing, K.G., Perreault, H.: Application of the StrOligo algorithm for the automated structure assignment of complex N-linked glycans from glycoproteins using tandem mass spectrometry. Rapid Commun. Mass Spectrom. **17**, 2713–20 (2003)
19. Tang, H., Mechref, Y., Novotny, M.V.: Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics **21**(Suppl 1), i431–9 (2005)
20. Bocker, S., Kehr, B., Rasche, F.: Determination of glycan structure from tandem mass spectra. IEEE/ACM Trans. Comput. Biol. Bioinform. **8**, 976–86 (2011)
21. Shan, B., Ma, B., Zhang, K., Lajoie, G.: Complexities and algorithms for glycan sequencing using tandem mass spectrometry. J. Bioinform. Comput. Biol. **6**, 77–91 (2008)
22. Sun, W., Lajoie, G.A., Ma, B., Zhang, K.: Bioinformatics research and applications, pp. 320–330. Springer International Publishing, Switzerland (2015)
23. Dong, L., Shi, B., Tian, G., Li, Y., Wang, B., Zhou, M.: An accurate de novo algorithm for glycan topology determination from mass spectra. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**, 568–78 (2015)
24. Kumozaki, S., Sato, K., Sakakibara, Y.: A machine learning based approach to de novo sequencing of glycans from tandem mass spectrometry spectrum. IEEE/ACM Trans. Comput. Biol. Bioinform. **12**, 1267–74 (2015)
25. Korte, B., Vygen, J.: Combinatorial optimization: theory and algorithms. Springer-Verlag, Berlin Heidelberg (2006)
26. Kreyszig, E. Advanced Engineering Mathematics (4th ed.) Wiley, (1979)
27. Sun, W.; Lajoie, G. A.; Ma, B.; Zhang, K. in *Bioinformatics Research and Applications*. Springer International Publishing, 2015, vol. 9096.
28. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**, 119–139 (1997)
29. Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. Classification and Regression Trees. Chapman and Hall/CRC, (1984)
30. Ciucanu, I., Kerek, F.: A simple and rapid method for the permethylation of carbohydrates. Carbohydr. Res. **131**, 209–217 (1984)
31. Ciucanu, I., Costello, C.E.: Elimination of oxidative degradation during the per-O-methylation of carbohydrates. J. Am. Chem. Soc. **125**, 16213–16219 (2003)
32. Koster, C.; Holle, A. presented in part at ASMS annual conference, Dallas, TX., 1999
33. Budnik, B.A., Haselmann, K.F., Elkin, Y.N., Gorbach, V.I., Zubarev, R.A.: Applications of electron-ion dissociation reactions for analysis of polycationic chitooligosaccharides in Fourier transform mass spectrometry. Anal. Chem. **75**, 5994–6001 (2003)
34. Adamson, J.T., Hakansson, K.: Electron capture dissociation of oligosaccharides ionized with alkali, alkaline earth, and transition metals. Anal. Chem. **79**, 2901–2910 (2007)
35. Wolff, J.J., Amster, I.J., Chi, L., Linhardt, R.J.: Electron detachment dissociation of glycosaminoglycan tetrasaccharides. J. Am. Soc. Mass Spectrom. **18**, 234–244 (2007)
36. Devakumar, A., Mechref, Y., Kang, P., Novotny, M.V., Reilly, J.P.: Laser-induced photofragmentation of neutral and acidic glycans inside an ion-trap mass spectrometer. Rapid Commun. Mass Spectrom. **21**, 1452–1460 (2007)
37. Zhao, C., Xie, B., Chan, S.Y., Costello, C.E., O'Connor, P.B.: Collisionally activated dissociation and electron capture dissociation provide complementary structural information for branched permethylated oligosaccharides. J. Am. Soc. Mass Spectrom. **19**, 138–150 (2008)
38. Wolff, J.J., Leach, F.E., Laremore, T.N., Kaplan, D.A., Easterling, M.L., Linhardt, R.J., Amster, I.J.: Negative electron transfer dissociation of glycosaminoglycans. Anal. Chem. **82**, 3460–3466 (2010)
39. Han, L., Costello, C.E.: Electron transfer dissociation of milk oligosaccharides. J. Am. Soc. Mass Spectrom. **22**, 997–1013 (2011)
40. Yu, X., Huang, Y., Lin, C., Costello, C.E.: Energy-dependent electron activated dissociation of metal-adducted permethylated oligosaccharides. Anal. Chem. **84**, 7487–7494 (2012)
41. Yu, X., Jiang, Y., Chen, Y., Huang, Y., Costello, C.E., Lin, C.: Detailed glycan structural characterization by electronic excitation dissociation. Anal. Chem. **85**, 10017–10021 (2013)
42. Gao, J., Thomas, D.A., Sohn, C.H., Beauchamp, J.: Biomimetic reagents for the selective free radical and acid–base chemistry of glycans: application to glycan structure determination by mass spectrometry. J. Am. Chem. Soc. **135**, 10684–10692 (2013)
43. Desai, N., Thomas, D.A., Lee, J., Gao, J., Beauchamp, J.: Eradicating mass spectrometric glycan rearrangement by utilizing free radicals. Chem. Sci. **7**, 5390–5397 (2016)
44. Pu, Y., Ridgeway, M.E., Glaskin, R.S., Park, M.A., Costello, C.E., Lin, C.: Separation and identification of isomeric glycans by selected accumulation-trapped ion mobility spectrometry-electron activated dissociation tandem mass spectrometry. Anal. Chem. **88**, 3440–3443 (2016)
45. Huang, Y., Pu, Y., Yu, X., Costello, C.E., Lin, C.: Mechanistic study on electronic excitation dissociation of the cellobiose-Na+ complex. J. Am. Soc. Mass Spectrom. **27**, 319–328 (2016)
46. Brüll, L., Kovácik, V., Thomas-Oates, J., Heerma, W., Haverkamp, J.: Sodium-cationized oligosaccharides do not appear to undergo 'internal residue loss' rearrangement processes on tandem mass spectrometry. Rapid Commun. Mass Spectrom. **12**, 1520–1532 (1998)
47. Harvey, D.J., Mattu, T.S., Wormald, M.R., Royle, L., Dwek, R.A., Rudd, P.M.: "Internal residue loss": rearrangements occurring during the fragmentation of carbohydrates derivatized at the reducing terminus. Anal. Chem. **74**, 734–740 (2002)