Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-024-05500-9

PAPER IN FOREFRONT

Navigating the maze of mass spectra: a machine-learning guide to identifying diagnostic ions in *O*-glycan analysis

James Urban[1,2] · Roman Joeres[1,2,3,4] · Luc Thomès[5] · Kristina A. Thomsson[6] · Daniel Bojar[1,2]

Received: 28 June 2024 / Revised: 6 August 2024 / Accepted: 9 August 2024

Abstract

Structural details of oligosaccharides, or glycans, often carry biological relevance, which is why they are typically elucidated using tandem mass spectrometry. Common approaches to distinguish isomers rely on diagnostic glycan fragments for annotating topologies or linkages. Diagnostic fragments are often only known informally among practitioners or stem from individual studies, with unclear validity or generalizability, causing annotation heterogeneity and hampering new analysts. Drawing on a curated set of 237,000 *O*-glycomics spectra, we here present a rule-based machine learning workflow to uncover quantifiably valid and generalizable diagnostic fragments. This results in fragmentation rules to robustly distinguish common *O*-glycan isomers for reduced glycans in negative ion mode. We envision this resource to improve glycan annotation accuracy and concomitantly make annotations more transparent and homogeneous across analysts.

Keywords Machine learning · Mass spectrometry · Carbohydrates · Bioinformatics · Computational biology

Introduction

Glycans decorate proteins and lipids and are present in all biological taxa [1]. Molecules interacting with glycans, such as lectins, are frequently sensitive to the three-dimensional conformation of a glycan [2], largely dictated by its constituent monosaccharides and the linkages joining them together. Small changes in linkage or hydroxyl group orientation can lead to different 3D structures with significant biological effects as consequences, for instance by differentially stabilizing a protein depending on the sialic acid linkage [3] or yielding qualitative differences in lectin binding depending on the exact glycan sequence [4]. This makes detailed characterization of glycan sequences in glycomics data crucial for uncovering the roles of specific isomers in particular biological systems. While many methods can be used for this purpose, we will focus our attention here on the most common approach: tandem mass spectrometry, usually preceded by liquid chromatography to separate isomeric structures.

Diagnostic fragments—only, or at least preferentially, occurring in one isomer—comprise a substantial part of current and preferred annotation strategies, due to their ease of use compared to alternative strategies such as exoglycosidase digestion. Examples here include diagnostic fragments to distinguish sialic acid linkage in *N*-glycans [5] or for the distinction of Lewis A and X structures [6]. Despite this, the usage of diagnostic fragments is neither standardized nor formalized, creating a lack of transparency and an entry barrier for analysts. No central databases or resources exist to catalog or compare such diagnostic fragments. Further, this lack of formalization also means that no quantitative confidence value can be attached to an individual human

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

✉ Daniel Bojar
daniel.bojar@gu.se

[1] Department of Chemistry and Molecular Biology, University of Gothenburg, Gothenburg, Sweden

[2] Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden

[3] Helmholtz Institute for Pharmaceutical Research Saarland, Helmholtz Centre for Infection Research, Saarbrücken, Germany

[4] Center for Bioinformatics, Saarland University, Saarbrücken, Germany

[5] ULR 7364 - RADEME - Maladies RAres du DÉveloppement embryonnaire et du Métabolisme, CHU Lille, University Lille, 59000 Lille, France

[6] Proteomics Core Facility at Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Published online: 24 August 2024

annotation, withholding necessary context and hampering transparency.

Several comprehensive studies to identify diagnostic fragmentation have been carried out before [5–8]. Typically, isomer-specific rules are devised or evaluated based on spectra obtained in a single experiment, often one carried out for the express purpose of finding these fragments and collected by the same person(s) that then analyzes it to that effect [6–8]. In rare cases, these rules are then validated on different experimental set-ups [5]. Yet, often, little information exists about whether, or to which extent, commonly used diagnostic ions are generalizable to different set-ups. Further, the quantitative efficacy of most rules is typically unknown, as well as the efficacy of combining multiple rules derived from disparate experiments, making them essentially soft rules, in which the (prominent) presence of an indicated fragment is associated with an undetermined annotation confidence.

Given the prevalent use of single/double fragment presences to determine structural details, evaluating and quantifying the performance of such criteria could not only improve annotation accuracy but also attach a confidence level to each annotation, allowing for a proper evaluation of attached biological findings. Here, we will focus on *O*-glycans as a test case. *O*-Glycans are fantastically diverse in the context of mucin glycosylation [9] and very much dependent on diagnostic fragments in their annotation, due to a less rigid biosynthesis than *N*-glycans. Recent comparisons across different analysts in the area of *O*-glycoproteomics have highlighted substantial heterogeneity [10] and it is to be expected that a similar situation arises in *O*-glycomics, especially for new analysts, due to the lack of resources and challenging nature of the problem, as less firm biosynthetic assumptions can be made compared to *N*-glycans. Although automated *O*-glycan annotation approaches have been proposed to aid the determination of isomeric structures [11], the exact decisions made by such approaches are not clearly interpretable, potentially affecting transparency.

For the related area of lectin-glycan binding specificities, an approach combining rule-based machine learning with expert curation has resulted in widely used and robust guidelines for a hitherto scattered field [4]. Thus, here we present a new workflow using interpretable machine learning on a large, curated set of > 237,000 *O*-glycomics spectra to derive an actionable set of rules used to identify common *O*-glycan topologies and structural isomers from tandem mass spectrometry data of reduced glycans in negative ion mode. We then couple the identification of diagnostic peaks with our automated fragment annotation method CandyCrumbs [11], to obtain human-understandable fragmentation events that can be used for annotation. Importantly, these rules are assessed across a wide array of experimental set-ups and analysts, resulting in (i) quantifiable rule performance, (ii) rules that are designed to work in combination with each other, and (iii) annotation confidence values of isomers identified with these rules.

Throughout this work, we also compare where our rules confirm or deviate from existing diagnostic fragments from the academic literature. We show that most *O*-glycan isomers can be confidently separated with a small number of diagnostic features, including ratios of fragment peaks, and even identify fragmentation patterns that are generalizably indicative of the same structural feature across many different glycans. We envision that this work will improve *O*-glycomics annotation accuracy, transparency, homogeneity, and accessibility, leading to new biological discoveries of the role of fine-structural details in glycans.

Materials and methods

Dataset construction

The herein used dataset of glycan tandem mass spectra was extended from a previously curated dataset [11]. Briefly, MS raw files were retrieved from, predominantly, GlycoPOST [12] and converted into mzML format, and MS^2 spectra were extracted into a tabular format. We then filtered our dataset to include only MS^2 spectra of *O*-glycans (containing a reducing end GalNAc or Fuc, as well as *O*-glycan peeling products), measured in negative ion mode, and only including structures which had undergone reductive β-elimination. All annotations by experts in this dataset were assumed to be true. The final dataset consisted of 237,931 spectra and 1647 unique glycans across 121 unique datasets (comprising 1442 glycomics raw files).

Data processing

Spectra were normalized by expressing their intensity as a percentage of the highest peak in the spectrum, in accordance with common practice, to facilitate direct usage of intensity threshold in obtained rules. Spectra were then binned by summing their intensities in *m/z* windows spanning 0.5 Da. Keeping track of the *m/z* difference between bin edge and peak allowed us to reconstruct the exact *m/z* later in the process [11]. Finally, we also formed relevant ratios between all bins of at least a mean value of 0.01 (i.e., 1%), as potent interaction features. Both normalized bins and ratios were available as features to the model trained to distinguish isomers.

Decision trees based on Shannon entropy

In this work, we build one decision tree–based model per mass group (± 0.5 Da around the theoretical mass of a composition) that uses the input spectra to predict the isomers from the group. Following the divide-and-conquer idea, we do not train one decision tree for the whole problem setting; instead, we first classify the topology, if applicable, and then build separate decision trees for each topology. In early experiments, we found the performance of this approach to be superior over fitting single trees per mass group. Additionally, growing smaller trees of depths two to three was often sufficient to achieve excellent prediction performance between isomers, ensuring the practical applicability of derived rules.

Decision trees follow the idea of splitting the set of samples into two parts at each node, maximizing the purity of the partition. This means each node in a tree formulates a classification problem for the subset of samples resulting from the last splitting. The problem is solved by selecting the feature and the splitting value that best solves it. Different ways exist to measure how well a classification problem is solved. We use the Information Gain; a popular alternative is the Gini-Impurity. The Information Gain of a decision is computed as the difference in the Shannon entropy of the node(s) above and below a split.

Figure 1A depicts how to compute the Shannon entropy H as the sum over the classes $x_i \in X$, with $p(x)$ being the proportion of class x in the respective node. In this way, we can measure how pure a node is, as the presence of a few dominant classes (high $p(x)$) will lead to a low H. More evenly distributed class proportions result in high values for H. H can then be used to calculate the information gain IG of a split A, where A represents the splitting value of a feature, as described above. After splitting the samples based on A, the best feature and splitting value are selected by maximizing the information gain where H(X|A) is the weighted sum over the child nodes. Figure 1B visualizes that this scheme can be applied recursively until a stopping criterion is reached [13].

Each tree was trained using scikit-learn v1.4.2, followed by processing using glycowork v1.3 [14]. The available data per classification task was split into 70% training, 20% validation, and 10% test data. We used DataSAIL [15] for splitting to combine a similarity measure based on GlycoPOST ID and filename with stratification, to ensure each class was present in each of the splits. The trees were trained with default parameters of scikit-learn and only optimized towards their depth with the validation set. All code is available on GitHub (https://github.com/BojarLab/FragmentFactory).

Calculating confidence and coverage

Confidence is defined here as the likelihood of a correct annotation when following the rule(s) and was calculated by strictly applying a rule to all relevant spectra for a group of isomers and dividing the number of correct annotations by the number of spectra. Coverage is defined as how many spectra of an isomer A follow the proposed rule(s). Coverage was calculated by strictly applying a rule to all relevant spectra and dividing the number of correct isomer A annotations by the total number of isomer A spectra.

Deriving rules from trees

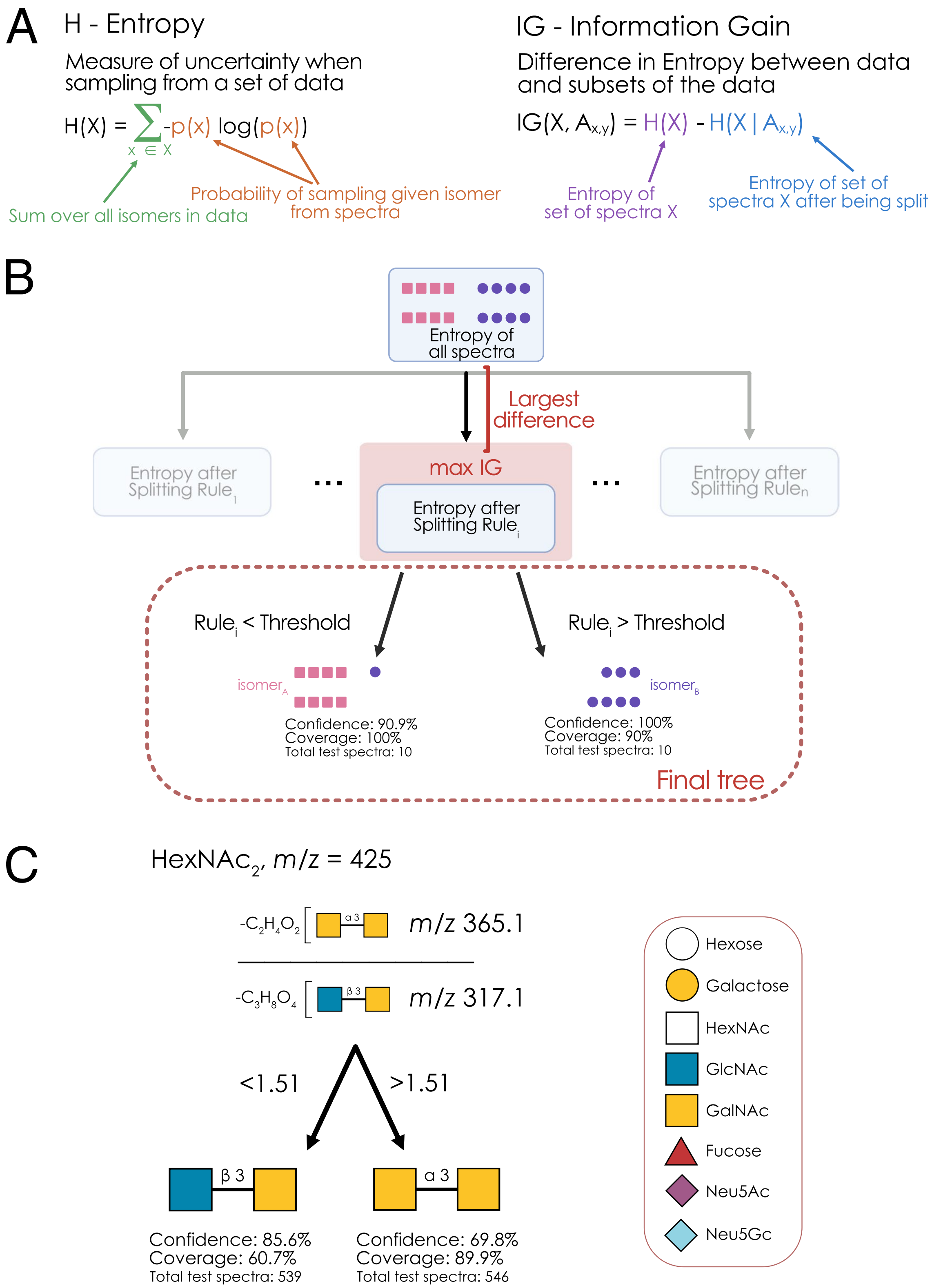
For each tree (both isomer and topology trees), we chose the best decision path per isomer as the source for derived annotation rules. Here, "best" was determined by a score comprising the product of confidence and coverage in a leaf node for that isomer, evaluated on the independent test data (not used in any way for building the tree). Then, bins used for splitting within that decision path were mapped back to their exact *m/z* values, followed by their annotation as candidate fragments via CandyCrumbs [11], which were then visualized via GlycoDraw [16]. This resulted in a set of fragments, with corresponding decision thresholds, that could be used as annotation rules.

Sample preparation of additional MS^n analyses

The sample containing HexNAc?1-?Galβ1-3(Neu5Acα2-6)GalNAc used to produce MS^3 of the *m/z* 800 fragment and MS^4 of the *m/z* 597 fragment was prepared from porcine gastric mucin according to the method reported in Bechtella et al. (2024) [17]. The sample containing Galβ1-3(Neu5Acα2-6)GalNAc used to produce MS^3 of the *m/z* 597 fragment was prepared in gilthead seabream mucin as reported in Thomsson et al. (2024) [18].

Glycans were resuspended in water (15 µL) and injected (2 µL) onto a liquid chromatography-electrospray ionization tandem mass spectrometry (LC-ESI/MS). The HPLC was a Vanquish Neo (Thermo Scientific). The oligosaccharides were separated on a column (10 cm × 250 µm) packed in-house with 5-µm porous graphite particles (PGC, Hypercarb, Thermo-Hypersil, Runcorn, UK) and a flow rate of 6 µL/min. The oligosaccharides were eluted with the following gradient: 5–20 min 1–25% B, wash 21–31 min 99% B, then equilibration between 32 and 52 min with 1% B. Buffer A was 10 mM ammonium bicarbonate (ABC) and buffer B was 10 mM ABC in 80% acetonitrile.

The samples were analyzed in negative ion mode on an Orbitrap mass spectrometer (Fusion, Thermo Electron, San José, CA). Compressed air was used as nebulizer gas. The heated capillary was kept at 325 °C. Full scan (MS^1) was set to *m/z* 670–680 (sea bream (SB) sample) or *m/z* 877–880 (PGM sample), and the resolution was 60,000. Two microscans were performed, maximum injection time was 118 ms, and AGC target was set to 800,000 (sea bream sample) or 400,000 (PGM sample). Selected CID MS^n scans using the precursor ion list function were performed as follows for the SB sample (MS^2 → MS^3, *m/z* 675.245 → 597.2) and the PGM sample



($MS^2 \rightarrow MS^3 \rightarrow MS^4$, m/z 878.33 → 800.2 → 597.2). AGC target was set to 30,000, with normalized collision energy of 35%, isolation window of 2 units, activation $q = 0.25$, and activation time 30 ms.

Data availability

All relevant data, including their data provenance with accession IDs, can be found on Zenodo under the https://

◂**Fig. 1** Rule-based machine learning to uncover diagnostic fragments. **A** Definition of Entropy as a measure of sample uncertainty, as well as the Information Gain as the reduction in sample uncertainty after a given decision. **B** Schema of decision tree construction indicating the greedy optimization of information gain at each node until the maximum depth is reached. **C** Machine learning–derived rule for distinguishing core 3 and core 5 *O*-glycans ($HexNAc_2$, *m*/*z* 425). The best decision tree for isomers of *m*/*z* 425 is shown, with the decision threshold representing values of the ratio between the two fragment ions. Confidence indicates the likelihood of a correct annotation when following the rule(s), whereas coverage designates how many spectra of that isomer follow the rule(s). The number of test spectra (not used in training the model and stemming from different experiments) for each isomer is provided in all decision trees as well. All fragments in this work are written in Domon-Costello nomenclature [20] and are visualized via GlycoDraw [16], adhering to the Symbol Nomenclature For Glycans (SNFG)

doi.org/10.5281/zenodo.12177170 [19]. Acquired mass spectrometry data are available at GlycoPOST, under the ID GPST000457.

Code availability

All relevant code for this work can be found at https://github.com/BojarLab/FragmentFactory.

Results

Rule-based machine learning yields widely usable diagnostic fragments

A systematic approach to identify generalizable diagnostic fragments requires, at least, two things: (i) a large set of MS^2 spectra from different experimental set-ups and different analysts, and (ii) an algorithm producing effective, but human-interpretable, rules to determine the correct isomer based on the MS^2 spectrum. For our previous work [11], we have curated a large set of annotated MS^2 spectra, which we have updated for this work with a special focus on *O*-glycomics data from reduced glycans in negative ion mode. Within these parameters, this dataset can be viewed as representative for a great variety of analysts and their respective set-ups. We then engaged in a rigorous data splitting procedure using DataSAIL [15] (see "Materials and methods"), to ensure that we only evaluated identified rules on experiments that differed from the ones used to generate the rules. This was important to (i) ensure the generalizability of obtained annotation rules and (ii) gain accurate performance metrics (confidence and coverage) for each set of rules.

With this, we could train machine learning models to predict the annotated isomer for a spectrum, given its fragment ions (Fig. S1). To achieve a set of annotation rules that was both performant and small, we trained a decision tree–based model for each group of isomers that minimized Shannon entropy (Fig. 1A), where each best split was considered one annotation rule (Fig. 1B). Importantly, for each isomer, this provided us with confidence and coverage values, where confidence indicated the proportion of true positives when using those rules and coverage indicated how many spectra of that isomer fell under those rules.

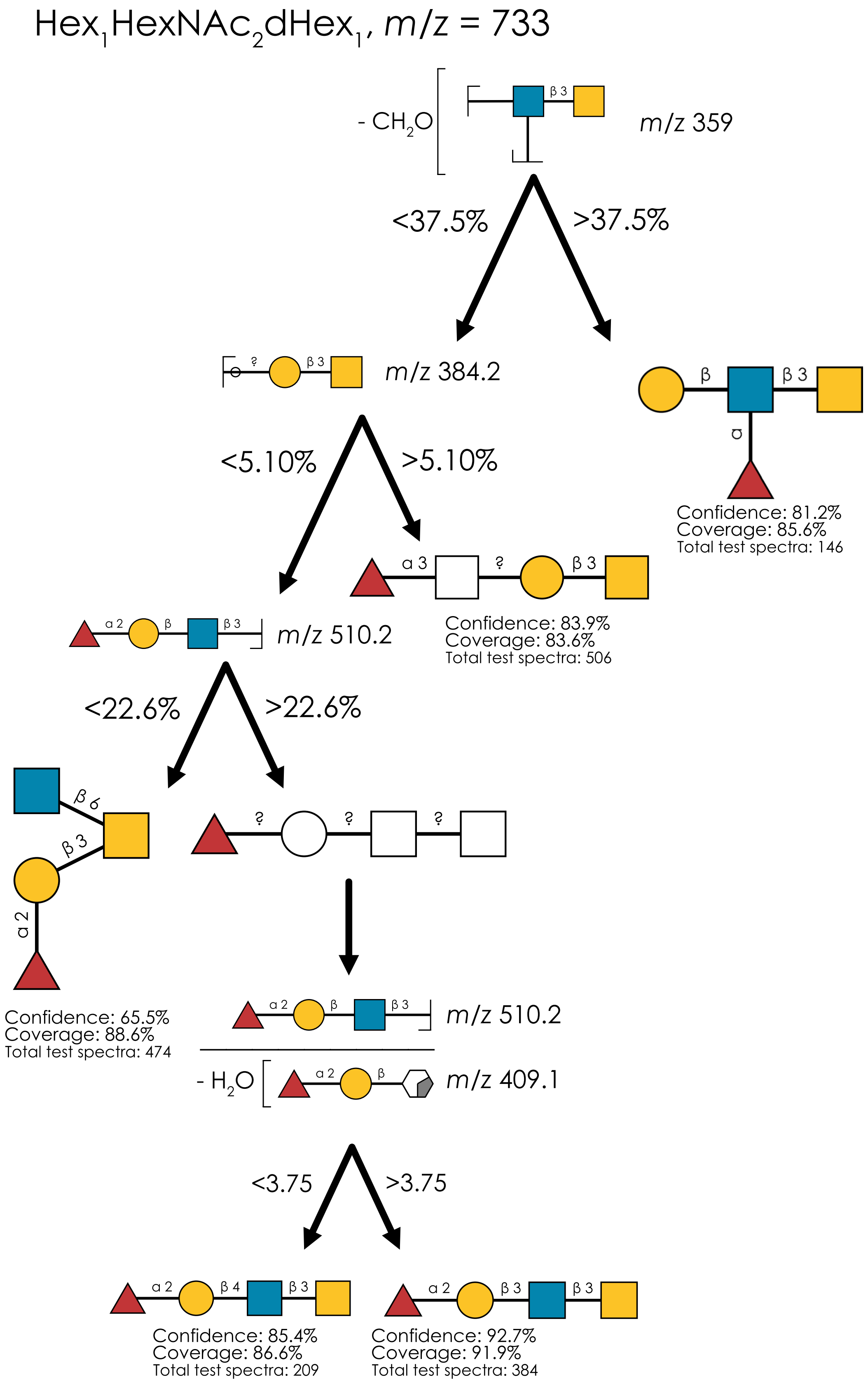
In general, this allowed us to construct sets of rules for many common *O*-glycan isomers where, in most cases, one or two rules were sufficient to achieve excellent confidence and coverage. One example can be seen in the model distinguishing the core 3 from the core 5 structure, where a single rule (the ratio between *m*/*z* 365.1 and *m*/*z* 317.1) was enough to effectively disambiguate between the two isomers (Fig. 1C). A value of above 1.5 here indicated the core 5 structure, allowing for easy application of this rule in practice. While there are no commonly used/accepted diagnostic fragments to distinguish these two isomers, past research comparing core 3 and core 5 structures in seabream mucin [18] supports our use of *m*/*z* 365, yet we here show that this can be improved by combining it with the *m*/*z* 317 fragment into a ratio, highlighting the potential value of this approach.

Of course, some isomeric differences, such as for the structure group $Hex_1HexNAc_1dHex_1$ (*m*/*z* 530), are very robust and can be almost considered to be "solved." In this case, the prominent presence of a $HexNAc_1dHex_1$ Z-ion (*m*/*z* 350.1) typically indicates an *O*-Fuc isomer (most often Galβ1-4GlcNAcβ1-3Fuc), in contrast to the standard *O*-GalNAc type isomer (Fucα1-2Galβ1-3GalNAc), in which this Z-ion would be topologically impossible. We were thus reassured to see that our new machine learning–based approach recovered these well-known effects and indeed chose *m*/*z* 350.1 as the best feature to distinguish these features (Fig. S2), resulting in 100% confidence and coverage at the best intensity splitting threshold. We then further aimed to distinguish type 1 and type 2 LacNAc isomers of this *O*-Fuc isomer and present *m*/*z* 488.2 as a potential new diagnostic fragment (Fig. S2), which indicates Galβ1-4GlcNAcβ1-3Fuc when present prominently and Galβ1-3GlcNAcβ1-3Fuc by its relative absence (given that the isomer Galβ1-?GlcNAcβ1-3Fuc has been already chosen due to *m*/*z* 350.1).

Distinguishing topology and linkage differencesvia a divide-and-conquer approach

Many *O*-glycan structure groups comprise both topologically different isomers, as well as those differing in a single linkage, presenting a multitude of challenges to annotators. A common example of such mass groups can be found in the, still relatively modest, composition of $Hex_1HexNAc_2dHex_1$ (*m*/*z* 733), which can form Lewis antigens, blood group epitopes, as well as three different core structures.

Here, we would like to showcase our divide-and-conquer approach of combining topology-level with linkage-level


Hex1HexNAc2dHex1, m/z = 733
- CH2O
β 3
m/z 359
<37.5%
>37.5%
?
β 3
m/z 384.2
β
β 3
α
Confidence: 81.2%
Coverage: 85.6%
Total test spectra: 146
<5.10%
>5.10%
α 3
?
β 3
Confidence: 83.9%
Coverage: 83.6%
Total test spectra: 506
α 2
β
β 3
m/z 510.2
<22.6%
>22.6%
β 6
β 3
α 2
?
?
?
Confidence: 65.5%
Coverage: 88.6%
Total test spectra: 474
α 2
β
β 3
m/z 510.2
- H2O
α 2
β
m/z 409.1
<3.75
>3.75
α 2
β 4
β 3
α 2
β 3
β 3
Confidence: 85.4%
Coverage: 86.6%
Total test spectra: 209
Confidence: 92.7%
Coverage: 91.9%
Total test spectra: 384

◂**Fig. 2** Distinguishing topologies and isomers with a divide-and-conquer approach. For the isomer group at *m/z* 733 ($Hex_1HexNAc_2dHex_1$), we used our decision tree–based approach to find rules distinguishing topologies and, finally, isomers. The combined decision tree with all rules is shown. Rules are visualized via the SNFG-depiction of Domon-Costello fragments and their corresponding threshold values for decision-making. The number of independent test spectra, as well as the therein achieved confidence and coverage, is shown for each isomer in its respective leaf node

models to obtain effective annotation rules (Fig. 2). A fragment containing the core 3 structure (*m/z* 359) was sufficient to separate Lewis-type structures from everything else. Then, we could separate core 1 isomers of *m/z* 733 via the presence of a Y ion containing the core 1 epitope itself (*m/z* 384.2). This was then followed by the separation of blood group core 2 and core 3 structures via *m/z* 510.2, the prominent presence of which as a B-ion indicated the linear core 3 structures. Finally, a ratio of this B-ion with an A-type cross-ring fragment on the GlcNAc residue (*m/z* 409.1) was sufficient to separate type 1 and type 2 LacNAc isomers of this structure (i.e., Fucα1-2Galβ1-**3**GlcNAcβ1-3GalNAc vs Fucα1-2Galβ1-**4**GlcNAcβ1-3GalNAc).

We were excited to see that this obtained decision scheme exhibited excellent confidence and coverage for all identified isomers. Specifically, the presented rules covered well over 80% of all spectra that contained the annotated isomers, making them extremely robust and applicable in most experimental settings. Combined with an annotation confidence of, in most cases, 80–90%, we envision these rules to raise annotation quality. We caution that, in this case, we did not identify satisfactory diagnostic features to distinguish Lewis A and Lewis X on the Lewis-type core 3 structure. The disambiguation of Lewis structures in reduced glycans presents a challenging problem in general [8, 21], which is compounded by the relative rarity of Lewis-type core 3 structures in our dataset. As discussed later, we also do want to point out that, for other mass groups such as *m/z* 895 ($Hex_2HexNAc_2dHex_1$), our models are, in fact, capable of identifying robust indicators for Lewis A and X, respectively (Fig. S11).

A guide to annotate common O-glycan isomers

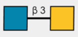
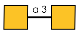
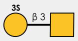
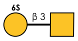
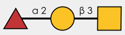
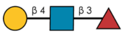
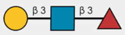
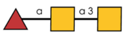
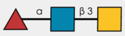
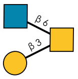
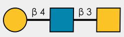
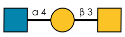
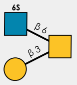
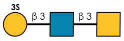
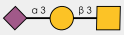
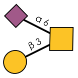
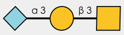
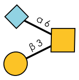
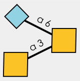
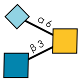
Having demonstrated the capabilities of both our rule-based machine learning approach in general, as well as its extension via the divide-and-conquer approach, we then moved on to extend this potent new approach to common sets of *O*-glycan isomers. We here present a comprehensive set of quantitatively identified and characterized annotation rules for common *O*-glycan isomers (Fig. 3). We note that we only included structures in this analysis that have known and relevant isomers (e.g., no rules were constructed for sialyl-Tn antigen annotation, due to the lack of alternative isomers).

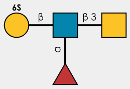
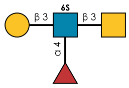
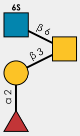
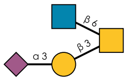
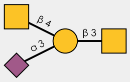
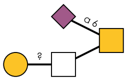
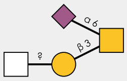
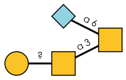
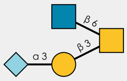
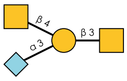
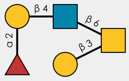
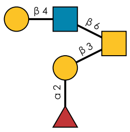
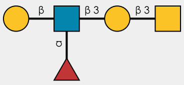
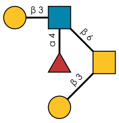
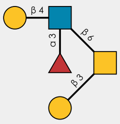
Sulfated structures can be especially difficult to correctly annotate, which is why we are enthusiastic that in some cases, such as $Hex_1HexNAc_1S_1$ (*m/z* 464; Fig. S3), our models could even identify diagnostic ratios to distinguish sulfate positioning on the galactose (Gal3S vs Gal6S) with satisfactory performance (> 70% confidence and coverage). This was then extended in $Hex_1HexNAc_2S_1$ (*m/z* 667; Fig. S6), in which we identified the ratio between *m/z* 444.1 and *m/z* 487.1 as most performant to distinguish core 2 and core 3 isomers of this composition. Other relevant examples that include new insights into diagnostic fragmentation behavior include $Hex_1HexNAc_2dHex_1S_1$ (*m/z* 813), a common sulfated structure group that can form either Lewis structures or an H-type 3 blood group epitope. Next to these topological distinctions, the sulfate moiety can be found on either the GlcNAc or the Gal residue, further complicating annotation. We find that a ratio of the sulfo-Lewis moiety (*m/z* 590.1) and the sulfated core 6 substructure (*m/z* 505.1) was sufficient to separate the scenarios of sulfated Gal and GlcNAc, respectively, which then was further refined via another ratio to separate Lewis and blood group structures (Fig. S8).

Overall, we note that many of the best models to distinguish isomers used ratios of fragment ion intensities as annotation features. We thus conclude, in accordance with much of the academic literature on this topic, that ratios are powerful diagnostic features and are optimistic that more complex combinations of fragment intensities, balanced with ease of use by humans, will allow for even more confident annotations. We also would like to point out that the formation of ratios is (i) more robust to systematic shifts in intensities and (ii) mitigates some of the compositional nature of relative intensities, increasing generalizability across datasets [22].

Derived rules can generalize beyond individual structures

In general, when seeking to distinguish two specific glycan motifs or isomers, the simplest approach would be to utilize “topologically exclusive” fragments, i.e., fragment masses that are only possible in a single glycan topology. Such fragments might be specific to a topology or glycan substructure, producing a high confidence value, but they are not guaranteed to occur in every experimental set-up, e.g., due to preferred alternative fragmentation pathways, yielding low coverage values. To take one example, the mass of a Neu5Ac-HexNAc fragment (*m/z* 513.2) is exclusive to the topologies containing core GalNAc sialylation. This fragment has been previously described [6] as diagnostic of this type of sialylation. Yet, when tested across a more diverse set of experiments, we found it to be a rather low-coverage rule to indicate a Neu5Ac-GalNAc core motif (Fig. S9B). Specifically, the presence of *m/z* 513.2 resulted in an 86% confidence of Neu5Ac-GalNAc annotation, yet this rule

m/z	Glycan	Rules (*m/z*)	Conf. Cov.
425		$\frac{365.1}{317.1} < 1.51$	86% 61%
425		$\frac{365.1}{317.1} > 1.51$	70% 90%
464		$\frac{282.1}{241.0} < 3.44$ $\frac{350.1}{422.1} < 3.56$	77% 71%
464		$\frac{282.1}{241.0} < 3.44$ $\frac{350.1}{422.1} > 3.56$	66% 72%
530		350.1 < 20.1%	100% 100%
530		350.1 > 20.1% 488.2 > 23.1%	89% 91%
530		350.1 > 20.1% 488.2 < 23.1%	91% 89%
571		$\frac{290.1}{511.2} < 0.987$	70% 99%
571		$\frac{290.1}{511.2} > 0.987$	99% 57%
587		$\frac{364.1}{407.1} < 0.07$	86% 93%
587		$\frac{364.1}{407.1} > 0.07$ $\frac{407.1}{527.1} < 1.55$	82% 49%
587		$\frac{364.1}{407.1} > 0.07$ $\frac{407.1}{527.1} > 1.55$	50% 69%
667		$\frac{444.1}{487.1} < 0.57$	88% 87%
667		$\frac{444.1}{487.1} > 0.57$	88% 88%
675		615.2 < 10.9%	86% 61%
675		615.2 > 10.9%	86% 61%
691		597.2 < 14.3%	86% 61%
691		597.2 > 14.3%	86% 61%
732		638.2 > 18.9%	86% 92%
732		638.2 < 18.9%	66% 52%

m/z	Glycan	Rules (*m/z*)	Conf. Cov.
813		$\frac{590.1}{505.1} > 2.11$	98% 96%
813		$\frac{590.1}{505.1} < 2.11$ $\frac{667.2}{282.0} > 41.1$	62% 58%
813		$\frac{590.1}{505.1} < 2.11$ $\frac{667.2}{282.0} < 41.1$	59% 60%
878		818.2 < 43.1% 384.2 < 0.6%	75% 90%
878		818.2 < 43.1% 384.2 > 0.6%	87% 89%
878		818.2 > 43.1% $\frac{675.2}{513.2} < 0.38$	66% 71%
878		818.2 > 43.1% $\frac{675.2}{513.2} > 0.38$	83% 58%
894		800.2 > 58.5%	100% 90%
894		800.2 < 58.5% 306.1 > 77.3%	90% 84%
894		800.2 < 58.5% 306.1 < 77.3%	68% 89%
895		$\frac{384.2}{389.2} < 1.20$ 569.2 < 93.1%	73% 91%
895		$\frac{384.2}{389.2} < 1.20$ 569.2 > 93.1%	92% 86%
895		$\frac{384.2}{389.2} > 1.20$ $\frac{389.2}{569.2} < 1.77$	86% 72%
895		$\frac{384.2}{389.2} > 1.20$ $\frac{389.2}{569.2} > 1.77$ 587.2 > 0.4%	91% 58%
895		$\frac{384.2}{389.2} > 1.20$ $\frac{389.2}{569.2} > 1.77$ 587.2 < 0.4%	69% 94%

◂**Fig. 3** A useful guide to *O*-glycan isomer annotation. For each isomer for which we could identify performant (> 60% confidence/coverage) as well as interpretable annotation rules here, we catalog the respective rules in a simplified manner. For the exact thresholds regarding intensity (individual fragments) or ratios, we refer to the respective supplementary figures (Fig. 1C, Fig. 2, Fig. 4A–D, Figs. S2–11), which list the exact models with all thresholds. Next to annotation rules, we here also depict the confidence (Conf.) and coverage (Cov.), assessed on an independent test set of experimental spectra, that result when annotating an isomer based on these rules

only covered 57% of Neu5Ac-GalNAc containing spectra, meaning that a large fraction of Neu5Ac-GalNAc containing spectra could not be classified with such a rule.

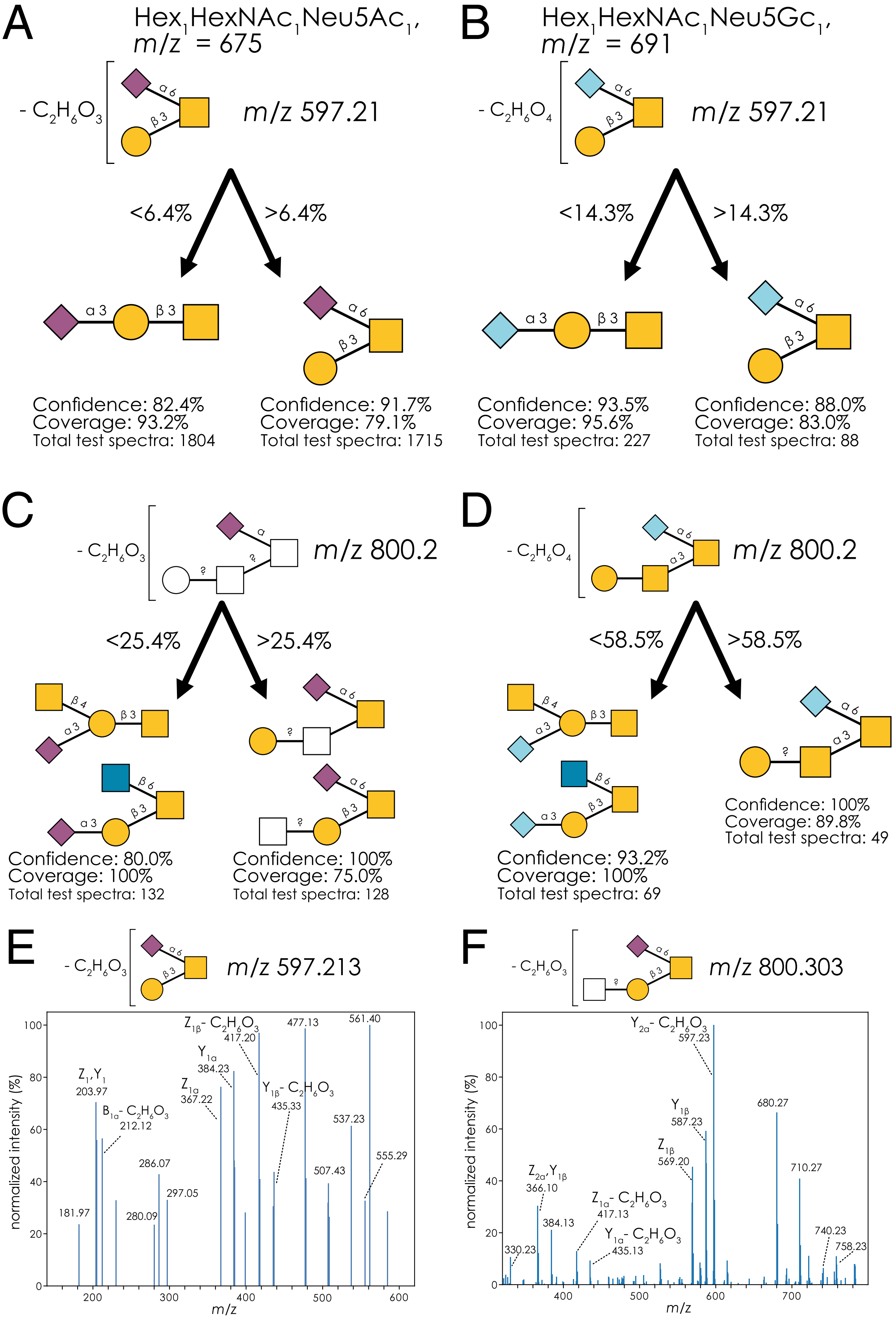
We posit that fragments such as *m/z* 513.2 are especially preferred because they are intuitive, as they are causally related to the topology/isomer that is to be annotated. Yet, as we have shown throughout this work, annotated MS^2 spectra contain many fragments that may not have such a clean explanation, making them less preferred for annotation, but that still offer excellent annotation quality. As a result of this, it is possible that there are many useful fragment ions not currently in use because their structure is either unknown or not intuitively thought to be connected to the isomeric difference. Our data-driven approach is designed to counteract precisely that, and we identified two such fragments that commonly occur in decision trees of sialylated structures. The fragment masses, at M-78 and M-94 for Neu5Ac/Neu5Gc, respectively, are seen in high abundance across a wide array of published MS^2 spectra. Even when mentioned, these fragments have not been fully characterized and are either labeled simply as M-$C_2H_4O_2$-H_2O or, most commonly, not labeled at all.

We found that this unexplained mass loss was effective in distinguishing reducing end GalNAc sialylation from branch Gal sialylation in both Neu5Ac- and Neu5Gc-containing structures (Fig. 4A, B), though we do caution that, in an *O*-glycan context, Sia-HexNAc/Sia-Hex is conflated with α2-6 vs α2-3 linkage of the sialic acid. We can further specify this phenomenon by examining α2-6 vs α2-3 linked Sia-HexNAc motifs in milk oligosaccharides with reducing end glucose [21]. Encouragingly, both linkage types of non-reducing end Sia-HexNAc showed very low or no abundance of the M-78 fragment masses, indicating reducing end HexNAc residues are involved in this loss.

With the example of low-coverage by *m/z* 513.2 (Fig. S9B), we show that M-$C_2H_4O_2$ (*m/z* 818.2) exhibited both higher coverage and higher confidence than the often-used *m/z* 513.2 fragment (Fig. S9C). In another work [23], this fragmentation pattern is also seen in branched sialylated trisaccharides (both Neu5Ac and Neu5Gc), as well as in larger molecules produced by extending these structures. Interestingly, Kdn-containing structures did also produce *m/z* 597 fragments, representing a loss of 36 Da (M-H_2O-H_2O), suggesting the losses at M-78 and M-94 to affect the C5 extension of Neu5Ac and Neu5Gc, as this moiety presents the only molecular difference. The distinguishing fragment masses in Neu5Ac and Neu5Gc differed by 16 Da, further indicating the loss to occur in the *N*-acetyl/*N*-glycolyl group of the sialic acids, due to the additional oxygen atom in Neu5Gc (Fig. 4A, B).

We thus propose that the specific fragmentation of M-78/M-94 here presents the loss of the acetyl/glycolyl group (M-C_2H_2O), paired with two water losses. We note that the order of acetyl loss and then water losses was also proposed in recent work on elucidating sialic acid fragmentation in glycoproteomics data [24]. These water losses could, for instance, occur via a lactonization of the carboxyl group of $C1_{Neu5Ac}$ with the hydroxyl group of $C4_{GalNAc}$. Importantly, $C4_{GalNAc}$ is axial in GalNAc, bringing the hydroxyl group into proximity of $C1_{Neu5Ac}$, which would not be possible in the case of GlcNAc, with an equatorial C4. Using glycan 3D structure information from GlycoShape [25], we could also show that the rotational flexibility of the hydroxyl group on $C4_{GalNAc}$ in this context was higher than that of the one on $C4_{Gal}$ (Fig. S12), potentially explaining the diagnostic behavior of this fragmentation pattern. Another water loss, for instance via 1,7-lactonization, would then result in the observed M-C_2H_2O-H_2O-H_2O in the case of Neu5Ac-containing structures. This pattern also extended to larger structures and generalized to multiple topologies, regardless of the terminal structure on the non-sialic acid branch (Fig. 4C, D). We also note that the utility of this rule encompassed structures with an additional terminal fucose, which also yielded a high relative abundance of M-78 ions after fragmentation [23, 26].

To confirm that the losses occurred in the sialic acid moiety and not somewhere else in the glycan, we acquired an MS^3 spectrum of this diagnostic fragment ion at *m/z* 597 (M-78; Fig. 4E). Abundant peaks at the masses representing $Z_{1β}$-$C_2H_6O_3$ and $Y_{1β}$-$C_2H_6O_3$ indicated that none of the indicated losses occurred on the galactose residue in the $Hex_1HexNAc_1Neu5Ac_1$ isomer. Further, a substantial abundance at *m/z* 212.1 represented the commonly seen $B_{1α}$ fragment at *m/z* 290.1, with a further loss of $C_2H_6O_3$. Finally, the presence of unmodified $Y_{1α}$ and $Z_{1α}$, corresponding to sialic acid loss, supports the finding that the fragmentation events of the -$C_2H_6O_3$ loss occur only within the sialic acid. To ensure the sialic acid fragmentation was not specific to this specific trisaccharide, we acquired a separate MS^3 spectrum of the same phenomenon in an extended structure, $Hex_1HexNAc_2Neu5Ac_1$ at *m/z* 800 (M-78; Fig. 4F). The most abundant peak, at *m/z* 597, represented the exact same fragment ion we originally found in $Hex_1HexNAc_1Neu5Ac_1$, which was confirmed by MS^4 (Fig. S13). There, we identified both simple sialic acid losses at their canonical masses ($Z_{1β}$ and $Y_{1β}$), along with the modified losses of Galβ1-3


A
$Hex_1HexNAc_1Neu5Ac_1$,
m/z = 675
- $C_2H_6O_3$
α 6
β 3
m/z 597.21
<6.4%
>6.4%
α 3
β 3
α 6
β 3
Confidence: 82.4%
Coverage: 93.2%
Total test spectra: 1804
Confidence: 91.7%
Coverage: 79.1%
Total test spectra: 1715
B
$Hex_1HexNAc_1Neu5Gc_1$,
m/z = 691
- $C_2H_6O_4$
α 6
β 3
m/z 597.21
<14.3%
>14.3%
α 3
β 3
α 6
β 3
Confidence: 93.5%
Coverage: 95.6%
Total test spectra: 227
Confidence: 88.0%
Coverage: 83.0%
Total test spectra: 88
C
- $C_2H_6O_3$
α
?
?
m/z 800.2
<25.4%
>25.4%
β 4
α 3
β 3
α 6
?
β 6
β 3
α 3
α 6
?
β 3
Confidence: 80.0%
Coverage: 100%
Total test spectra: 132
Confidence: 100%
Coverage: 75.0%
Total test spectra: 128
D
- $C_2H_6O_4$
α 6
α 3
m/z 800.2
<58.5%
>58.5%
β 4
α 3
β 3
β 6
α 3
β 3
α 6
α 3
?
Confidence: 100%
Coverage: 89.8%
Total test spectra: 49
Confidence: 93.2%
Coverage: 100%
Total test spectra: 69
E
- $C_2H_6O_3$
α 6
β 3
m/z 597.213
normalized intensity (%)
0
20
40
60
80
100
200
300
400
500
600
m/z
181.97
Z_1,Y_1
203.97
$B_{1α}$- $C_2H_6O_3$
212.12
280.09
286.07
297.05
$Z_{1α}$
367.22
$Y_{1α}$
384.23
$Z_{1β}$- $C_2H_6O_3$
417.20
$Y_{1β}$- $C_2H_6O_3$
435.33
477.13
507.43
537.23
555.29
561.40
F
- $C_2H_6O_3$
α 6
β 3
?
m/z 800.303
normalized intensity (%)
0
20
40
60
80
100
400
500
600
700
m/z
330.23
$Z_{2α}$,$Y_{1β}$
366.10
384.13
$Z_{1α}$- $C_2H_6O_3$
417.13
$Y_{1α}$- $C_2H_6O_3$
435.13
$Z_{1β}$
569.20
$Y_{1β}$
587.23
$Y_{2α}$- $C_2H_6O_3$
597.23
680.27
710.27
740.23
758.23

◀**Fig. 4** A generalizable diagnostic fragment for Sia-HexNAc annotation. **A**, **B** Discriminatory performance of classifying Neu5Acα2-3Galβ1-3GalNAc and Galβ1-3(Neu5Acα2-6)GalNAc with the M-78 ($C_2H_6O_3$) fragment (**A**) and Neu5Gcα2-3Galβ1-3GalNAc and Galβ1-3(Neu5Gcα2-6)GalNAc with the M-94 ($C_2H_6O_4$) fragment (**B**). **C**, **D** Discriminatory performance of distinguishing topologies of $Neu5Ac_1Hex_1HexNAc_2$ with and without a sialylated reducing GalNAc with the M-78 ($C_2H_6O_3$) fragment (**C**) and distinguishing topologies of $Neu5Gc_1Hex_1HexNAc_2$ with and without a sialylated reducing GalNAc with the M-94 ($C_2H_6O_4$) fragment (**D**). **E**, **F** $MS.^3$ spectrum of the M-78 (-$C_2H_6O_3$) fragment produced by Galβ1-3(Neu5Acα2-6)GalNAc in sea bream mucin (**E**) and HexNAc?1-?Galβ1-3(Neu5Acα2-6)GalNAc in porcine gastric mucin (**F**)

arm ($Y_{1\alpha}$-$C_2H_6O_3$ and $Z_{1\alpha}$-$C_2H_6O_3$), confirming a similar fragmentation pattern across different structures sharing this motif. While such a triple loss event would not commonly be viewed as the most parsimonious annotation explanation, we here show that it is extremely potent (high confidence), common (high coverage), and generalizable (different structural contexts), underscoring the importance of a data-driven approach to identifying diagnostic fragments in glycomics annotation.

Discussion

Here, we presented a comprehensive resource of quantifiably performant and human-actionable rules for *O*-glycan isomer annotation based on interpretable machine learning. One of the main strengths of this work is that our annotation rules have been derived from a dataset composed of many experimental set-ups and analysts, who used different equipment (i.e., mass analyzers, collision energy, collision gas, etc.), and whose samples were present in different biological contexts, with different coeluting solutes and different solvents. Since these rules were then also validated and tested on such a diverse dataset, we can be confident that they present a more robust/performant foundation for annotation. We emphasize that our focus on coverage, typically the most neglected metric in identifying diagnostic fragments, ensures the generalizability and utility of our presented annotation rules. In principle, this process could then even be synergistically extended further, such as with retention time libraries for isomers [27], if a specific liquid chromatography context is constant for an analyst.

We are also optimistic about the promise of the herein presented workflow for further applications. In principle, the exact same workflow can be applied to the identification of similar diagnostic fragments or features for *N*-glycans, glycosphingolipids, or milk oligosaccharides. At least for some of those, the curated full dataset [11] could even be used as a data source, providing a clear and actionable implementation path. Similarly, due to the flexibility of our algorithms and CandyCrumbs [11], even data collected in, e.g., positive ion mode can be analyzed with this workflow. In general, we stress the importance of both annotation quality (influencing rule confidence) and data diversity, with regard to both annotators and instruments (influencing rule coverage). As with any machine learning approach, generalizing to unseen types of data can be challenging, so we advise caution in using our rules if a given set-up is not represented among, for instance, GlycoPOST data.

We are especially enthusiastic about future work identifying further generalizable diagnostic fragments for biologically relevant motifs, similar to our efforts with Sia-HexNAc here. One example here can be found with Lewis structures, such as Lewis A and X, which currently are often only distinguished by separately analyzing non-reduced glycans [21], due to the reliance on reducing end cross-ring fragmentation as diagnostic fragments.

We caution that the herein identified rules for isomer annotation are restricted to negative ion mode and, likely, reduced glycans. As mentioned above, these are not restrictions of the workflow per se but rather restrictions of the scope that we set out for this article and, hence, stem from the used dataset. A limitation partly arising from the workflow is the possibility of additional isomers that were not considered in this analysis. A classic example could be the analysis of non-mammalian glycans [28], which may exhibit different isomers than the ones considered here, which then invalidates the use of some of the herein presented rules. We thus would like to state that the rules identified here assume that the isomers in a given tree are the only isomers that are present in major abundances in a given sample. We also advise special caution if values for ratios or individual fragments are very close to the cut-off values provided by the rules, as error rates are expected to decrease with the distance to these cut-off values.

It is important to keep in mind that human annotations, which have been used to derive the rules here, are imperfect, which likely means that rule with 100% coverage/confidence should be theoretically unobtainable, on average. Still, for our workflow to remain valid, only the majority of the input assignments need to be correct, with erroneous assignments being considered as noise during the derivation of rules. Hence, we would expect that a rigorous application of high-performance rules to existing GlycoPOST data could even improve the average annotation quality and correct some structural assignments, which could be catalogued in a companion database, similar to how PDB-REDO refines the structural information of glycoproteins from the PDB [29].

As stated above, the preferred fragmentation pathway (ignoring collision energy as a modulator) is a function of glycan 3D structure, which then allows for the existence of diagnostic fragments to distinguish isomers in the first place. Hence, analyzing the 3D structure of isomers via molecular dynamics simulation could provide mechanistic explanations for diagnostic fragmentation, such as we have shown in previous work for distinguishing $HexNAc_2Neu5Ac_1$/$HexNAc_2Neu5Gc_1$ isomers [11]. We envision that understanding these processes mechanistically then holds the potential of identifying more general diagnostic fragments that generalize across sequences. We are convinced there still is a need for such fragments, especially when their performance is quantified such as here, which provides (i) a standardized set of annotation rules that (ii) attaches a confidence value to annotations and (iii) overall improves the quality of annotation, leading to a more robust foundation for engaging in biological exploration of *O*-glycomics data.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-024-05500-9.

Author contribution D.B., J.U., and R.J. conceived the method. J.U., R.J., and L.T. performed computational analyses. K.A.T. performed experimental analyses. D.B. and J.U. curated the datasets for this study. All authors prepared figures. D.B. supervised. D.B. acquired the funding for this study. All authors wrote and edited the manuscript.

Funding Open access funding provided by University of Gothenburg. This work was funded by a Branco Weiss Fellowship – Society in Science awarded to D.B., by the Knut and Alice Wallenberg Foundation, and the University of Gothenburg, Sweden. We thank SciLifeLab and BioMS (Swedish research council) for providing financial support to the Proteomics Core Facility, Sahlgrenska Academy.

Declarations

Competing interests The authors declare no competing interests.

References

1. Varki A. Biological roles of glycans. Glycobiology. 2017;27:3–49. https://doi.org/10.1093/glycob/cww086.
2. McMahon CM, Isabella CR, Windsor IW, Kosma P, Raines RT, Kiessling LL. Stereoelectronic effects impact glycan recognition. J Am Chem Soc. 2020;142:2386–95. https://doi.org/10.1021/jacs.9b11699.
3. Zhang Z, Shah B, Richardson J. Impact of Fc N-glycan sialylation on IgG structure. mAbs. 2019;11:1381–90. https://doi.org/10.1080/19420862.2019.1655377.
4. Bojar D, Meche L, Meng G, Eng W, Smith DF, Cummings RD, Mahal LK. A useful guide to lectin binding: machine-learning directed annotation of 57 unique lectin specificities. ACS Chem Biol. 2022;acschembio.1c00689. https://doi.org/10.1021/acschembio.1c00689.
5. Ashwood C, Lin C-H, Thaysen-Andersen M, Packer NH. Discrimination of isomers of released *N*- and *O*- glycans using diagnostic product ions in negative ion PGC-LC-ESI-MS/MS. J Am Soc Mass Spectrom. 2018;29:1194–209. https://doi.org/10.1007/s13361-018-1932-z.
6. Everest-Dass AV, Abrahams JL, Kolarich D, Packer NH, Campbell MP. Structural feature ions for distinguishing ***N***- and ***O***- linked glycan isomers by LC-ESI-IT MS/MS. J Am Soc Mass Spectrom. 2013;24:895–906. https://doi.org/10.1007/s13361-013-0610-4.
7. Doohan RA, Hayes CA, Harhen B, Karlsson NG. Negative ion CID fragmentation of *O*- linked oligosaccharide aldoses—charge induced and charge remote fragmentation. J Am Soc Mass Spectrom. 2011;22:s13361–011–0102–3. https://doi.org/10.1007/s13361-011-0102-3.
8. Karlsson NG, Schulz BL, Packer NH. Structural determination of neutral O-linked oligosaccharide alditols by negative ion LC-electrospray-MS ^{n}. J Am Soc Mass Spectrom. 2004;15:659–72. https://doi.org/10.1016/j.jasms.2004.01.002.
9. Jin C, Kenny DT, Skoog EC, Padra M, Adamczyk B, Vitizeva V, Thorell A, Venkatakrishnan V, Lindén SK, Karlsson NG. Structural diversity of human gastric mucin glycans. Mol Cell Proteomics. 2017;16:743–58. https://doi.org/10.1074/mcp.M117.067983.
10. Kawahara R, Chernykh A, Alagesan K, Bern M, Cao W, Chalkley RJ, Cheng K, Choo MS, Edwards N, Goldman R, Hoffmann M, Hu Y, Huang Y, Kim JY, Kletter D, Liquet B, Liu M, Mechref Y, Meng B, Neelamegham S, Nguyen-Khuong T, Nilsson J, Pap A, Park GW, Parker BL, Pegg CL, Penninger JM, Phung TK, Pioch M, Rapp E, Sakalli E, Sanda M, Schulz BL, Scott NE, Sofronov G, Stadlmann J, Vakhrushev SY, Woo CM, Wu H-Y, Yang P, Ying W, Zhang H, Zhang Y, Zhao J, Zaia J, Haslam SM, Palmisano G, Yoo JS, Larson G, Khoo K-H, Medzihradszky KF, Kolarich D, Packer NH, Thaysen-Andersen M. Community evaluation of glycoproteomics informatics solutions reveals high-performance search strategies for serum glycopeptide analysis. Nat Methods. 2021;18:1304–16. https://doi.org/10.1038/s41592-021-01309-x.
11. Urban J, Jin C, Thomsson KA, Karlsson NG, Ives CM, Fadda E, Bojar D. Predicting glycan structure from tandem mass spectrometry via deep learning. Nat Methods. 2024. https://doi.org/10.1038/s41592-024-02314-6.
12. Watanabe Y, Aoki-Kinoshita KF, Ishihama Y, Okuda S. GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. Nucleic Acids Res. 2021;49:D1523–8. https://doi.org/10.1093/nar/gkaa1012.
13. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.
14. Thomès L, Burkholz R, Bojar D. Glycowork: a Python package for glycan data science and machine learning. Glycobiology. 2021;cwab067. https://doi.org/10.1093/glycob/cwab067.
15. Joeres R, Blumenthal DB, Kalinina OV. DataSAIL: Data Splitting Against Information Leakage. 2023. https://doi.org/10.1101/2023.11.15.566305.
16. Lundstrøm J, Urban J, Thomès L, Bojar D. GlycoDraw: a python implementation for generating high-quality glycan figures. Glycobiology. 2023;cwad063. https://doi.org/10.1093/glycob/cwad063.
17. Bechtella L, Chunsheng J, Fentker K, Ertürk GR, Safferthal M, Polewski Ł, Götze M, Graeber SY, Vos GM, Struwe WB, Mall MA, Mertins P, Karlsson NG, Pagel K. Ion mobility-tandem mass spectrometry of mucin-type O-glycans. Nat Commun. 2024;15:2611. https://doi.org/10.1038/s41467-024-46825-4.
18. Thomsson KA, Benktander JA, Toxqui-Rodríguez S, Piazzon MC, Lindén SK. Gilthead seabream mucus glycosylation is complex,

differs between epithelial sites and carries unusual poly N-acetylhexosamine motifs. 2024. https://doi.org/10.2139/ssrn.4823066
19. Urban J, Joeres R, Thomès L, Thomsson KA, Bojar D. Navigating the maze of mass spectra: a machine-learning guide to identifying diagnostic ions in O-glycan analysis. 2024. https://doi.org/10.1101/2024.06.28.601175.
20. Domon B, Costello CE. A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconjugate J. 1988;5:397–409. https://doi.org/10.1007/BF01049915.
21. Jin C, Lundstrøm J, Korhonen E, Luis AS, Bojar D. Breast milk oligosaccharides contain immunomodulatory glucuronic acid and LacdiNAc. Mol Cell Proteomics. 2023;22:100635. https://doi.org/10.1016/j.mcpro.2023.100635.
22. Bennett AR, Lundstrøm J, Chatterjee S, Thaysen-Andersen M, Bojar D (2024) Ratios in disguise, truths arise: glycomics meets compositional data analysis. https://doi.org/10.1101/2024.06.09.598163.
23. Jin C, Padra JT, Sundell K, Sundh H, Karlsson NG, Lindén SK. Atlantic salmon carries a range of novel *O* -glycan structures differentially localized on skin and intestinal mucins. J Proteome Res. 2015;14:3239–51. https://doi.org/10.1021/acs.jproteome.5b00232.
24. Geiszler DJ, Polasky DA, Yu F, Nesvizhskii AI. Detecting diagnostic features in MS/MS spectra of post-translationally modified peptides. Nat Commun. 2023;14:4132. https://doi.org/10.1038/s41467-023-39828-0.
25. Ives CM, Singh O, D'Andrea S, Fogarty CA, Harbison AM, Satheesan A, Tropea B, Fadda E. Restoring protein glycosylation with GlycoShape. 2023. https://doi.org/10.1101/2023.12.11.571101.
26. Zhang T, Wang W, Wuhrer M, De Haan N. Comprehensive *O* -glycan analysis by porous graphitized carbon nanoliquid chromatography–mass spectrometry. Anal Chem. 2024;96:8942–8. https://doi.org/10.1021/acs.analchem.3c05826.
27. Abrahams JL, Campbell MP, Packer NH. Building a PGC-LC-MS N-glycan retention library and elution mapping resource. Glycoconj J. 2018;35:15–29. https://doi.org/10.1007/s10719-017-9793-4.
28. Staudacher E. Mucin-type O-glycosylation in invertebrates. Molecules. 2015;20:10622–40. https://doi.org/10.3390/molecules200610622.
29. Van Beusekom B, Lütteke T, Joosten RP. Making glycoproteins a little bit sweeter with *PDB-REDO*. Acta Crystallogr F Struct Biol Commun. 2018;74:463–72. https://doi.org/10.1107/S2053230X18004016.

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

James Urban is a PhD student in the group of Dr. Daniel Bojar at the University of Gothenburg, Sweden. There, he focuses on computational mass spectrometry and artificial intelligence, with a particular emphasis on complex carbohydrates or glycans.

Roman Joeres has been a PhD student at Saarland University since May 2022 and since May 2024 has been on a break to work as a researcher in the Bojar Lab. He mainly works on graph neural networks for drug target interaction prediction and explainability of deep learning models. Apart from research, he is active in teaching science to students and high schoolers.

Luc Thomès is a research engineer at the Université de Lille, France, specializing in developmental disease studies using sequencing data. He holds a PhD in bioinformatics, where he studied vertebrate stress response evolution, and completed postdoctoral research in glycobioinformatics and machine learning at the University of Gothenburg, Sweden, in the team of Dr. Daniel Bojar. His work integrates computational methods to explore systemic biological processes, from molecular interactions to organismal evolution, in both fundamental and medical research.

Kristina A. Thomsson is a postdoctoral researcher working at the Proteomics Core Facility at Gothenburg University and also associated with Professor Sara Lindén's research group. She works with sample preparation and LC/MS analyses of glycans as well as developing protocols and tools to improve glycan sequencing and semiquantitation using mass spectrometry.

Daniel Bojar became Assistant Professor of Bioinformatics at the University of Gothenburg, Sweden, in January 2021. His research is focused on developing and applying artificial intelligence and data science to better understand the biological roles of complex carbohydrates or glycans. He has pioneered AI-driven glycobiology and was featured on the 2022 *Forbes* 30 Under 30 Europe list for work in Science & Healthcare.