Direct Glycan Analysis of Biological Samples and Intact Glycoproteins by Integrating Machine Learning Driven- Surface-Enhanced Raman Scattering (SERS) and Boronic Acid Array

Qiang Hu[1], Hung-Jen Wu[1,*]

[1] The Artie McFerrin Department of Chemical Engineering, Texas A&M University, TX 77843, USA

* Corresponding author

Abstract

Frequent monitoring of glycan patterns is a critical step for studying glycan-mediated cellular processes. However, the current glycan analysis tools are resource-intensive and less suitable for routine use in standard laboratories. We developed a novel glycan detection platform by integrating surface-enhanced Raman spectroscopy (SERS), boronic acid (BA) receptors, and machine learning (ML) tools. This sensor monitors the molecular fingerprint spectra of BA binding to cis-diol-containing glycans. Different types of BA could yield different stereoselective reactions toward different glycans and exhibit unique vibrational spectra. By integrating the Raman spectra collected from different BA receptors, the structural information can be enriched, eventually improving the accuracy of glycan classification and quantification. Here, we established a SERS-based sensor incorporating multiple different BA receptors. This sensing platform could directly analyze the biological samples, including the whole milk and intact glycoproteins (fetuin and asialofetuin), without tedious glycan release and purification steps. The results demonstrate the platform's ability to classify milk oligosaccharides with remarkable classification accuracy, despite the presence of other non-glycan constituents in the background. This sensor could also directly quantify sialylation levels of fetuin/asialofetuin mixture without glycan release procedures. Moreover, by selecting appropriate BA receptors, the sensor exhibits an excellent performance of differentiating between α2,3 and α2,6 linkages of sialic acids. This low-cost, rapid, and highly accessible sensor will provide the scientific community with an invaluable tool for routine glycan screening in standard laboratories.

Introduction:

Glycans are highly abundant biomolecules found in all living organisms. They form dense layers on cell membranes, proteins, and other biomolecules, facilitating a wide range of biochemical reactions[1]. The glycosylation processes are highly sensitive to various factors, such as environmental conditions, cell activities, nutrition, cell growth cycles, cell health, etc[2, 3]. To investigate glycan-mediated cellular processes, frequent monitoring of glycan changes in biological samples is essential. However, glycan analysis poses significant challenges due to their intricate structures, including complex isomeric forms, glycosidic linkages, and branched structures.

The comprehensive glycan sequencing tools, such as mass spectrometry (MS)-based techniques, are effective but highly resource-intensive; therefore, these methods are less suitable for routine use in standard laboratories[4-6]. Staining samples with lectins (i.e., glycan binding proteins) is another popular tool for comparative glycan analysis. Because the protocol is relatively simple, lectin staining can be applied in standard laboratories[7]. However, lectin-glycan interactions are not highly specific[8-11]; a lectin often binds to various glycan structures with different affinities[12-14]. In addition, the relatively small lectin library could not cover all of the essential glycan structures. Thus, there is a growing demand for a low-cost, rapid, and highly accessible tool that empowers researchers to frequently monitor dynamic changes in glycosylation[4, 6].

Our prior research developed a machine learning (ML)-driven surface enhanced Raman spectroscopy (SERS) sensor capable of classifying the selected glycans with remarkable accuracy exceeding 99%[15]. This new sensing platform includes three major components: (1) boronic acid (BA) receptors, (2) SERS, and (3) ML program. (Scheme shown in Figure 1) BA can reversibly bind with cis-diol-containing carbohydrates, leading to the formation of boronate esters[16]. It is worth noting that BA binding is not highly specific to a particular glycan structure. Thus, the classic yes/no confirmative response (e.g., fluorescence) in the staining assay is not sufficient to distinguish different glycan structures. To further identify the glycan structures, the sensor monitors molecular fingerprint spectra of BA-glycan reaction complex.

Raman spectroscopy was chosen to monitor molecular fingerprint spectra for several reasons. First, Raman spectroscopy not only provide fingerprint information of molecules but also can distinguish isomeric structures, allowing for isomeric glycan detection[17]. Second, the availability of low-cost Raman spectrometers will enable widespread adoption[18]. Additionally, a low-cost plasmonic SERS substrate (<$0.08 per test), called nanopaper, are used to enhance Raman signals[19, 20].

In our previous work, we had evaluated two commercially available BA receptors, namely 4-mercaptophenylboronic acid (4MBA) and 1-thianthrenylboronic acid (1TBA), which effectively captured glycan molecules through cis-diol chemical reactions[15]. Utilizing advanced ML algorithms, we analyzed spectral variations across a wide frequency range for glycan detection. This sensor successfully distinguished stereoisomers and structural isomers featuring different glycosidic linkages.

One of the key observations in our previous study is that the structural information of glycans can be enriched by integrating the spectra obtained from 4MBA and 1TBA. The collective Raman spectra could increase the accuracy of glycan classification and quantification. This discovery offers a strategy to improve the sensor performance by using an array of BA receptors. Here, we developed a glycan sensor containing up to 8 different BA receptors to directly analyze the complex biological samples without glycan purification steps, including whole milk and intact glycoproteins. Because SERS is a near-field phenomenon, the resulting Raman signals primarily originate from BA-glycan reaction complexes that are directly adsorbed on metallic nanoparticles. The influences of the background matrices were minimal. As expected, the collective spectra from different BA receptors achieved a remarkable 100% accuracy for classifying the milk oligosaccharides in the commercial dairy products.

In addition, we used this platform to directly analyze intact glycoproteins. Protein glycosylation analysis typically requires tedious sample preparation, such as enzymatically releasing glycans from glycoproteins and chemically labeling glycans for detection[21-23]. Direct analysis of intact glycoproteins will speed up the analysis procedure and benefit the glycobiology community. We evaluated the feasibility of quantifying sialylation levels of fetuin/asialofetuin mixture. The collective spectra once again demonstrated superior quantification performance, with R^2 and normalized mean square error (NMSE) values of 0.9941 and 0.005912, respectively. Moreover, we evaluated the sensor performance in distinguishing glycosylic linkages of sialic acids. Sialic acid is an important monosaccharide for mammalians due to its functionality in nervous system development, immune regulation, and involvements in many diseases[24-26]. We tested the sensor's classification performance on the most common α2,3 and α2,6 sialic acid linkage[27, 28]. By using various BA receptors, the sensor could achieve 100% accuracy in the classification of sialic acid linkages. In summary, this sensor could work as a user-friendly platform to directly detect glycan profiles in biological samples and intact glycoproteins without time-consuming glycan purification steps.

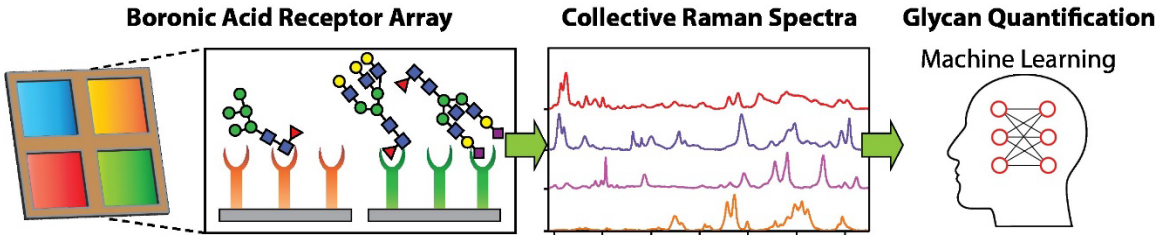

Figure 1. Schematic of the sensor. Glycans are captured by various boronic acids (BAs) printed on the SERS substrates. Each BA-glycan reaction complex offers a unique molecular fingerprint spectrum. The structural information is enriched by integrating the fingerprint spectra from different BAs, called the collective Raman spectra. The complex collective spectra are processed by the advanced machine learning technique for glycan classification and quantification.

Methods:

Materials:

2-(N-morpholino)ethanesulfonic acid (MES), fetuin, asialofetuin, 2,3 sialyllactose (3-SLA), *N*-acetylneuraminic acid (Neu5Ac/sialic acid), 2,6 sialyllactose (6-SLA), 4-mercaptophenylboronic acid (4MBA), 1-thianthrenylboronic acid (1TBA), 3-mercaptophenylboronic acid (3MBA), 4-aminophenylboronic acid (4ABA), pyridine-4-boronic acid (PyriBA), pyrene-1-boronic acid (PyreBA), 2-(hydroxymethyl)phenylboronic acid cyclic monoester (HBACM), and benzo[b]thien-2-ylboronic acid (BBA) were purchased from Sigma Aldrich (boronic acid structure are shown in SI Figure 1). Glass microfiber paper (GF-C, binder free, 100 mm circles) was acquired from Whatman. Silver nitrate (99.99995%), ammonia, dextrose and 2-propanol were purchased from ThermoFisher scientific. All milk products were purchased from local markets. All chemicals were ACS grade or higher and used without further purification.

Nanopaper fabrication:

Nanopapers were fabricated as previously reported[15, 19, 20]. In brief, Tollens' reagent containing 300 mM ammonia and 50 mM silver nitrate was prepared in a 2 L glass beaker in a 55 °C water bath. Glass microfiber papers were immersed in the solution, and 500 mm glucose solution was added to initiate the silver mirror reaction. After the reaction was complete, the filter papers were rinsed thoroughly with deionized water and 2-propanol. The resulting products, i.e., the nanopapers, were stored in 2-propanol at room temperature. The storage container was covered with aluminum foil and placed in drawers to prevent light exposure.

Surface modification

The surface modification was performed as previously reported[15]. In brief, nanopapers were cut into a 1 cm × 0.5 cm rectangular shape, and then immersed in 50 mM 1TBA or 0.1 mM for other boronic acids (BAs) in methanol for 1 hour. Before glycan measurement, the BA-coated nanopaper was air-dried. For the glycan measurement, the BA-coated nanopapers were spotted with the aqueous solutions containing glycans or glycoproteins in 100 mM MES buffer (pH 5) and incubated for 1 hour. Before Raman measurement, the paper was dried in an oven at 75°C for 5 minutes.

Raman measurement

Raman spectra were collected with a ThermoFisher Scientific DXR3 Raman microscope using laser excitation with a wavelength of 785 nm and an output power of 1 mW. This instrument was equipped with an Olympus BX41 optical microscope and a

thermoelectrically cooled charge-coupled detector (ThermoFisher front-illuminated CCD system) with 1024 × 256 pixel format, operating at −70 °C. The signal was calibrated by an internal polystyrene standard and a 10× objective. The spot size was about 3.8 μm. 200 SERS spectra were collected with an exposure time of 1 s for 5 accumulations at different spots for each sample.

Milk oligosaccharide extraction and classification

The milk oligosaccharide was extracted using the traditional Folch extraction with slight modification[29]. Briefly, the milk was mixed with chloroform and methanol mixture (3:1, v/v) in a 1:4 (milk v/solvent v) ratio in the 50mL Nalgene™ Oak Ridge high-speed PTFE FEP centrifuge tubes. The mixture was shaken vigorously for 5 minutes until homogeneous, followed by 40 minutes of centrifuge at 4000 RPM. The upper layer of the solution was extracted and concentrated in the rotary evaporator under 55 ℃ until all solvents were evaporated. Total carbohydrate concentration was determined using the phenol-sulfuric acid assay[30]. The absorbance of the carbohydrate assay was detected by microplate reader (BMG Labtech FLUOstar Omega). After that, the oligosaccharide extracts were stored at -20 ℃ until usage. Three extractions were performed on each milk type on different days.

The dried milk oligosaccharide extracts were reconstituted with DDI water to reach the final total carbohydrate concentration of 0.34 mg/mL, which is the same concentration as the 1 mM lactose solution. The reconstituted oligosaccharide samples were measured under the same protocol as above, with 600 spectra collected for each milk type. Before data analysis, the spectra were averaged for three batches, resulting in 200 spectra for each milk type.

Whole milk glycan profiling

20 μL of cow milk, goat milk, oat milk, soy milk, or almond milk was spotted onto surface-modified nanopaper using micropipettes and left for 1 hour. The paper was then rinsed with 100 mM MES buffer (pH 5) to remove unbound proteins and non-glycan contents. Subsequently, the paper dried in an oven at 75℃ for 5 minutes and was measured by Raman spectrometer using the protocol reported above.

Intact protein quantification

Different ratios of 1 mM fetuin and 1 mM asialofetuin in 100 mM MES buffer (pH 5) were mixed at different ratios to prepare a titration curve ranging from 0% to 100% of 1 mM fetuin, with 20% intervals. The glycoprotein mixtures were spotted on the BA-coated nanopapers and incubated at room temperature for 1 hour. The nanopaper was then dried

in an oven at 75°C for 5 minutes and measured by Raman spectrometer using the same protocol as above.

Sialic acid linkage identification

Aqueous solutions of 1 mM of 2,3 sialyllactose (3-SLA), 1 mM of 2,6 sialyllactose (6-SLA), and an equal volume mixture of sialic acid (0.5 mM) and lactose (0.5 mM) were prepared to represent α2,3 sialic linkage, α2,6 sialic acid linkage, and no linkage between sialic acid and other glycans, respectively. The samples were spotted on the BA-coated nanopapers and incubated at room temperature for 1 hour. Afterward, the nanopaper was then dried in an oven at 75°C for 5 minutes and measured by Raman spectrometer using the same protocol reported above, with 200 spectra collected for each concentration.

Data processing

The data analysis was performed using the same methodology reported in the previous study[15]. Briefly, the spectra were first processed using asymmetric least square (ALS) baseline correction. Then, baselined spectra were vector normalized and smoothed using Savitzky–Golay filtering (4^{th} order polynomial, with a frame size of 37). Finally, multivariate analysis techniques and classification algorithms were applied in the spectral range 400-1650 cm^{-1}. Data processing was carried out using Matlab 2021b.

Multivariate analysis and machine learning

Prior to applying classifiers, the smoothed spectra underwent multivariate statistical analysis to reduce complexity and extract significant spectral features explaining the most variance. Discriminant analysis of principal components (DAPC) was used for this purpose[31]. Principal component analysis (PCA) was initially applied to reduce the data complexity, and then, a supervised analysis process, discriminant analysis, was used to further discriminate the dataset by correlating data variation with the sample information.

After feature extraction, common machine learning classifiers were used to classify SERS spectra. Support vector machine (SVM) was selected due to its superior performance in the prior Raman study[32]. A 5-fold cross-validation was conducted to assess the suitability of the classification algorithm[33]. In brief, the training and the validation sets were established by randomly selecting from the Raman spectral data. The training dataset was used to generate a classification model, and the model predicted the validation dataset to evaluate the performance. The cross-validation approach was repeated five times, wherein the validation set consists of and 800 and 480 randomly selected SERS spectra in repetition for the whole milk glycan study and sialic acid linkage study, respectively. The model's performance was evaluated by using classification accuracies, sensitivity, and selectivity. The collective spectra were constructed by combining the truncated boronic acid spectra

(400-1650 cm^{-1}). Then, the collective spectra went through the same multivariate analysis and classification algorithm as the individual spectra.

Regression analysis was conducted using Matlab 2021b. The gaussian process regression model was used to predict the percentage fraction of fetuin within the fetuin/asialofetuin mixture. DAPC was first performed on the dataset for the wavenumber from 400-1650 cm^{-1}, and then the resulting canonicals were used in regression analysis. A 5-fold cross-validation was performed on the model to evaluate the regression performance. The model was evaluated based on the normalized mean square error (NMSE) and coefficient of determination (R^2). For the collective spectra regression, the dataset was built in the same way as described in the classification. Then the spectra went through the same regression algorithm and were evaluated based on the same performance metrics (NMSE and R^2).

Statistical analysis

The data analysis was performed using the same methodology reported in the previous study[15]. For classification tasks, the performance was evaluated by accuracy, sensitivity, and selectivity. The classification accuracy, sensitivity, and selectivity are defined as:

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{All cases}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive+ False negative cases}} \quad (2)$$

$$\text{Selectivity} = \frac{\text{True negative}}{\text{True negative+ False positive}} \quad (3)$$

For quantitative analysis, the performance was evaluated based on the normalized mean square error (NMSE) and coefficient of determination (R^2).

$$\text{NMSE} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

where $\hat{y}_i$ is the predicted values, y_i is the actual values in the dataset, $\bar{\hat{y}}$ is the mean of the predicted values, $\bar{y}$ is the mean of the actual values, and the n is the number of spectra in the dataset.

Result and Discussion

Direct analysis of unprocessed milk samples

Our previous study demonstrated ML-driven SERS platform exhibits exceptional performance on identifying purified glycans[15]. However, the additional purification process is time-consuming and may result in the loss and degradation of glycans. A new approach allowing direct analysis of unprocessed biological samples and intact glycoproteins will benefit the scientific communities. For the proof-of-concept, we first evaluate the feasibility of analyzing milk oligosaccharides in the unprocessed milk samples.

Milk oligosaccharides are pivotal nutrients in the human health[34]. For example, sialic acids in milk are critical for supporting infant body development[35, 36]. Researchers commonly use liquid chromatography-mass spectrometry (LC-MS) to profile N-glycans in milk[34, 37]. However, the tedious sample preparation and time-consuming testing process limit its widespread use. In our previous study, we have demonstrated the integration of BAs, SERS, and ML program could identify and quantify oligosaccharides extracted from milks. Our approach offers a valuable platform for detecting milk adulteration. However, the oligosaccharide extraction process is time- and labor-intensive. We are now taking a step further to investigate the feasibility of detecting unprocessed milk samples. Direct detection not only reduces the processing time but also eliminates experimental variations during the oligosaccharide extraction process.

Because SERS is a near-field effect, we hypothesize that the major Raman signals are contributed by BA-glycan reaction complexes that directly attach to SERS substrates. To verify the hypothesis, we compared the SERS spectra of extracted oligosaccharides and unprocessed milk samples on 4MBA functionalized substrates (Cow milk example shown in Figure 2 & a detailed comparisons among other milk samples in SI Figure 2). Milk was dropped onto a surface-modified nanopaper, and then the nanopaper was rinsed with the buffer to remove unbound proteins and non-glycan contents. The spectral difference between processed and unprocessed milk samples is minimal. The small spectral differences were observed at 470 cm^{-1} (CCC out of plane bending), 607 cm^{-1} (CCC in-plane bending), 1075 cm^{-1} (CCC in-plane bending, CS stretching), and 1589 cm^{-1} (CC stretching, CH bending). Since the milk oligosaccharides were extracted using Folch method, lipids and proteins were removed during the separation process[38]. These minor spectral variations are likely contributed by the additional glycan compounds, such as glycoproteins and glycolipids in the unprocessed milk samples. The similarity of SERS spectra between extracted oligosaccharides and unprocessed samples suggests that BA is capable of capturing milk oligosaccharides and the major SERS signals were contributed by BA-glycan reaction complexes, despite the presence of other non-glycan constituents in the background. This discovery allows us to eliminate the time-consuming glycan extraction procedure in the detection protocol.

After demonstrating the influence of background molecules is minimal, we evaluated the sensor's capability of classifying commercial dairy products, including cow, goat, soy, oat, and almond milk. In our previous research, we demonstrated that the collective SERS

spectra from different BA receptors could enrich the structural information and improve the classification accuracy[15]. To improve the sensor performance, we establish an array of BA receptors. This array consists of 5 different BAs, including two BAs used in our previous study (4MBA, 1TBA) and three additional BAs (3MBA, 4ABA, and PyriBA). The position difference of the mercapto group (3MBA vs. 4MBA), the substitution of mercapto group with amine (4MBA vs. 4ABA), and the incorporation of pyridine group in PyriBA could result in distinct molecular vibrations, leading to unique Raman spectra of BA-glycan reaction complexes. The Raman spectra and confusion matrices of these BAs were shown in SI Figure 3 ~ SI Figure 12. Among these BAs, 4MBA exhibited the best performance but still misclassified one sample from all cases. The collective spectra of 5 boronic illustrate the remarkable 100% classification accuracy. (SI Figure 13)

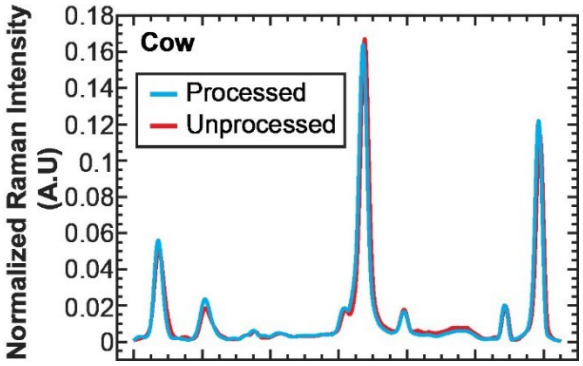

Figure 2. Average normalized SERS spectra (n=200) of the unprocessed cow milk sample and the purified milk oligosaccharides on 4MBA coated substrates. The milk oligosaccharides were extracted from cow milk using Folch method. The spectral difference between processed and unprocessed samples is minimal.

Direct analysis of intact glycoproteins

The current techniques for glycoprotein analysis require tedious sample preparation, including enzymatically releasing glycans from glycoproteins and chemically labeling glycans for detection[21-23]. For example, the glycosidase (PNGase F/PNGase A) could be used to cleave the N-glycans from glycoproteins, and then the glycans are purified via hydrophilic interaction liquid chromatography (HILIC)[39, 40]. According to the detection techniques, chemical labeling of the glycans may be required for liquid chromatography (LC) or liquid chromatography-mass spectrometry (LC-MS)[40, 41]. For O-glycan analysis, the procedure can be more complicated due to the lack of enzymatic cleavage methods. While chemical release approaches exist, there is a risk of glycan structure degradation during the release process[42]. Therefore, direct analysis of intact glycoproteins without tedious sample preparation is highly desired. For the proof-of-concept, we selected fetuin and asialofetuin as a model system[43]. Bovine fetuin is known to contain three N-glycosylation sites and five O-glycosylation sites[44], while asialofetuin shares the same protein structure and glycosylation sites but lacks terminal sialic acids[45]. Sialic acid plays a crucial role in the central nervous system and the immune system, making it an essential glycan building block[24]. Here, we used our sensing platform to directly quantify sialylation levels of fetuin/asialofetuin mixtures without sample pretreatment.

Figure 3 displays normalized SERS spectra of fetuin and asialofetuin binding to 4MBA, and the spectra of 1TBA, 3MBA, 4ABA, and PyriBA are shown in SI Figure 14. For 4MBA spectra, sialic acid residues in fetuin cause the peak shifts to higher wavenumber and the increases of intensity at 475 cm^{-1} (CCC out of plane bending) and 613 cm^{-1} (CCC in plane bending). Similarly, the presence of sialic acid residues also results in the variations of signal intensities at 1015 cm^{-1} (CC stretching, OH stretching) and at 1589 cm^{-1} (CC stretching, CH bending). These spectral changes are consistent with the data of sialic acid monosaccharide observed in our prior study[15]. For 1TBA, the spectra differences among fetuin, asialofetuin, and negative control (no glycoprotein) are more significant. With sialic acid residue, the lower spectral signals were observed at 431 cm^{-1} (CCCC torsion, SCCC out-of-plane bending), 669 cm^{-1} (CCCC torsion, CCC in-plane bending), 1037 cm^{-1} (CC stretching), 1125 cm^{-1} (CC stretching, HCC bending), and 1194 cm^{-1} (CC stretching). In contrast, the signals raise at 1081 cm^{-1} (CC stretching) and 1554 cm^{-1} (CC stretching) when sialic acid residues present. This changes are consistent with our prior observations of sialic acid[15].

Since the presence of sialic acid residues in SERS spectra was observable, we evaluated the capability of ML tool to quantify sialic acid levels in the samples containing both fetuin and asialofetuin. Fetuin and asialofetuin were mixed with various molar ratios in total of 1 mM, and SERS spectra of the mixed samples were collected on the nanopapers functionalized with different BAs. The sialic acid levels were quantified using Gaussian regression models. SI Figure 15 illustrates the regression results using 4MBA, while SI Figure 16 to SI Figure 19 show the results with 1TBA, 3MBA, 4ABA, and PyriBA. Among the five BAs, 3MBA exhibit the best quantification performance, with R^2 of 0.9920 and NMSE of 0.007982. This is probably because more spectral differences between fetuin and asialofetuin were observed on 3MBA substrates, such as the peak intensity change around 785 cm^{-1} and 999 cm^{-1}. We also quantified the sialic acid levels using the collective spectra. (Figure 4). As expected, the quantification result from the collective spectra of five BA spectra is better than the analysis from the individual BA spectra.

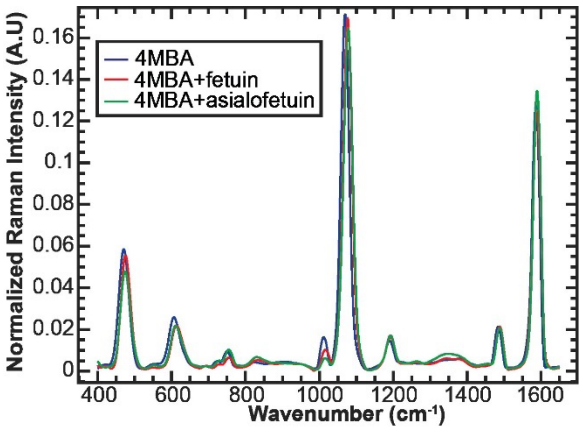

Figure 3. Average normalized SERS spectra (n=200) of fetuin and asialofetuin on 4MBA coated substrates. Differences could be observed among 4MBA, 4MBA+fetuin, and 4MBA+asialofetuin spectra.

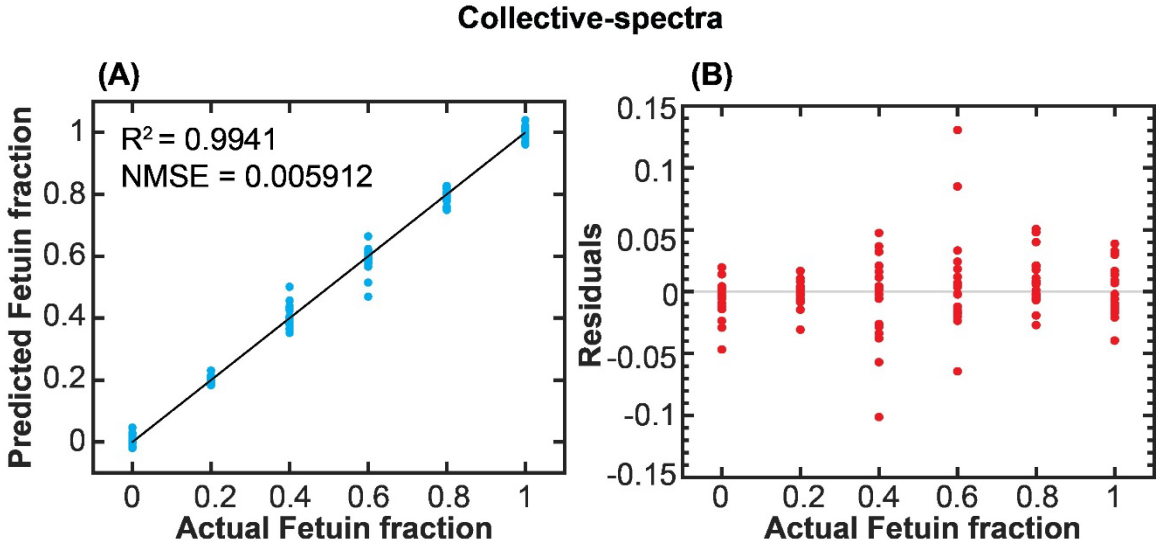

Figure 4 Predict fetuin fraction vs. actual fetuin fraction (A) and residual vs. actual fetuin fraction (B) for collective-spectra (4MBA, 1TBA, 3MBA, 4ABA, PyriBA) regression on fetuin fractions.

Sialic acid linkage identification

We have demonstrated this sensing platform can quantify sialylation levels of intact glycoproteins (mixtures of fetuin and asialofetuin). In our quest for deeper insights, we evaluate the capability of this sensing platform for distinguish sialic acid linkages. Sialic acids could link with other glycan entities in multiple linkage forms (α2,3, α2,6, α2,8 or α2,9)[25, 27, 28], and the most common linkages for sialic acids to other glycans are α2,3 and α2,6[25]. It is crucial to differentiate sialic acids glycosidic linkages since the linkages could influence biological activities. For example, α2,3 linkage may promote selectin binding, and it is related to several cancers and higher patient death rates[46]. Cancer cells are known to have high levels of α2,6-linked sialic acid with galactose[25] and α2,6 linkage can block

the apoptosis inducing galectin proteins interactions with glycan, which improves cell survival[27]. The sialic acid linkage also impacts the anti-inflammatory properties of IgG. IgG Fc fragments containing α2,6-linked sialic acid only show a 10-fold increase in anti-inflammatory activity compared to those containing both α2,3- and α2,6-linked sialic acids[47]. In contrast, Fc fragments containing only α2,3-linked sialic acid show no anti-inflammatory activity at all[47]. It is crucial to ensure that the IgG antibodies used for anti-inflammatory treatments have the appropriate sialic acid linkages through quality control. IgG could be produced by HEK or CHO cell lines. However, CHO cell lines only produce α2,3-linked sialic acids while HEK cell lines could produce both types of linkage[48]. As such, identification of the sialic acid linkage is crucial for quality control in the pharmaceutical industry. However, conventional LC-MS struggles to distinguish between α2,3 and α2,6 linkages since they share the same molecular weight and result in identical m/z values[49, 50]. Having the ability to distinguish linkages with our platform will significantly benefit the pharmaceutical industry and the glycobiology community.

The stereoselective reactions between BAs and glycans are determined by the spatial orientations and intermolecular distance between BA moieties[51, 52]. The recognition of α2,3 and α2,6 linkages could be improved by selecting appropriate BA receptors. To explore this, we introduced three new BAs: PyreBA, HBACM, and BBA, alongside the five BAs that are employed in the previous sections (4MBA, 1TBA, 3MBA, 4ABA, PyriBA) (Structures are in SI Figure 1). PyreBA contains four aromatic rings, known for their distinctive Raman vibrations[53]. Using PyreBA as a model helps us understand the role of aromatic ring vibrations in our platform. HBACM has been previously used in detecting monosaccharides, such as glucose and fructose, as well as oligosaccharides like stachyose and nystose[54, 55]. Similarly, BBA closely resembles the structure of HBACM, with the exception of a sulfur atom replacing the oxygen atom.

We examined the performance of the selected BAs in differentiating between the following samples: 2,3 sialyllactose (3-SLA), 2,6 sialyllactose (6-SLA), and sialic acid monosaccharide mixed with lactose (SA+Lac). (structure information are shown in SI Figure 1). 3-SLA consists of a sialic acid residue attached to galactose in lactose molecule with α2,3 linkage, and 6-SLA has a α2,6 linked sialic acid with lactose. To observe the influence of sialic acid without linkage, we mixed the equal molar concentrations of sialic acid monosaccharides with lactose.

SI Figure 20 and SI Figure 21 illustrate the confusion matrix and spectra for 4MBA and 1TBA. Both 1TBA and 4MBA successfully distinguished the existence of glycosidic linkages. However, when it came to distinguishing the two different linkages, 4MBA outperformed 1TBA. The difference may be attributed to the poor performance of 1TBA in lactose identification shown in the previous study[15]. When we analyzed the collective spectra, which incorporated the spectral data from both 4MBA and 1TBA, the accuracy was improved. SI Figure 22 and SI Figure 23 showcase the confusion matrix and spectra for 3MBA, 4ABA, and PyriBA, while SI Figure 24 and SI Figure 25 present the confusion matrix and spectra for the new BAs, including PyreBA, HBACM, and BBA. Notably,

PyreBA did not exhibit strong classification performance. Conversely, both HBACM and BBA delivered satisfactory results compared to the other tested BAs. Of particular interest, 3MBA, 4ABA, and BBA achieved a remarkable 100% accuracy. Finally, Figure 5 shows the confusion matrix for the collective spectra of 8 boronic acids for sialic acid linkage identification, which resulted in 100% accuracy as well. This discovery demonstrated that the detection performance of the specific glycan structures can be improved by selecting appropriate BA receptors, as well as the potential of using up to 8 boronic acids as an array for glycan profiling.

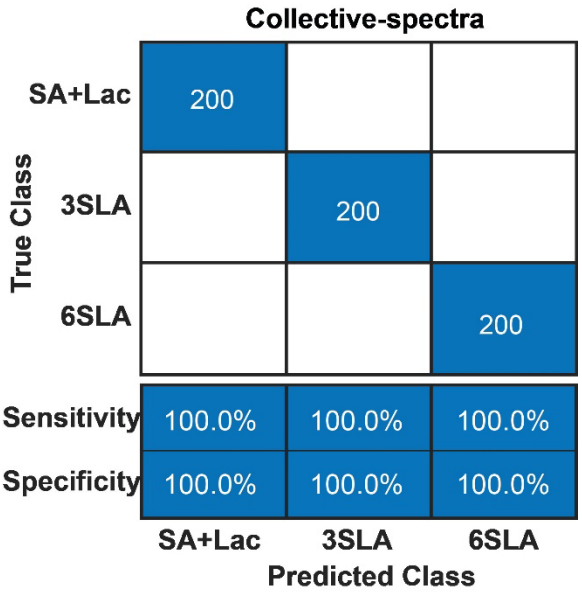

Figure 5 Confusion matrix of sialic acid linkage classification using collective spectra method (100% accuracy) with all 8 boronic acids (4MBA, 1TBA, 3MBA, 4ABA, PyriBA, PyreBA, HBACM, and BBA).

Conclusion

We have introduced a ML-driven SERS glycan sensor capable of classifying and quantifying purified glycans with high accuracy[15]. Expanding upon this foundation, we assessed the platform's ability to profile glycans in the presence of non-glycan entities. Because SERS is a near-field phenomenon, the Raman signals are majorly contributed by BA-glycan reaction complexes directly attached to the SERS substrates, and the influences of background matrices are minimal. We successfully classified glycans in unprocessed milk samples as well as quantified the sialylation level of intact glycoproteins (the mixture of fetuin and asialofetuin). This discovery allows us to directly analyze biological samples without time-consuming glycan releasing and extraction procedures. The elimination of sample preparation steps would minimize the loss and degradation of glycans, eventually reducing experimental variations.

BAs can reversibly bind with cis-diol in glycan molecules. By controlling spatial orientations and intermolecular distance between BA moieties, BAs can bind to different pairs of hydroxyl groups on a glycan with different binding affinities[51, 52]. Different types of BAs could yield different stereoselective reactions. In addition, most BAs contain Raman active structures, so the BA-glycan reaction complexes could exhibit unique and strong Raman spectral shifts. We hypothesize that detection accuracy can be improved by integrating the Raman spectra collected from different BA receptors. To validate the hypothesis, we established a sensor containing five different BAs to classify milk oligosaccharides and quantify the sialylation levels of fetuin/asialofetuin mixtures. As expected, the detection accuracy was significantly improved by integrating the spectral data obtained from different BA receptors. This result offers a systematic strategy to improve the sensor performance when the complexity of glycan sample increases.

We also explored how BA structures influence the detection of the glycosidic linkages. We examined eight BA receptors containing various functional groups and aryl structures. The appropriate BA receptors, including 3MBA, 4ABA, and BBA, could exhibit an excellent performance of differentiating between α2,3 and α2,6 linkages of sialic acids. This discovery suggested that the careful selection of BA receptors is crucial to improving the detection accuracy.

In summary, the combination of SERS, BA receptors, and ML-driven chemometrics offers a rapid and efficient approach for comparative glycan detection. This study demonstrated this sensing platform could directly analyze unprocessed biological samples and intact glycoproteins without glycan purification steps. We also demonstrate that the detection accuracy can be improved by using multiple BA receptors. This sensor can serve as a rapid, low-cost, and valuable tool for routine glycosylation analysis in standard laboratories.

Supporting Information

Peak assignments for 4MBA and 1TBA, Glycan and boronic acid structure, Scheme for machine learning, SERS spectra for monosaccharides, electrostatic potential distribution from DFT simulation, PCA contribution plots and confusion matrix for classification cases, regression model performance plots (PDF)

Author Contributions

Q. Hu conducted the experiments and analyzed the data. Q. Hu and H.-J. Wu conceived idea and wrote the manuscript.

Acknowledgements

The authors gratefully acknowledge the support from National Science Foundation (CBET-2114203).

Notes

The authors declare no competing financial interest.

References

(1) Gagneux, P.; Hennet, T.; Varki, A. Biological Functions of Glycans. *Essentials of Glycobiology [Internet]. 4th edition* **2022**.

(2) Costa, A. R.; Rodrigues, M. E.; Henriques, M.; Oliveira, R.; Azeredo, J. Glycosylation: impact, control and improvement during therapeutic protein production. *Critical Reviews in Biotechnology* **2014**, *34* (4), 281-299. DOI: 10.3109/07388551.2013.793649.

(3) Viinikangas, T.; Khosrowabadi, E.; Kellokumpu, S. N-Glycan Biosynthesis: Basic Principles and Factors Affecting Its Outcome. In *Antibody Glycosylation*, Pezer, M. Ed.; Springer International Publishing, 2021; pp 237-257.

(4) Cummings, R. D.; Pierce, J. M. The challenge and promise of glycomics. *Chem Biol* **2014**, *21* (1), 1-15. DOI: 10.1016/j.chembiol.2013.12.010 From NLM Medline.

(5) Marino, K.; Bones, J.; Kattla, J. J.; Rudd, P. M. A systematic approach to protein glycosylation analysis: a path through the maze. *Nat Chem Biol* **2010**, *6* (10), 713-723. DOI: 10.1038/nchembio.437 From NLM Medline.

(6) Mimura, Y.; Katoh, T.; Saldova, R.; O'Flaherty, R.; Izumi, T.; Mimura-Kimura, Y.; Utsunomiya, T.; Mizukami, Y.; Yamamoto, K.; Matsumoto, T.; Rudd, P. M. Glycosylation engineering of therapeutic IgG antibodies: challenges for the safety, functionality and efficacy. *Protein Cell* **2018**, *9* (1), 47-62. DOI: 10.1007/s13238-017-0433-3 From NLM Medline.

(7) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. Essentials of Glycobiology. In *Essentials of Glycobiology*, Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., Etzler, M. E. Eds.; 2009.

(8) Veillon, L.; Huang, Y.; Peng, W.; Dong, X.; Cho, B. G.; Mechref, Y. Characterization of isomeric glycan structures by LC-MS/MS. *ELECTROPHORESIS* **2017**, *38* (17), 2100-2114. DOI: https://doi.org/10.1002/elps.201700042.

(9) Choi, H. K.; Lee, D.; Singla, A.; Kwon, J. S.; Wu, H. J. The influence of heteromultivalency on lectin-glycan binding behavior. *Glycobiology* **2019**, *29* (5), 397-408. DOI: 10.1093/glycob/cwz010 From NLM Medline.

(10) Krishnan, P.; Singla, A.; Lee, C. A.; Weatherston, J. D.; Worstell, N. C.; Wu, H. J. Hetero-multivalent binding of cholera toxin subunit B with glycolipid mixtures. *Colloids Surf B Biointerfaces* **2017**, *160*, 281-288. DOI: 10.1016/j.colsurfb.2017.09.035.

(11) Worstell, N. C.; Krishnan, P.; Weatherston, J. D.; Wu, H. J. Binding Cooperativity Matters: A GM1-like Ganglioside-Cholera Toxin B Subunit Binding Study Using a Nanocube-Based Lipid Bilayer Array. *PLoS ONE* **2016**, *11* (4), e0153265. DOI: 10.1371/journal.pone.0153265.

(12) Worstell, N. C.; Singla, A.; Saenkham, P.; Galbadage, T.; Sule, P.; Lee, D.; Mohr, A.; Kwon, J. S.; Cirillo, J. D.; Wu, H. J. Hetero-Multivalency of Pseudomonas aeruginosa Lectin LecA Binding to Model Membranes. *Scietific Reports* **2018**, *8* (1), 8419. DOI: 10.1038/s41598-018-26643-7.

(13) Hirabayashi, J.; Yamada, M.; Kuno, A.; Tateno, H. Lectin microarrays: concept, principle and applications. *Chem Soc Rev* **2013**, *42* (10), 4443-4458, 10.1039/C3CS35419A. DOI: 10.1039/C3CS35419A.

(14) Dang, K.; Zhang, W.; Jiang, S.; Lin, X.; Qian, A. Application of Lectin Microarrays for Biomarker Discovery. *ChemistryOpen* **2020**, *9* (3), 285-300. DOI: https://doi.org/10.1002/open.201900326.

(15) Hu, Q.; Kuai, D.; Park, H.; Clark, H.; Balbuena, P. B.; Kwon, J. S.-I.; Wu, H.-J. Advancing Glycan Analysis: A New Platform Integrating SERS, Boronic Acids, and Machine Learning Algorithms. *Advanced Sensor Research* **2023**, *2* (12), 2300052. DOI: https://doi.org/10.1002/adsr.202300052.

(16) Wang, X.; Xia, N.; Liu, L. Boronic Acid-based approach for separation and immobilization of glycoproteins and its application in sensing. *Int J Mol Sci* **2013**, *14* (10), 20890-20912. DOI: 10.3390/ijms141020890 PubMed.

(17) Melchiorre, M.; Ferreri, C.; Tinti, A.; Chatgilialoglu, C.; Torreggiani, A. A Promising Raman Spectroscopy Technique for the Investigation of trans and cis Cholesteryl Ester Isomers in Biological Samples. *Appl Spectrosc* **2015**, *69* (5), 613-622. DOI: 10.1366/14-07706 From NLM Medline.

(18) Cho, Y. C.; Ahn, S. I. Fabricating a Raman spectrometer using an optical pickup unit and pulsed power. *Scientific Reports* **2020**, *10* (1), 11692. DOI: 10.1038/s41598-020-68650-7.

(19) Weatherston, J. D.; Seguban, R. K. O.; Hunt, D.; Wu, H.-J. Low-Cost and Simple Fabrication of Nanoplasmonic Paper for Coupled Chromatography Separation and Surface Enhanced Raman Detection. *ACS Sensors* **2018**, *3* (4), 852-857. DOI: 10.1021/acssensors.8b00098 (acccessed 2021-03-19T21:24:07).

(20) Weatherston, J. D.; Yuan, S.; Mashuga, C. V.; Wu, H.-J. Multi-functional SERS substrate: Collection, separation, and identification of airborne chemical powders on a single device. *Sensors and Actuators B: Chemical* **2019**, *297*, 126765. DOI: https://doi.org/10.1016/j.snb.2019.126765.

(21) Saunders, M. J.; Woods, R. J.; Yang, L. Simplifying the detection and monitoring of protein glycosylation during in vitro glycoengineering. *Scientific Reports* **2023**, *13* (1), 567. DOI: 10.1038/s41598-023-27634-z.

(22) Xu, M.; Yang, A.; Xia, J.; Jiang, J.; Liu, C.-F.; Ye, Z.; Ma, J.; Yang, S. Protein glycosylation in urine as a biomarker of diseases. *Translational Research* **2023**, *253*, 95-107. DOI: https://doi.org/10.1016/j.trsl.2022.08.001.

(23) Costa, J.; Hayes, C.; Lisacek, F. Protein glycosylation and glycoinformatics for novel biomarker discovery in neurodegenerative diseases. *Ageing Research Reviews* **2023**, *89*, 101991. DOI: https://doi.org/10.1016/j.arr.2023.101991.

(24) Ghosh, S. Chapter 1 - Sialic acid and biology of life: An introduction. In *Sialic Acids and Sialoglycoconjugates in the Biology of Life, Health and Disease*, Ghosh, S. Ed.; Academic Press, 2020; pp 1-61.

(25) Lewis, A. L.; Chen, X.; Schnaar, R. L.; Varki, A. Sialic acids and other nonulosonic acids. *Essentials of Glycobiology [Internet]. 4th edition* **2022**.

(26) Burzyńska, P.; Sobala, Ł. F.; Mikołajczyk, K.; Jodłowska, M.; Jaśkiewicz, E. Sialic acids as receptors for pathogens. *Biomolecules* **2021**, *11* (6), 831.

(27) Wasik, B. R.; Barnard, K. N.; Parrish, C. R. Effects of Sialic Acid Modifications on Virus Binding and Infection. *Trends in Microbiology* **2016**, *24* (12), 991-1001. DOI: https://doi.org/10.1016/j.tim.2016.07.005.

(28) Schauer, R.; Kamerling, J. P. Chapter One - Exploration of the Sialic Acid World. In *Advances in Carbohydrate Chemistry and Biochemistry*, Baker, D. C. Ed.; Vol. 75; Academic Press, 2018; pp 1-213.

(29) Folch, J.; Lees, M.; Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J Biol Chem* **1957**, *226* (1), 497-509. DOI: https://doi.org/10.1016/S0021-9258(18)64849-5 From NLM Medline.

(30) Nielsen, S. S. Phenol-Sulfuric Acid Method for Total Carbohydrates. In *Food Analysis Laboratory Manual*, Food Science Texts Series, Springer US, 2010; pp 47-53.

(31) Jombart, T.; Devillard, S.; Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **2010**, *11* (1), 94. DOI: 10.1186/1471-2156-11-94 (acccessed 2021-01-18T22:30:52). From NLM Medline.

(32) Hu, Q.; Sellers, C.; Kwon, J. S.-I.; Wu, H.-J. Integration of surface-enhanced Raman spectroscopy (SERS) and machine learning tools for coffee beverage classification. *Digital Chemical Engineering* **2022**, *3*, 100020. DOI: https://doi.org/10.1016/j.dche.2022.100020.

(33) Berrar, D. Cross-Validation. In *Encyclopedia of Bioinformatics and Computational Biology*, Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. Eds.; Academic Press, 2019; pp 542-545.

(34) Guan, B.; Zhang, Z.; Chai, Y.; Amantai, X.; Chen, X.; Cao, X.; Yue, X. N-glycosylation of milk proteins: A review spanning 2010–2022. *Trends in Food Science & Technology* **2022**, *128*, 1-21. DOI: https://doi.org/10.1016/j.tifs.2022.07.017.

(35) van Leeuwen, S. S.; te Poele, E. M.; Chatziioannou, A. C.; Benjamins, E.; Haandrikman, A.; Dijkhuizen, L. Goat Milk Oligosaccharides: Their Diversity, Quantity, and Functional Properties in Comparison to Human Milk Oligosaccharides. *Journal of Agricultural and Food Chemistry* **2020**, *68* (47), 13469-13485. DOI: 10.1021/acs.jafc.0c03766.

(36) de Sousa, Y. R. F.; da Silva Vasconcelos, M. A.; Costa, R. G.; de Azevedo Filho, C. A.; de Paiva, E. P.; Queiroga, R. d. C. R. d. E. Sialic acid content of goat milk during lactation. *Livestock Science* **2015**, *177*, 175-180. DOI: https://doi.org/10.1016/j.livsci.2015.04.005.

(37) Sheng, B.; Thesbjerg, M. N.; Glantz, M.; Paulsson, M.; Nielsen, S. r. D.; Poulsen, N. A.; Larsen, L. B. Phosphorylation and glycosylation isoforms of bovine κ-casein variant E in homozygous Swedish Red cow milk detected by liquid chromatography-electrospray ionization mass spectrometry. *Journal of Dairy Science* **2022**, *105* (3), 1959-1965. DOI: https://doi.org/10.3168/jds.2021-21172.

(38) Folch, J.; Lees, M.; Stanley, G. H. S. A SIMPLE METHOD FOR THE ISOLATION AND PURIFICATION OF TOTAL LIPIDES FROM ANIMAL TISSUES. *Journal of Biological Chemistry* **1957**, *226* (1), 497-509. DOI: https://doi.org/10.1016/S0021-9258(18)64849-5.

(39) Fischler, D. A.; Orlando, R. N-linked Glycan Release Efficiency: A Quantitative Comparison between NaOCl and PNGase F Release Protocols. *J Biomol Tech* **2019**, *30* (4), 58-63. DOI: 10.7171/jbt.19-3004-001 From NLM.

(40) Kinoshita, M.; Yamada, K. Recent advances and trends in sample preparation and chemical modification for glycan analysis. *Journal of Pharmaceutical and Biomedical Analysis* **2022**, *207*, 114424. DOI: https://doi.org/10.1016/j.jpba.2021.114424.

(41) Tiwold, E. K.; Gyorgypal, A.; Chundawat, S. P. S. Recent Advances in Biologic Therapeutic N-Glycan Preparation Techniques and Analytical Methods for Facilitating Biomanufacturing Automation. *J Pharm Sci-Us* **2023**, *112* (6), 1485-1491. DOI: https://doi.org/10.1016/j.xphs.2023.01.012.

(42) Wilkinson, H.; Saldova, R. Current Methods for the Characterization of O-Glycans. *Journal of Proteome Research* **2020**, *19* (10), 3890-3905. DOI: 10.1021/acs.jproteome.0c00435.

(43) Shipman, J. T.; Nguyen, H. T.; Desaire, H. So You Discovered a Potential Glycan-Based Biomarker; Now What? We Developed a High-Throughput Method for Quantitative Clinical Glycan Biomarker Validation. *ACS Omega* **2020**, *5* (12), 6270-6276. DOI: 10.1021/acsomega.9b03334.

(44) Lin, Y.-H.; Franc, V.; Heck, A. J. R. Similar Albeit Not the Same: In-Depth Analysis of Proteoforms of Human Serum, Bovine Serum, and Recombinant Human Fetuin. *Journal of Proteome Research* **2018**, *17* (8), 2861-2869. DOI: 10.1021/acs.jproteome.8b00318.

(45) Černocká, H.; Římánková, L.; Ostatná, V. Fetuin and asialofetuin at charged surfaces: Influence of sialic acid presence. *Journal of Electroanalytical Chemistry* **2021**, *902*, 115801. DOI: https://doi.org/10.1016/j.jelechem.2021.115801.

(46) Schultz, M. J.; Swindall, A. F.; Bellis, S. L. Regulation of the metastatic cell phenotype by sialylated glycans. *Cancer and Metastasis Reviews* **2012**, *31* (3), 501-518. DOI: 10.1007/s10555-012-9359-7.

(47) Anthony, R. M.; Nimmerjahn, F.; Ashline, D. J.; Reinhold, V. N.; Paulson, J. C.; Ravetch, J. V. Recapitulation of IVIG Anti-Inflammatory Activity with a Recombinant IgG Fc. *Science* **2008**, *320* (5874), 373-376. DOI: doi:10.1126/science.1154315.

(48) Blundell, P. A.; Lu, D.; Dell, A.; Haslam, S.; Pleass, R. J. Choice of Host Cell Line Is Essential for the Functional Glycosylation of the Fc Region of Human IgG1 Inhibitors of Influenza B Viruses. *The Journal of Immunology* **2020**, *204* (4), 1022-1034. DOI: 10.4049/jimmunol.1901145 (acccessed 1/29/2024).

(49) Manz, C.; Mancera-Arteu, M.; Zappe, A.; Hanozin, E.; Polewski, L.; Giménez, E.; Sanz-Nebot, V.; Pagel, K. Determination of Sialic Acid Isomers from Released N-Glycans Using Ion Mobility Spectrometry. *Analytical Chemistry* **2022**, *94* (39), 13323-13331. DOI: 10.1021/acs.analchem.2c00783.

(50) Suzuki, N.; Abe, T.; Natsuka, S. Quantitative LC-MS and MS/MS analysis of sialylated glycans modified by linkage-specific alkylamidation. *Analytical Biochemistry* **2019**, *567*, 117-127. DOI: https://doi.org/10.1016/j.ab.2018.11.014.

(51) Tommasone, S.; Allabush, F.; Tagger, Y. K.; Norman, J.; Köpf, M.; Tucker, J. H. R.; Mendes, P. M. The challenges of glycan recognition with natural and artificial receptors. *Chemical Society Reviews* **2019**, *48* (22), 5488-5505, 10.1039/C8CS00768C. DOI: 10.1039/C8CS00768C.

(52) Fang, G.; Wang, H.; Bian, Z.; Sun, J.; Liu, A.; Fang, H.; Liu, B.; Yao, Q.; Wu, Z. Recent development of boronic acid-based fluorescent sensors. *RSC Advances* **2018**, *8* (51), 29400-29427, 10.1039/C8RA04503H. DOI: 10.1039/C8RA04503H.

(53) Nyquist, R. Chapter 9 - Benzene and Its Derivatives. In *Interpreting Infrared, Raman, and Nuclear Magnetic Resonance Spectra*, Nyquist, R. A. Ed.; Academic Press, 2001; pp 351-423.

(54) Arimitsu, K.; Iwasaki, H.; Kimura, H.; Yasui, H. Strong Binding Affinity of d-Allulose and Allulosides to Boronic Acids and the Structural Characterization of Their Sugar-boronate Complexes. *Chemistry Letters* **2021**, *50* (8), 1470-1474. DOI: 10.1246/cl.210181 (acccessed 1/11/2024).

(55) Tommasone, S.; Tagger, Y. K.; Mendes, P. M. Targeting Oligosaccharides and Glycoconjugates Using Superselective Binding Scaffolds. *Adv Funct Mater* **2020**, *30* (31), 2002298. DOI: https://doi.org/10.1002/adfm.202002298.