Synthetic heparan sulfate standards and machine learning facilitate the development of solid-state nanopore analysis

Ke Xia[a,b], James T. Hagan[c], Li Fu[b], Brian S. Sheetz[c], Somdatta Bhattacharya[d], Fuming Zhang[d], Jason R. Dwyer[c,1], and Robert J. Linhardt[a,b,d,1]

[a]Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, NY 12180-3590; [b]Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY 12180-3590; [c]Department of Chemistry, University of Rhode Island, Kingston, RI 02881; and [d]Howard P. Isermann Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180

Edited by Robert Langer, Massachusetts Institute of Technology, Cambridge, MA, and approved January 25, 2021 (received for review November 23, 2020)

The application of solid-state (SS) nanopore devices to single-molecule nucleic acid sequencing has been challenging. Thus, the early successes in applying SS nanopore devices to the more difficult class of biopolymer, glycosaminoglycans (GAGs), have been surprising, motivating us to examine the potential use of an SS nanopore to analyze synthetic heparan sulfate GAG chains of controlled composition and sequence prepared through a promising, recently developed chemoenzymatic route. A minimal representation of the nanopore data, using only signal magnitude and duration, revealed, by eye and image recognition algorithms, clear differences between the signals generated by four synthetic GAGs. By subsequent machine learning, it was possible to determine disaccharide and even monosaccharide composition of these four synthetic GAGs using as few as 500 events, corresponding to a zeptomole of sample. These data suggest that ultrasensitive GAG analysis may be possible using SS nanopore detection and well-characterized molecular training sets.

solid-state nanopore | polysaccharide | glycosaminoglycan | sequencing | single-molecule analysis

Glycosaminoglycans (GAGs) are linear anionic polysaccharides found on cell surfaces and in the extracellular matrix in all animals. GAGs comprise an important class of biopolymers that are ubiquitous in nature and exhibit a number of critical functional roles including biological recognition and signaling (1–3). Such processes play critical roles in physiology, such as in development and wound healing, and pathophysiology, such as cancer and infectious disease. Sulfated GAGs result from template-independent synthesis in the Golgi of animal cells (4, 5) and are polydisperse, heteropolysaccharides comprising variable disaccharide repeating units that are classified by these repeating units. Like nucleic acids, sulfated GAGs are made up of repeating units that comprise a linear sequence (Fig. 1). Unlike the nucleic acids, GAGs have far more complicated structures and number of possible sequences and they present severe challenges to both synthesis and characterization. Thus, we undertook to chemoenzymatically synthesize defined GAGs and characterize these using solid-state nanopore analysis.

Despite their structural complexities, sulfated GAGs often contain well-defined domain structures that are responsible for their diverse biological functions, yet even this level of structural complexity poses a significant general challenge to structural analysis and sequencing. The simple, short-chain, chondroitin sulfate GAG component of bikunin has been sequenced using liquid chromatography–tandem mass spectrometry (LC-MS/MS) (6). While LC-MS/MS is capable of sequencing such simple, short-chain GAGs, it is not yet able to distinguish all of the many isobaric isomers of the variably sulfated saccharide residues and uronic acid epimers commonly encountered in more structurally complex GAGs, such as heparan sulfate (HS) (7). NMR has been applied to determine GAG structures but often requires milligram amounts of samples. HS/heparin is made up of →4)-β-D-glucuronic acid (GlcA) [or α-L-iduronic acid (IdoA)] (1→4)-α-D-glucosamine (GlcN) [1→ repeating units with 2-*O*-sulfo (S) groups on selected uronic acid residues and 3- and/or 6-*O*-S and *N*-S or *N*-acetyl (Ac) group substitutions on the glucosamine residues] (Fig. 1). GAG structural analysis presents challenges beyond their chemical complexity. There are no amplification methods to detect small numbers of GAG chains, whereas nucleic acid analysis can rely on PCR. Similarly, there are few GAG-specific antibodies or aptamers (8), and no natural GAG chromophores or fluorophores (9), in contrast to the many used for protein sensing. Ultrasensitive (zeptomole) detection methods of modified GAGs, based on fluorescence resonance energy transfer (FRET) (10), DNA bar coding (11), and dye-based nanosensors (12) have been demonstrated, but their application to sequencing is particularly challenging because of the high level of structural complexity of sulfated GAGs.

Nanopore single-molecule detection is now routinely applied to DNA (13, 14) and RNA (15–17) biopolymers, and is increasingly applied to protein characterization (18–22). In brief, a nanopore is a nanofluidic channel ~10 nm long and <100 nm in diameter, serving as the sole fluid connection between two reservoirs of electrolyte separated by an otherwise impermeable

Significance

Glycans, proteins, and nucleic acids are the three most important biomolecule classes. Nanopore DNA sequencing is moving toward maturity as a competitive alternative to commonly used technologies. A few nanopore studies have made successful forays into the realm of glycan analysis, motivating the work described in our paper. However, these previous studies have been limited by the absence of highly controlled/well-regulated standards, which now have been chemoenzymatically synthesized by us. Without using amplification, with pioneering solid-state nanopore platform and machine-learning–based signal, we clearly differentiate four synthetic heparan sulfates and, astonishingly, determine disaccharide and even monosaccharide composition on as little as 1 zeptomole of sample.

Author contributions: J.R.D. and R.J.L. designed research; K.X., J.T.H., L.F., B.S.S., and S.B. performed research; L.F. contributed new reagents/analytic tools; K.X., J.T.H., B.S.S., F.Z., J.R.D., and R.J.L. analyzed data; and K.X., J.R.D., and R.J.L. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: jason_dwyer@uri.edu or linhar@rpi.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2022806118/-/DCSupplemental.

Published March 9, 2021.

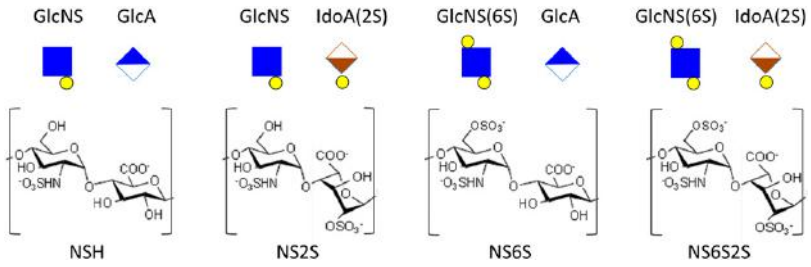


Fig. 1. Structures of four synthetic GAG samples. Polysaccharide NSH is made up of *N*-sulfoglucosamine (GlcNS) and glucuronic acid (GlcA),NS2S is made up with GlcNS and 2-*O*-sulfo-iduronic acid (IdoA2S), NS6S is made up with 6-*O*-sulfo-*N*-sulfoheparosan (GlcNS6S) and GlcA, and NS6S2S is made up with GlcNS6S and IdoA2S.

membrane (Fig. 2*A*). On applying a voltage across this nanopore, the passage of supporting electrolyte ions results in a "baseline," or open-pore current, i_0. The passage of a biopolymer analyte through this nanopore disrupts the flow of supporting electrolyte ions, often as a current blockage. This temporary reduction in ionic current is called an "event," and its magnitude (mean blockage ratio over the dwell time, $\langle f_b \rangle = \langle i \rangle_{Td}/\langle i_0 \rangle$) and its temporal features [dwell time (Td)] (Fig. 2 *B* and *C*) depend on the size and shape of the nanopore, the biopolymer analyte, and the applied voltage and interfacial charge distributions. Indeed, the passage of DNA through engineered protein nanopore devices produces current blockages that can be applied in sequencing, and the widespread use of these commercial protein nanopore DNA sequencing devices is increasing (23, 24). Despite this success with protein nanopores, the potential benefits of (abiotic) solid-state (SS) nanopores have continued to drive development efforts. Such a transition to the freely size-tunable SS platform (25, 26), however, is vital for the application of nanopores to the characterization of branched glycans (27). Yet the use of SS nanopores in even the better-established DNA sensing regime remains challenging. The application of nanopore sensing to glycans, while promising, remains profoundly exploratory using nanopores of any kind. The transition to the SS nanopores is accompanied by significant changes in pore geometry, chemistry, characteristics, and potential analyte–pore interactions and sensing modalities, so that there is a critical need for studies in the realm of nanopore glycomics (27, 28). For example, outcomes of early nanopore studies on a structurally simple unsulfated GAG, hyaluronan (HA, →4)- β -GlcA (1 → 3)- β -GlcNAc (1→), while providing some information on HA size does not provide definitive structural information (29, 30). SS nanopore analysis of two sulfated GAGs, heparin and a heparin contaminant, oversulfated chondroitin sulfate, using a silicon nitride SS nanopore was able to qualitatively identify these GAGs by either the magnitude or duration of characteristic current blockages (28). SS nanopore data on GAGs, analyzed using a machine-learning (ML) algorithm (i.e., a support vector machine [SVM]), distinguished heparin and chondroitin sulfate oligosaccharides and unfractionated heparin and low molecular weight heparin with >90% accuracy (31).

Nanopore studies on GAGs, and glycans more broadly, have been severely limited by the lack of a library of structurally defined standards. The uniformity of sulfated GAGs prepared from animal sources is difficult to control and exhibits significant sequence heterogeneity and polydispersity (32). HS is particularly problematic as even for a small HS hexasaccharide, composed of an IdoA/GlcA:GlcNS/GlcNAc sequence with 12 available sites for random sulfation, there are 32,768 possible sequences. Recently, chemoenzymatic synthesis has made inroads in the preparation of high-purity sulfated HS GAGs from heparosan (→4)- β -GlcA (1→4)- β -GlcNAc (1→) (33). HS GAGs of approximately the same chain length and polydispersity and having a single repeating disaccharide unit (*SI Appendix*, Table S1) including, NSH (→4)- β -GlcA (1→4)- β -GlcNS (1→), NS2S (→4)- α α -IdoA2S (1 → 4)- β -GlcNS (1→), NS6S (→4)- β -GlcA (1→4)- β -GlcNS6S(1→)), NS6S2S (→4)- α -IdoA2S (1→4)- β -GlcNS6S (1→) have been prepared (see *Materials and Methods* and ref. 34) (Fig. 1). Here we use our recently developed synthetic technique, which has proven difficult to benchmark, in conjunction with a nanopore technique, which has only just begun to be applied to glycomics and has been severely challenged by the lack of available high-quality samples, to develop a fully integrated approach for the nanopore analysis of complex carbohydrates.

Results and Discussion

We prepared a carefully designed group of chemoenzymatically synthesized HS GAG chains and have synthesized these GAGs following a similar strategy as described in our previous publication (34). Biosynthetically enriched heparosan precursor was obtained from microbial culture. Subsequent steps were nearly complete chemical de-*N*-acetylation using aqueous NaOH and *N*-sulfation with $(CH_3)_3N{\cdot}SO_3$, enzymatic epimerization, and sulfation using recombinant heparin biosynthetic enzymes (*SI Appendix*, Fig. S1). Those model GAGs have uniform length (~40 disaccharide units) (*SI Appendix*, Table S1), compositions, and sequences (Fig. 1). Each of the four synthetic GAG samples comprised a different disaccharide unit, having from one to three sulfate groups by using a combination of four monosaccharide units. As confirmed by NMR (*SI Appendix*, Fig. S2) and disaccharide compositional analysis (*SI Appendix*, *Methods*), all synthetic steps are highly efficient; NSH was first synthesized from a nearly completely de-*N*-acetylated heparosan, resulting in 97% *N*-sulfation and 3% residual *N*-acetyl. NS2S was then synthesized from this NSH, with 90% 2-O-sulfation, and NS6S was similarly synthesized from NSH, but with 86% 6-O-sulfation. Finally, NS6S2S was synthesized as an 86% 6-O-sulfation product from NS2S. Considering the challenge in synthesis, this is the by far the purest polysaccharide GAG library reported for nanopore study.

The nanopore experimental configuration is illustrated in Fig. 2*A*. Detectable events with good signal-to-noise ratio, appreciable event frequencies, and the lowest probability of occurrence of signal instabilities generally ascribed to analyte sticking to the pore surface were achieved in 4 M KCl electrolyte at pH 4.1. Signal was not detected in 100 mM KCl electrolyte (NS2S; ϕ4 nm; pH 9; ±100, 400 mV), nor was signal detected in 1 M KCl (NHS; ϕ11 nm; pH 4.1, 9; ±150, 200 mV). While events could be detected at pH 7 and 9 in 4 M KCl electrolyte, sticking was irreversible at pH 7 (NHS) and reversible but detrimentally frequent at pH 9 (notably for NS6S). At pH 7 and pH 9, the nanopore surface is net negatively charged: Electrostatic repulsion of the anionic GAGs would, thus, disfavor analyte entry to the pore, and electroosmosis would oppose electrophoresis (27, 28).

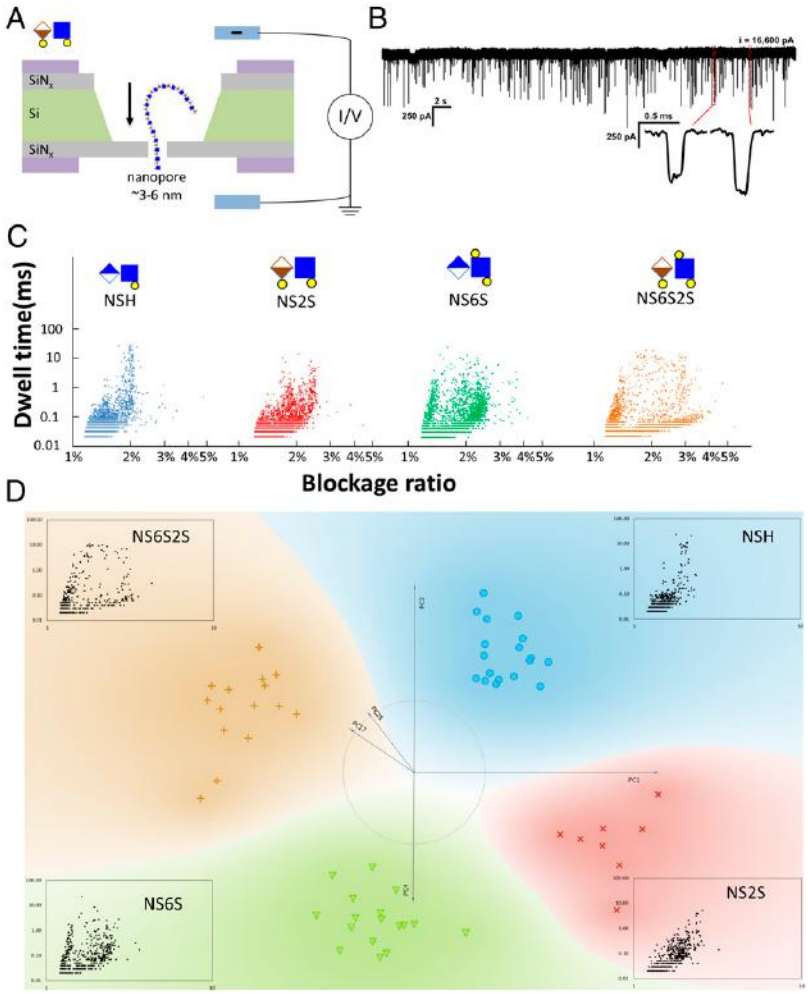


Fig. 2. Nanopore characteristics of four samples. (*A*) Schematic of the nanopore configuration. Anionic GAGs driven by electrophoresis to and through the pore with a negative applied voltage would be detected if they perturbed the open-pore current. (*B*) A representative current trace and events from polysaccharide NS6S2S test using an ~6-nm-diameter nanopore. Measurements were collected using a −150-mV applied-voltage (details in *Results and Discussion*, and *Materials and Methods*) (*C*) Scatter plots of dwell time vs. current blockage ratio for four polysaccharides. To remove the bias of event numbers in human image recognition, all plots contain only the first 2,475 events. (*D*) PCA visualization of the embedded images from the four unique GAGs. The blue circles and region represent NSH, the red X and region represents NS2S, the green triangle and region represents NS6S, and the brown cross and region represents NS6S2S. The algorithm clusters signals from each GAG based on scatter plot images. Each insert shows one 500-events image from each sample class. All 500-events images are in *SI Appendix*, Fig. S12.

At pH 4.1, the amphoteric SiN_x pore is positive and electroosmotic and electrophoretic driving forces both contribute to analyte translocation through the pore. The solution was not acidified appreciably beyond the SiN_x isoelectric point of pI~4.3 to prevent chemical degradation of either the analyte or the nanopore surface.

All four GAGs could be detected at pH 4.1 in 4 M KCl using an ~6-nm-diameter silicon nitride SS nanopore (Fig. 2*A*). A constant voltage of −150 mV (the electrophoretic polarity) was applied for ~10 min except when an occasional voltage pulse was needed to reverse transient clogging (28). Unstable sections of the nanopore signal owing to such events were removed from further consideration. At least 2,475 unique events were recorded for each sample (*SI Appendix*, Table S2). Representative sections of current traces of all GAGs are plotted (Fig. 2*B* and *SI Appendix*, Fig. S3).

An automated extraction method based on current blockage levels, $f_b = i/\langle i_0 \rangle$—the ratio of the instantaneous current to the mean open-pore current in each consecutive analysis window—was used to separate the events from the open-pore current, as detailed in *Materials and Methods*. In brief, continuous ranges of time for which $f_b \leq f_{threshold}$ were identified as single events, and were reported as $(\langle f_b \rangle_{Td}, Td)$. Histograms of these data are shown in *SI Appendix*, Fig. S6. Critically, a single $f_{threshold} = 0.988$ was used for all samples. Scatter plots showing the durations (Td) and blockage ratios of the first 2,475 events from each sample, NSH, NS2S, NS6S, and NS6S2S, were created (Fig. 2*C* and *SI Appendix*, Fig. S4). Sophisticated data analysis methods, including ML, play an important enabling role in nanopore analysis. It is well-established for DNA sequencing (13, 35, 36), and was recently first applied to the analysis of clinically relevant samples of heparin (31). That study used 16 different signal parameters, including the maximum event amplitude, average event amplitude, duration, blockade (blockage) ratio, and the average magnitude of the cepstrum spectra down-sampled into

51 equal windows of the event. Without overlooking the potential benefits of exhaustively mining the information content of the nanopore signal and its possible transformations, we were interested in exploring the sensitivity of the canonical minimal representation of the nanopore signal, the event current and duration, to GAG identity. Moreover, we aimed to challenge the ability of ML to extract information using this highly restricted set of nanopore signal parameters. In keeping with this ethos, we undertook two data analysis approaches: to see, first, if the analytes could be distinguished at the level of fingerprint (27, 28), and then to determine if sequence-specific information, to the level of monosaccharide, could be extracted.

The scatter plots have the potential to serve as analyte fingerprints, and the samples were easily distinguishable by scatter plot comparisons using only the human eye (Fig. 2*C*). We decided to treat such scatter plots as pure images and use image analysis approaches to make this visual fingerprint examination more exacting and less affected by human subjectivity. We took monochrome deidentified scatter plots of as few as 500 events for each image, with common figure size and positions, and used Google's deep neural network for image embedding, converting the image into feature vectors. Principal component analysis (PCA) was then used to compress these feature vectors and to classify them (37). The resulting clusters showed clear differentiation by sample type (Fig. 2*D*). This graphical analysis of 58 fingerprint images of 500 events each of the 4 synthetic GAG samples—varying by disaccharide composition—consistently identified the correct GAG. Use of the entropy-MDL method to create a heat map using all events provided further quantitative differentiation between the signal characteristics of each sample (*SI Appendix*, Figs. S4–S6). A heat map of the four synthetic GAG samples for these nine states shows the relatedness of each sample (*SI Appendix*, Fig. S7). For example, the monosulfated NSH is primarily associated with weak blockage (e.g., refs. 1–3) and trisulfated NS6S2S is primarily associated with strong blockage (e.g., refs. 7–9). Moreover, the two disulfated GAG structures, NS2S and NS6S, show intermediate but distinctly different patterns in their heat maps.

After the successful use of image analysis algorithms to differentiate between the four GAG samples by their fingerprints, we wished to delve more deeply into the less apparent information content of the nanopore signals using ML. We continued at the minimal level of Td and blockage ratio, only, but classified the 9,250 events into 9 total states: by Td as slow, intermediate, or fast events, and by blockage ratio as weak, moderate, or strong events. The classification boundaries were established using an equal division frequency along each axis (Fig. 3*A* and *SI Appendix*, Fig. S8). For ML the nanopore current traces for each ~20-min sample read were each divided into eight short reads having an equal number of events (Fig. 3*B*). The 32 short reads (four synthetic GAG samples each with eight short reads) were normalized into the same nine states (Fig. 3*A* and *SI Appendix*, Table S3 and Fig. S9). A robust and low sampling bias 100-repeats-stratified-random-sampling, was applied with a 75% training set and a 25% external validation set (EVS) split (Fig. 4*A*). The eight short reads in EVS were randomly split from the 32 short reads. The remaining 24 short reads were used for development of seven widely used models that were applied to predict the eight short reads in the EVS. This entire process was repeated 100 times, and the accumulated results were used for evaluation. The seven models were compared, with the most accurate model SVM giving an accuracy of 91.2% (Fig. 4*B*). A confusion table for the SVM model successfully predicted most short reads in the EVS, as shown in the blue highlighted diagonal in Fig. 4*C*. The receiver-operating characteristic (ROC) curves for each of the four synthetic GAG samples for the top three models based on accuracy are shown (Fig. 4*D*) with SVM showing an average area under the curve of 98.3% (100% representing all correct). To rule out the possibility of random correlations in data, a randomly labeled dataset was generated by randomly class-labeling the 32 short reads. The same ML strategy applied to randomly labeled dataset resulted in a significantly lower predictive accuracy level of 25–45% and an area under the curve (AUC) ~40 and 60% (*SI Appendix*, Fig. S11*C*). These results ruled out the possibility of random correlations in data. Although the prediction from single short reads was already very promising, if we combine the saccharide calls from all eight short reads of the same sample, the consensus accuracy is close to 100%. The validated predictive ability was further examined by the "stratified k-fold cross-validation" (SKCV) and "leave one out cross-validation" (LOOCV) methods (38) (*SI Appendix*, Fig. S11 *A* and *B*), which both demonstrated high accuracy (93.7–96.9%). Overall, the reliability of our ML has been well validated, thus demonstrating accuracy and robustness of our approach.

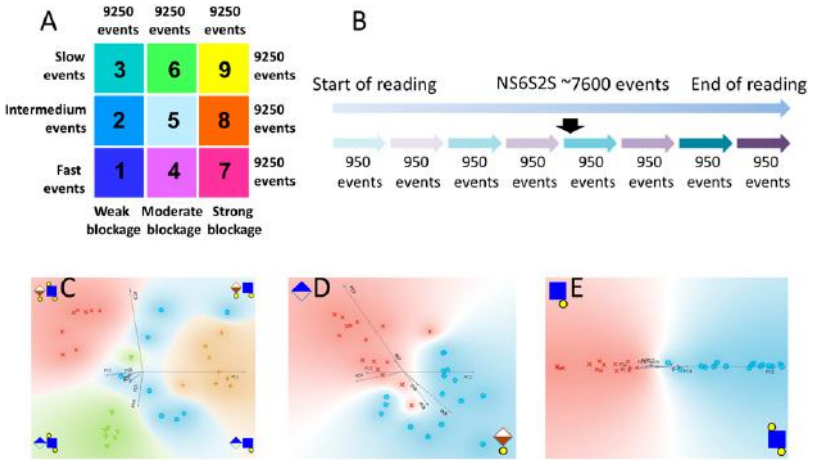
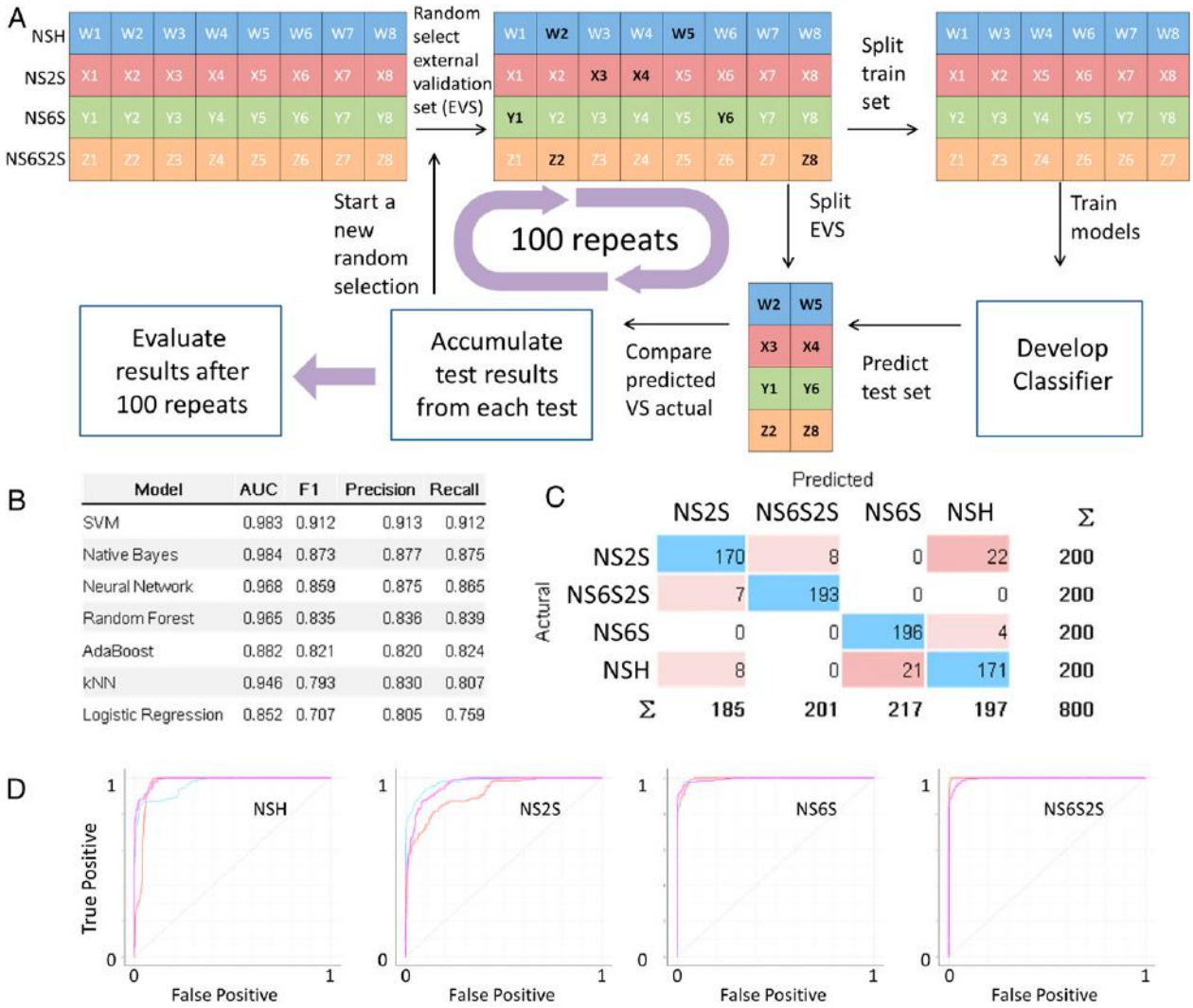


Fig. 3. Data process and normalization. (*A*) All ~27,750 events from four samples (the entire dataset) were binned into nine states. (*B*) The NS6S2S long read, for example, was divided into eight short reads by equal frequency. (*C*) PCA clustering of the 32 short reads from four GAG classes of samples. Common colored points and regions represent each GAG class, showing how the algorithm clusters the four GAG classes based on states of short reads. (*D*) PCA clustering of the 32 short reads, showing how the algorithm clusters the GlcA vs. IdoA(2S) based on states of short reads. (*E*) PCA clustering of the 32 short reads, showing how the algorithm clusters the GlcNS vs. GlcNS(6S) based on states of short reads. Algorithm clusters of the GlcNS vs. GlcNS(6S).



Model	AUC	F1	Precision	Recall
SVM	0.983	0.912	0.913	0.912
Native Bayes	0.984	0.873	0.877	0.875
Neural Network	0.968	0.859	0.875	0.865
Random Forest	0.965	0.835	0.836	0.839
AdaBoost	0.882	0.821	0.820	0.824
kNN	0.946	0.793	0.830	0.807
Logistic Regression	0.852	0.707	0.805	0.759

Actural \ Predicted	NS2S	NS6S2S	NS6S	NSH	Σ
NS2S	170	8	0	22	200
NS6S2S	7	193	0	0	200
NS6S	0	0	196	4	200
NSH	8	0	21	171	200
Σ	185	201	217	197	800

Fig. 4. Identify test samples by ML. (*A*) Stratified random sampling with 100 repeats of train and test (75 and 25% random split) was used for a robust evaluation of identification ability and avoids sampling bias. (*B*) Seven different developed models were used and evaluated using the AUC, F1 measure (the harmonic mean of precision and recall), precision [TP/(TP+FP)], and recall [TP/(TP+FN)]. (*C*) The confusion table of the best model's (SVM) classification of the test runs. The model returned successful predictions for most test runs. (*D*) ROC curve for each sample class. ROC of top three models (magenta-colored SVM, blue-colored native Bayes, orange-colored neural network). The average AUC for SVM was 98.3%, providing both excellent specificity and selectivity.

The four GAGs comprised unique dimers drawn from pairings of two from the four options (Fig. 1). Analysis of event depth and duration, alone, clearly differentiated between GAG signal characteristics in pure and mixed sample sets, thus permitting the characterization of samples at the level of disaccharide composition. This successful determination prompted us to ask whether the signal analysis would support further refinement to recognize monosaccharide composition. That is, having successfully differentiated the nanopore signals on the basis of GAG disaccharide composition, whether signal analysis could reveal, for example, that NSH and NS2S shared a common monosaccharide (GlcNS).

We next asked the question whether these data could differentiate between the monosaccharide making up each of the four GAG samples. PCA analysis of these data was performed (Fig. 3 *C–E* and *SI Appendix*, Table S3). Surprisingly, not only could the disaccharide units of each of the four synthetic GAGs be resolved but also the samples could be classified based on their monosaccharide units.

Summary

The combined use of an SS nanopore device, composition-controlled, high-purity synthetic GAG samples, and ML has allowed for the clear classification of four GAG molecules based on patterns of events in only a minimal representation framework. Differentiation of signals on the basis of disaccharide composition was demonstrated in measurements of pure experimental samples. Moreover, further examination of these data shows selectivity of signal for individual monosaccharide units, thus foreshadowing the potential use of this tool for single-unit identification. Especially if further development could slow down glycans and be able to read more information from each event, a single-molecule reading would be achieved. In that scenario, this approach could be extent to sequence variation in disaccharides along a molecule.

Sulfation differences would be expected to alter event frequency. However, considering the local concentration of analytes near the nanopore would also contribute to the event frequency,

we intentionally avoid use the event frequency in the classification. In future development, if local concentration of analytes could be well controlled, events frequency difference could give another dimension of data contributing to the classification of samples.

Despite synthetic challenges, our samples are the purest GAG polysaccharides yet applied in nanopore studies. The 86% purity of NS6S and NS6S2S indicates these samples contain only minor components mixture associated with their less-sulfated precursors. However, the accurate classification of NS6S and NS6S2S suggests our method is also suitable for such mixtures and does not require standards of 100% purity.

The nonuniform character of the scatter plots for a single type of molecule is surprising. The same molecule might give different types of single events due to conformation differences or orientation as it passes through the nanopore. For example, translocation starting from the nonreducing end would undoubtedly give a different type of signal than starting from the reducing end. Further development of our nanopore sequencing method might make it possible to slow down glycan translocation providing more details of a single event, such as sublevels of current blockage and translocation duration. This could in turn provide a more refined level of sequencing information. One benefit of our method is that it relies on the entire scatter plot distribution, instead of using the details of a single event, making it more robust in handling a sample having many different conformations.

Overall our method is ultrasensitive requiring 500 events corresponding to 1 zeptomole and requires no modification of the GAG structure while providing structural information in mono- and disaccharide units.

Materials and Methods

Nanopore Formation and Characterization. Nanopores were formed by controlled dielectric breakdown (26) as described previously (28). Briefly, four 8-V direct current (DC) potentials were applied across the SiNx membrane in pH~7, 1 M KCl electrolyte. Once formed, the pore's (Ohmic) conductance, G, measured from −200–200 mV was used to infer a nanopore diameter from a conductance model accounting for bulk, surface, and access resistance terms for a cylindrical nanopore geometry (28).

$$G = \left(\frac{1}{G_{bulk} + G_{surface}} + \frac{1}{G_{access}} \right)^{-1}.$$

Nanopores used for measurements produced stable open-pore (analyte-free) currents in the electrolyte solutions used.

Nanopore Data Collection. All nanopore measurements were performed using an Axopatch 200B amplifier (Axon Instruments) in voltage clamp mode, configured as detailed in earlier work (28). Current-versus-time measurements were collected over 15–20 min at 100-kHz acquisition rates with Bessel filtering using a built-in low-pass filter set to 10 kHz. To measure the conductance, the acquisition rate was set to 10 kHz and low-pass filter set to 1 kHz (more details in *SI Appendix*).

NSH to NS6S2S Sample Preparation Major Pathway. We have synthesized GAGs following similar strategy as our previous publication (34). Biosynthetically enriched heparosan precursor is obtained from microbial culture. Subsequent steps were chemical de-*N*-acetylation using aqueous NaOH and *N*-sulfonated with $(CH_3)_3N{\cdot}SO_3$, enzymatic epimerization and sulfation with recombinant heparin biosynthetic enzymes (*SI Appendix*, Fig. S1).

Briefly, NSH samples were treated with immobilized enzymes, 2-OST-1 and C5-Epi are used to generate NS2S while NSH and NS2S were treated with immobilized enzymes 6-OST-1 and 6-OST-3 to generate NS6S and NS6S2S. The detailed reaction conditions are as follows: substrate concentration of 1 mg/mL, each enzyme concentration of 0.5 mg/mL for 50% slurry, 3′-phosphoadenosine-5′-phosphosulfate (PAPS) cofactor concentration of 3 mM, all reactions were incubated at 37 °C for 40 h in 50 mM 2-(*N*-morpholino) ethanesulfonic acid (MES) buffer (pH 7.2). After the reactions were complete, the mixtures were filtered to remove enzyme resin, dialyzed using 3K-Da molecular weight cutoff centrifugal membrane units with distilled water to remove PAPS, MES salt, and other small-molecule impurities, and the retentates were lyophilized for further synthesis and analysis. The samples were free of proteins and after dissolving each in nanopore buffer were filtered through a 0.2-μm filter (more detail in *SI Appendix*).

Image Recognition. Every 500 events from each sample were plotted as a two-dimensional (2D) scatter plot of event characteristics Td against blockage ratio. There were 18 images for NSH, 17 images for NS6S, and 15 images for NS6S2S. For NS2S, eight images were generated including four images from interpolation (events 251–750, 751–1250, 1251–1750,1751–2250). To minimize possible sources of confounding bias, all plot formatting and layout parameters were identical (*SI Appendix*, Fig. S12).

Images are embedded by Orange image analytics with InceptionV3, which is Google's deep neural network for image recognition, which consists of 48 layers (37). It is trained on the ImageNet data set and performs with 5.6% top-5 error (39, 40) at the ImageNet dataset. We used the pretrained networks implementation from the Tensorflow's models repository. PCA was used to compress the feature vectors output by the deep neural network to evaluate data clustering and to permit straightforward visual assessment of the analysis (details of PCA in *SI Appendix*).

ML. ML-based classifications were performed using Orange data-mining software (version 3.23) (37, 41). Seven ML algorithms were used as follows: 1) adaptive boosting; 2) k-nearest neighbors; 3) naive Bayes; 4) neural network; 5) random forest; 6) logistic regression; and 7) support vector machine (SVM). For the adaptive boosting, the base estimator and classification algorithm were tree and SAMMER, respectively, and the regression loss function was linear. For the k-nearest neighbors, the number of neighbors was five, with a Euclidean metric and uniform weight. For the naive Bayes, the algorithm was typical. For the neural network, the number of hidden layers, activation algorithm, solver, regularization parameter (alpha), and the maximum number of iterations were 200, logistic, Adam, 0.0001, and 200, respectively. For random forest, the number of trees and the split limit were 10 and 5, respectively. For logistic regression, the regularization type was Ridge(L2) and strength was C = 1. For the SVM, the cost, kernel, numerical tolerance, and iteration limit were 1, RBF, 0.001, and 100, respectively.

Following model development, the performance of ML models on unseen test datasets was evaluated and compared using the AUC. In addition to AUC, the F1 measure (the harmonic mean of precision and recall), precision, and recall were calculated as well.

We performed a rigorous validation procedure to establish the robustness of the sequence-specific analysis and whether it might be used to determine homogeneous samples of unknown composition and sequence. The 32 normalized short reads were split into a training set and an EVS. ML techniques were applied to the training set to construct classifiers, while the EVS was reserved unseen to ensure reliability (41). These classifiers were used to predict the classification of the unseen short reads in EVS. The predictive ability was examined by the SKCV and LOOCV methods (38) (*SI Appendix*, Fig. S11 *A* and *B*), which both demonstrated high accuracy (93.7–96.9%).

Data Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. We are thankful for valuable suggestions on ML from Prof. Pingkun Yan and his student Hanqing Chao. Funding was provided in the form of grants from the NIH (CA231074, DK111958, and HL125371 to R.J.L.) and the NSF (CHE 1808344 to J.R.D.).

1. A. Varki *et al*., *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, ed. 2, 2009).
2. P. H. Seeberger, Chemical glycobiology: Why now? *Nat. Chem. Biol.* **5**, 368–372 (2009).
3. M. Dalziel, M. Crispin, C. N. Scanlan, N. Zitzmann, R. A. Dwek, Emerging principles for the therapeutic exploitation of glycosylation. *Science* **343**, 1235681 (2014).
4. J. D. Esko, S. B. Selleck, Order out of chaos: Assembly of ligand binding sites in heparan sulfate. *Annu. Rev. Biochem.* **71**, 435–471 (2002).
5. R. J. Linhardt, 2003 Claude S. Hudson Award address in carbohydrate chemistry. Heparin: Structure and activity. *J. Med. Chem.* **46**, 2551–2564 (2003).
6. M. Ly *et al*., The proteoglycan bikunin has a defined sequence. *Nat. Chem. Biol.* **7**, 827–833 (2011).
7. M. Ly, T. N. Laremore, R. J. Linhardt, Proteoglycomics: Recent progress and future challenges. *OMICS* **14**, 389–399 (2010).
8. M. Kizer *et al*., RNA aptamers with specificity for heparosan and chondroitin glycosaminoglycans. *ACS Omega* **3**, 13667–13675 (2018).
9. J. Zaia, Mass spectrometry and glycomics. *OMICS* **14**, 401–418 (2010).
10. Y. Chang *et al*., Ultrasensitive detection and quantification of acidic disaccharides using capillary electrophoresis and quantum dot-based fluorescence resonance energy transfer. *Anal. Chem.* **85**, 9356–9362 (2013).

11. S. J. Kwon *et al*., Signal amplification by glyco-qPCR for ultrasensitive detection of carbohydrates: Applications in glycobiology. *Angew. Chem. Int. Ed. Engl.* **51**, 11800–11804 (2012).
12. M. Kalita *et al*., A nanosensor for ultrasensitive detection of oversulfated chondroitin sulfate contaminant in heparin. *J. Am. Chem. Soc.* **136**, 554–557 (2014).
13. M. T. Noakes *et al*., Increasing the accuracy of nanopore DNA sequencing using a time-varying cross membrane voltage. *Nat. Biotechnol.* **37**, 651–656 (2019).
14. Y. Feng, Y. Zhang, C. Ying, D. Wang, C. Du, Nanopore-based fourth-generation DNA sequencing technology. *Genomics Proteomics Bioinf.* **13**, 4–16 (2015).
15. C. Shasha *et al*., Nanopore-based conformational analysis of a viral RNA drug target. *ACS Nano* **8**, 6425–6430 (2014).
16. D. R. Garalde *et al*., Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
17. O. K. Zahid, F. Wang, J. A. Ruzicka, E. W. Taylor, A. R. Hall, Sequence-specific recognition of microRNAs and other short nucleic acids with solid-state nanopores. *Nano Lett.* **16**, 2033–2039 (2016).
18. J. Saharia *et al*., Molecular-level profiling of human serum transferrin protein through assessment of nanopore-based electrical and chemical responsiveness. *ACS Nano* **13**, 4246–4254 (2019).
19. E. C. Yusko *et al*., Real-time shape approximation and fingerprinting of single proteins using a nanopore. *Nat. Nanotechnol.* **12**, 360–367 (2017).
20. D. J. Niedzwiecki *et al*., Observing changes in the structure and oligomerization state of a helical protein dimer using solid-state nanopores. *ACS Nano* **9**, 8907–8915 (2015).
21. Z. Dong, E. Kennedy, M. Hokmabadi, G. Timp, Discriminating residue substitutions in a single protein molecule using a sub-nanopore. *ACS Nano* **11**, 5440–5452 (2017).
22. P. Waduge *et al*., Nanopore-based measurements of protein size, fluctuations, and conformational changes. *ACS Nano* **11**, 5706–5716 (2017).
23. J. Quick *et al*., Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
24. S. L. Castro-Wallace *et al*., Nanopore DNA sequencing and genome assembly on the 10 international space station. *Sci. Rep.* **7**, 18022 (2017).
25. M. Waugh *et al*., Solid-state nanopore fabrication by automated controlled breakdown. *Nat. Protoc.* **15**, 122–143 (2020).
26. H. Kwok, K. Briggs, V. Tabard-Cossa, Nanopore fabrication by controlled dielectric breakdown. *PLoS One* **9**, e92880 (2014).
27. B. I. Karawdeniya *et al*., Challenging nanopores with analyte scope and environment. *J. Anal. Test.* **3**, 61–79 (2019).
28. B. I. Karawdeniya, Y. M. N. D. Y. Bandara, J. W. Nichols, R. B. Chevalier, J. R. Dwyer, Surveying silicon nitride nanopores for glycomics and heparin quality assurance. *Nat. Commun.* **9**, 3278–3285 (2018).
29. A. Fennouri *et al*., Single molecule detection of glycosaminoglycan hyaluronic acid oligosaccharides and depolymerization enzyme activity using a protein nanopore. *ACS Nano* **6**, 9672–9678 (2012).
30. F. Rivas *et al*., Label-free analysis of physiological hyaluronan size distribution with a solid-state nanopore sensor. *Nat. Commun.* **9**, 1037 (2018).
31. J. Im, S. Lindsay, X. Wang, P. Zhang, Single molecule identification and quantification of glycosaminoglycans using solid-state nanopores. *ACS Nano* **13**, 6308–6318 (2019).
32. U. Bhaskar *et al*., Engineering of routes to heparin and related polysaccharides. *Appl. Microbiol. Biotechnol.* **93**, 1–16 (2012).
33. X. Zhang, L. Lin, H. Huang, R. J. Linhardt, Chemoenzymatic synthesis of glycosaminoglycans. *Acc. Chem. Res.* **53**, 335–346 (2020).
34. B. F. Cress *et al*., Heavy heparin: A stable isotope-enriched, chemoenzymatically-synthesized, poly-component drug. *Angew. Chem.* **58**, 5962–5966 (2019).
35. S. Winters-Hilt *et al*., Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules. *Biophys. J.* **84**, 967–976 (2003).
36. S. Winters-Hilt, M. Akeson, Nanopore cheminformatics. *DNA Cell Biol.* **23**, 675–683 (2004).
37. P. Godec *et al*., Democratized image analytics by visual programming through integration of deep models and small-scale machine learning. *Nat. Commun.* **10**, 4551 (2019).
38. K. C. Chou, H. B. Shen, Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **3**, 153–162 (2008).
39. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the inception architecture for computer vision" in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2016), pp. 2818–2826.
40. N. Silberman, S. Guadarrama, TensorFlow-Slim image classification model library. GitHub. https://github.com/tensorflow/models/tree/master/research/slim. Data accessed 1 February 2020.
41. J. Demsar *et al*., Orange: Data mining toolbox in python. *J. Mach. Learn. Res.* **14**, 2349–2353 (2013).