JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Article

Predicting Structural Motifs of Glycosaminoglycans using Cryogenic Infrared Spectroscopy and Random Forest

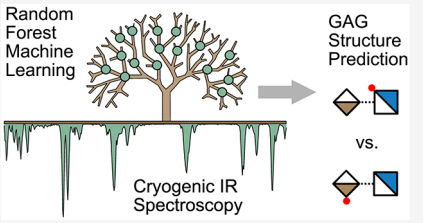
Jerome Riedel,° Maike Lettow,° Márkó Grabarics, Michael Götze, Rebecca L. Miller, Geert-Jan Boons, Gerard Meijer, Gert von Helden, Gergo Peter Szekeres,* and Kevin Pagel*

Cite This: *J. Am. Chem. Soc.* 2023, 145, 7859–7868

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In recent years, glycosaminoglycans (GAGs) have emerged into the focus of biochemical and biomedical research due to their importance in a variety of physiological processes. These molecules show great diversity, which makes their analysis highly challenging. A promising tool for identifying the structural motifs and conformation of shorter GAG chains is cryogenic gas-phase infrared (IR) spectroscopy. In this work, the cryogenic gas-phase IR spectra of mass-selected heparan sulfate (HS) di-, tetra-, and hexasaccharide ions were recorded to extract vibrational features that are characteristic to structural motifs. The data were augmented with chondroitin sulfate (CS) disaccharide spectra to assemble a training library for random forest (RF) classifiers. These were used to discriminate between GAG classes (CS or HS) and different sulfate positions (2-*O*-, 4-*O*-, 6-*O*-, and *N*-sulfation). With optimized data preprocessing and RF modeling, a prediction accuracy of >97% was achieved for HS tetra- and hexasaccharides based on a training set of only 21 spectra. These results exemplify the importance of combining gas-phase cryogenic IR ion spectroscopy with machine learning to improve the future analytical workflow for GAG sequencing and that of other biomolecules, such as metabolites.



Downloaded via FREIE UNIV BERLIN on June 8, 2023 at 08:09:58 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

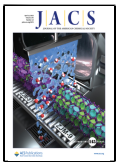
INTRODUCTION

Glycosaminoglycans (GAGs) are linear sulfated polysaccharides that are involved in a variety of biological processes such as cell adhesion, blood coagulation, cell-to-cell communication, and regulatory interactions with chemokines and growth factors.[1,2] Although GAGs consist of disaccharide units that are linked strictly linearly, they show great structural diversity arising from differences in chain length, monomer configuration, and sulfation degree and position. Generally, GAGs are divided into four major classes: heparin/heparan sulfate (Hp/HS), chondroitin sulfate/dermatan sulfate (CS/DS), keratan sulfate (KS), and hyaluronic acid (HA), with HS and CS exhibiting more diverse structures in comparison to KS and HA (Figure 1a).[3,4] Sequencing GAGs is of paramount importance to understand their physiological activity and interactions; however, it remains challenging due to their structural complexity and the limited availability of pure samples.[5] The overall composition of a GAG chain can be assessed using lyase digestion and subsequent liquid chromatography of the resulting disaccharides.[6] Mass spectrometry-based experiments coupled with liquid chromatography, ion mobility, and different fragmentation techniques can further be used to identify characteristic oligosaccharide motifs,[4,7−9] although the analysis of the fragmentation data is a complex task and may not always lead to unambiguous assignments.[4]

In the past few years, gas-phase vibrational spectroscopy of GAGs, and glycoconjugates in general, has emerged as a tool that provides complementary information to mass-spectrometry-based sequencing. Common gas-phase spectroscopic techniques range from infrared multiple photon dissociation experiments at ambient temperatures[10] to tagging spectroscopy in cryogenic ion traps[11] or helium nanodroplet spectroscopy.[12,13] The obtained vibrational fingerprint can be correlated with the structural motifs and functional groups present within the molecule,[13−15] which enables the identification and/or distinction of even minute structural details. Due to the ability of cryogenic gas-phase spectroscopy approaches to resolve even small shifts in band position, it has the potential to unravel the binding of GAGs to different target molecules. Here, changes in the IR spectrum of an intact and/or dissociated GAG complex might yield informative insights into GAG-specific interaction and binding patterns.

In this work, GAGs were studied using helium nanodroplet spectroscopy, where the ions are captured in superfluid helium nanodroplets and rapidly cooled down to 0.4 K. Irradiation with IR light leads to vibrational excitation of the ions in the droplet. The absorbed energy is redistributed from the ion to the matrix, resulting in partial evaporation and shrinking of the

Received: November 30, 2022
Published: March 31, 2023



ACS Publications

© 2023 The Authors. Published by American Chemical Society

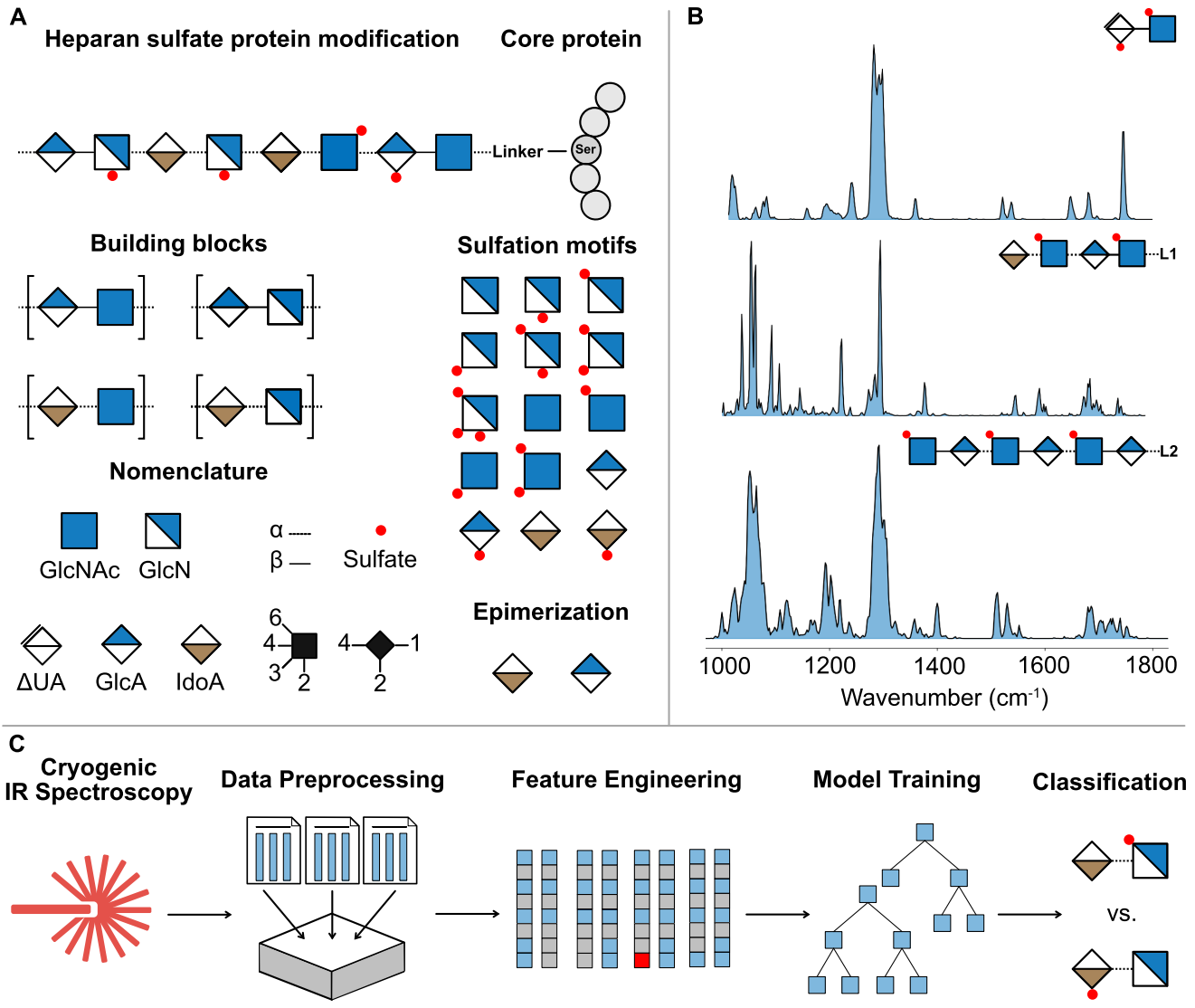


Figure 1. Random Forest analysis of glycosaminoglycan IR spectra. (a) Example of a heparan sulfate oligosaccharide linked to a protein. The structural complexity is defined by disaccharide building blocks, α/β-glycosidic bonds, and differences in sulfation degree and position.[31] (b) Helium nanodroplet spectra capture the molecular fingerprint of GAGs. Structural motifs can be directly inferred from the position and intensity of vibrational bands in the spectrum as shown here for di-, tetra-, and hexasaccharides. (c) Schematic representation of the pipeline for training a Random Forest model starting from raw experimental data.

helium nanodroplets, thus maintaining a temperature of 0.4 K. Eventually, the trapped ion is ejected and can be detected in a time-of-flight mass spectrometer.[13,16] Measuring molecules in a liquid-nitrogen-cooled ion trap that are picked up in helium nanodroplets at 0.4 K will purge the conformational ensemble from higher energy conformers. This results in a few low-energy conformers in their local minima, which effectively leads to decongestion of the IR spectrum.[17] Moreover, the continuous cooling of ions by the superfluid helium will always relax the ions back to their vibrational ground state, thus reducing the spectral broadening. Therefore, ion spectra at cryogenic temperatures enable the systematic evaluation of vibrational band occurrences across a wide range of structural motifs.

Machine learning, especially cluster and pattern recognition algorithms are now frequently used in science to analyze complex data sets, as well as to solve problems where resource consumption is overwhelming.[18−20] Random Forest (RF) modeling is among the more popular algorithms used for pattern recognition in IR and Raman spectroscopy, and has been successfully applied, e.g., to identify diseases in serum samples and to follow physiological processes in cells.[21−23] Since its development by Breiman,[24] RF became a well-established supervised machine learning technique and is known to perform well for medium to large size data sets, while having only minimal requirements on data type and feature correlation.[24−27] RF classifiers train an ensemble of decision trees (CART algorithm) to partition the feature space into a network of decision rules that, upon traversal, answer the classification task. Given a data set with n samples and k features, RF models are trained on training sets holding m samples, where $m \leq n$. A selection of features, which best partition the feature space, is directly incorporated in CARTs, but the selection can be suboptimal if the number of available samples is too low. In spectroscopic data, the dimensionality of the feature space regularly outweighs the number of samples available for training, which first requires feature reduction to avoid later misclassifications.[28] In the past, different feature

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

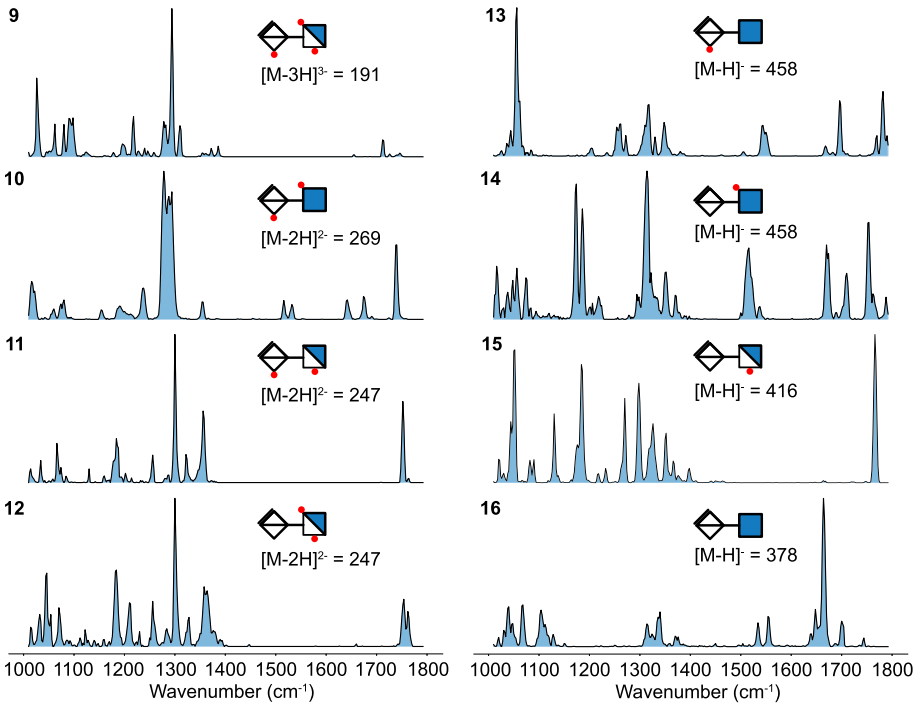


Figure 2. Helium nanodroplet spectra of HS disaccharides with varying sulfation patterns.

engineering strategies have been applied to reduce their number and to identify the most important features for the classification objective, e.g., principal component analysis, Wilcoxon testing, and evolutionary algorithms.[21,29,30]

Here, we present a workflow that establishes an IR spectral library of HS and CS di- and tetrasaccharides (Figure 1b) to train RF classifiers (Figure 1c) for predicting structural motifs in tetra- and hexasaccharide GAGs from their vibrational fingerprint. This approach opens up new possibilities to predict structural motifs in GAG oligosaccharides and helps with the identification of important GAG binding sequences for the design of novel pharmaceuticals. Moreover, it may serve as a blueprint for the analysis of other biomolecular classes, such as metabolites.

MATERIALS AND METHODS

A library of HS and CS disaccharide to hexasaccharide standards was assembled to cover the full chemical space of GAG features, including sulfation variations (*N*-sulfation versus *N*-acetylation, 2-*O*-, 4-*O*-, and 6-*O*-sulfation), epimerization (IdoA or GlcA) and backbone diversity (Hp/HS or CS/DS). The spectrum library includes 16 disaccharides (CS **1**−**8**, HS **9**−**16**), six tetrasaccharides (**17**−**22**), and one hexasaccharide (**23**) (Table S1). HS disaccharides were purchased from Iduron (Manchester, United Kingdom). All disaccharide standards have a reduced hexuronic acid (ΔUA) at the nonreducing end obtained by bacterial heparinase and -chondroitinase cleavage (β-elimination) of HS/CS oligosaccharides, respectively. The tetrasaccharides and hexasaccharide were synthesized chemically as described previously.[32,33] Solvents (HPLC grade) were purchased from Sigma-Aldrich (StLouis, USA). Prior to use, all glycans were dissolved in water/methanol (v/v, 50/50%) to yield 50 μM analyte solutions. The gas-phase IR spectra for the library were recorded on the custom-built helium nanodroplet instrument in the range of 1000−1800 cm^{-1}. A detailed description of the experimental setup can be found elsewhere.[13,16]

Computational Details. The experimental IR spectra were extracted from time-of-flight intensity values measured as a function of the laser wavelength. The ion signal was then normalized to the repetition rate and energy of the laser as a first order approximation. Linear interpolation with 2 cm^{-1} steps was employed to align spectra on a commonly shared wavenumber axis before finding the largest shared wavenumber range among spectra (1010−1786 cm^{-1}). In addition to the spectroscopic features, information about the charge and degree of sulfation was included, as well as on the presence of a linker when tetrasaccharides were part of the training set. The wavenumber features were binned by integration of the IR trace in windows of fixed width (15 cm^{-1}, Figure S6) for the purpose of feature reduction (see Supporting Information, Data Pre-Processing). After spectral binning, bins were normalized to [0,1] to minimize the influence of absolute intensity values as latent variables in growing the decision trees, since in helium nanodroplet spectroscopy, intensities are not linear due to the absorption of multiple photons before the ions are ejected. This can result in over- or underestimating absorption cross sections of vibrational modes.[34] For model training, feature selection was performed using an evolutionary optimization algorithm as implemented in the *GAFeatureSelectionCV* class provided by *sklearn-genetic-opt*.[35,36] In each generation, the algorithm uses reproduction, mutation, and selection stages to choose the locally optimal features that will be present in the offspring generation. Construction of the offspring generation is made according to the $\mu + \lambda$ algorithm (see Supporting Information, Model Training).[37] Additionally, elitism was used in each generation to maintain a set of the fittest features across evolutions. Classification was performed using an RF model trained by a Leave-One-Out cross-validation approach. Computationally, the *RandomForestClassifier* class provided by *scikit_learn* was used,[38] which uses an optimized version of the CART algorithm for model training.[39] The splitting criterion (partitioning of the sample space) at every node level was evaluated according to the entropy formulation. Detailed descriptions of the algorithms used for feature selection and model training can be found in the Supporting Information, along with a summary of the model training parameters used in the setup of feature selection and classifier classes (Tables S4 and S5 and Figures S6 and S7).

RESULTS AND DISCUSSION

The set of HS di-, tetra-, and hexasaccharides form a diverse spectral library, from which prominent structural motifs can be

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

inferred. The complete set of CS disaccharide and the **17**−**20** HS tetrasaccharide spectra were published previously (Figure S2 and S3).[14,15] In Figure 2, the IR spectra of HS disaccharides **9**−**16** are shown in the range of 1000 to 1800 cm^{-1}. Selected vibrational features (summarized in Table 1) were assigned based on previous experimental and theoretical results of glycan cations and anions.[14,40,41] In sulfated GAGs, the 1000 to 1150 cm^{-1} range is populated by ν(C−O) and ν(C−C) modes, which overlap with the symmetric stretching of charged sulfates. The strong antisymmetric SO_3^- stretching vibrations appear between 1150 to 1350 cm^{-1}. These features serve as a unique fingerprint of the sulfation pattern in glycans. Further minor bands between 1200 and 1500 cm^{-1} can be assigned to C−H deformation and O−H bending modes. The distinct amide I−III modes between 1300 and 1700 cm^{-1} (of which amide III is generally very weak in IR spectra) stem from *N*-acetylation in GAGs. The amide I, with major contribution from ν(C═O), is typically the most intense amide region in the spectra, and it is highly sensitive to the intramolecular hydrogen bonding pattern associated with the conformation of the glycan.[15] The amide II region is mainly populated by the N−H in-plane bending vibration with decreasing contributions from the C−N stretching motion and other minor components. The weak stretching mode ν(C═C) around 1600 cm^{-1} is characteristic for uronic acid monosaccharides derived from lyase digestion, which carry a C═C bond at the nonreducing end. Above 1700 cm^{-1}, the C═O stretching vibration comes from the neutral carboxyl group in the hexuronic acid moiety. In nonsulfated GAGs, the charged carboxylate functional group yields a strong peak around 1700 cm^{-1}, corresponding to the antisymmetric COO^- stretching.

All HS disaccharides **9**−**16** show unique IR signatures. *N*-acetylation (and therewith the lack of *N*-sulfation) is easy to determine due to the presence of the amide I (1600−1700 cm^{-1}) and amide II (1500−1600 cm^{-1}) vibrations in the IR fingerprints of disaccharides **10**, **13**, **14**, and **16**. A silent amide region can indicate that HS is either *N*-sulfated or has a free primary amine. This can later be circumvented by analyzing the 1294−1300 cm^{-1} range, where a single strong mode at ≈1300 cm^{-1} (or red-shifted in case of high sulfation) can indicate *N*-sulfation. The *O*-sulfation motif in these ions (2-*O*- or 6-*O*-sulfation) is more challenging to differentiate based only on the IR spectra.

The number of conformers captured in the cryogenic ion trap can vary even for ions of very similar chemical structure.[15] A double band in the IR spectrum of disaccharide **10** with a shift in the amide I band of ≈35 cm^{-1} can indicate that two conformers (or anomers) were present in the trap. The number of conformers in the cryogenic ion trap for disaccharides **9**, **11**, and **15** appeared to be rather low because the neutral carboxylic acid only yields a single ν(C═O) mode. In the sulfated HS disaccharides, this band shows an apparent tendency: with increasing sulfation, the ν(C═O) stretching vibration red-shifts from the 1753−1781 cm^{-1} region (singly sulfated) to the 1739−1762 cm^{-1} region in the doubly sulfated HS ions, and eventually to 1712 cm^{-1} in the triply sulfated HS ion. This red-shift is most probably a contribution of the electron-withdrawing characteristics of sulfates, with increasing sulfation reducing the electron density around the carboxyl moiety. In the spectra of **11**, **12**, and **14**, the ν(C═O) band appears at 1753 and 1763 cm^{-1} as a double peak, while the region below 1400 cm^{-1} is more populated with (often weak) modes compared to the other spectra. This leads to two

Table 1. Tentative Assignment of Selected Vibrational Modes[a] in Figure 2

Frequencies (cm^{-1})	9	10	11	12	13	14	15	16
$\nu_a(SO_3^-)$	1217, 1279, 1294, 1311	1238, 1278, 1287, 1294, 1355	1185, 1256, 1300, 1324, 1356	1183, 1212, 1316, 1283, 1301, 1327, 1361	1260, 1272, 1218, 1330, 1349	1173, 1186, 1218, 1296, 1314, 1323, 1334, 1351, 1370	1185, 1217, 1233, 1271, 1297, 1326, 1351, 1367, 1397	—
amide II	—	1517, 1533	—	—	1543, 1550	1515, 1538	—	1534, 1555
amide I	—	1641, 1675	—	—	1697	1670, 1710[b]	—	1647, 1701
ν(C═O)	1713	1739	—	1753, 1762	1769, 1782, 1790	1754, 1763, 1788	1766	—
$\nu_a(COO^-)$	—	—	—	—	—	—	—	1664

[a]ν_a designates the (antisymmetric) stretching mode. [b]or ν(C═O) mode of the neutral carboxylic acid

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

conclusions: these three ions can exist in a more extended conformational space than the other ions, while they also share a highly similar chemical environment around the carboxyl moiety. The ν(C═O) region in the spectrum of **13** contains several additional bands, with unique band positions compared to the previous case pointing toward a different chemical environment for the carboxylic moiety. This can result from the fact that the only charged sulfate in this case is located on the same monosaccharide unit as the carboxyl group itself. The three charged sulfates in disaccharide **9** show intense $\nu_a(SO_3^-)$ vibrational features in the IR spectrum, which can overpower the ν(C═O) and account for its relatively low intensity. It is important to note that the relative intensities only serve as a qualitative measure in helium nanodroplet spectroscopy.[34] Interestingly, the most intense band in the IR spectrum of disaccharide **13**, which carries 2-*O*-sulfation, is at 1055 cm^{-1}. The ν(C−O) and ν(C−C) of the glycan core and the $\nu_s(SO_3^-)$ in this region are difficult to differentiate; therefore, this strong mode remains unassigned. The cryogenic IR spectra of the HS tetrasaccharides **21** and **22** are shown in Figure 3.

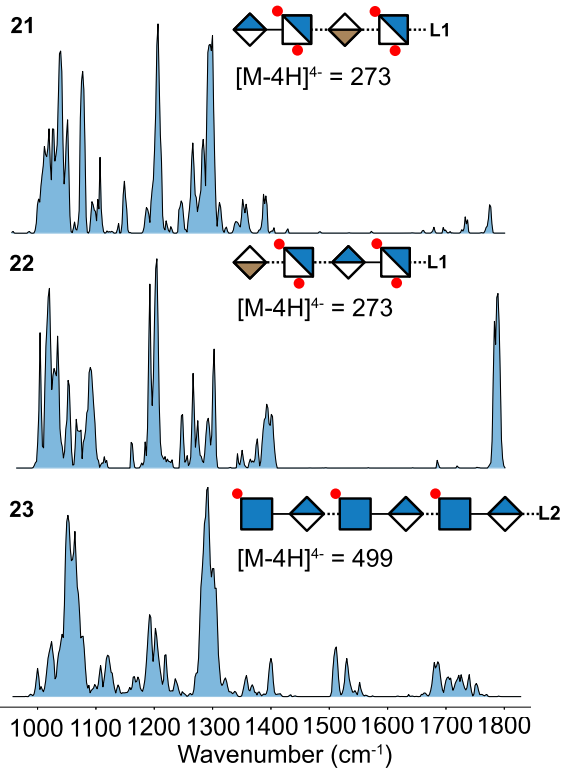


Figure 3. Helium nanodroplet spectra of HS tetrasaccharides and hexasaccharide from 1000 to 1800 cm^{-1}. The tetrasaccharide diastereomers were labeled with an aminopentyl linker (L1), while the hexasaccharide carried a *p*-methoxyphenyl linker (L2).

The two diastereomers can be differentiated from their IR signature (Table 2). The aminopentyl linker at the reducing end has additional contributions in the 1200−1500 cm^{-1} region corresponding to C−H deformation, while the scissoring motion of neutral, primary amines are typically found around 1600 cm^{-1} and are of very low intensity (Table 2).[42,43] This highly flexible linker can form interactions of different strength with several different functional groups in the molecules, which may distribute their vibrational contribution over an extended wavenumber range, thus resulting in an overall weak NH_2 signal.

Table 2. Tentative Assignment of Selected Diagnostic Modes[a] in the Spectra of HS Tetra- and Hexasaccharides 21−23[14]

Frequencies (cm^{-1})	21	22	23
$\nu_a(SO_3^-)$	1150, 1207, 1267, 1286, 1296, 1301, 1357, 1393	1160, 1191, 1203, 1247, 1266, 1274, 1292, 1302, 1392, 1400	1192, 1202, 1290, 1301, 1400[b]
amide II	—	—	1511, 1530
amide I	—	—	1684, 1704[b]
ν(C═O)	1739, 1780	1781, 1787	1725, 1740

[a] $\nu(\nu_a)$ designates the (antisymmetric) stretching mode. [b] or ν(C═O) mode of the neutral carboxylic acid

Random Forest. Since the data set of di-, tetra-, and hexasaccharide samples is small compared to those usually used in RF model training, one of the main tasks was to optimize the feature-to-sample ratio. Otherwise, RF models are prone to overfitting, which in the case of IR spectra results in the selection of vibrational features that are not associated with the structural motif in question. For classification models working on high correlation data sets, such as IR spectra, in which neighboring IR channels are highly correlated, the optimal feature-to-sample ratio can be approximated by $k \approx \sqrt{m}$ features selected in the final model.[28,44] Furthermore, at every node level during training of the decision trees in the RF model, a set of $\sqrt{k}$ features is considered for partitioning, where k is the total number of features in the training set.[44,45] With a total number of 59 features after spectral binning, additional feature selection by an evolutionary algorithm had to take place to constrain the size of the feature space and mitigate overfitting effects. This step resulted in seven features used per decision function on average, with a maximum of 23 features used for HS/CS classifications.

In the joined data set of disaccharides and tetrasaccharides, the set of HS and CS disaccharides presented the purest, most complete, and most informative subset based on the number of unique sulfation patterns covered. However, as RF performance correlates with the size of the data set, training runs were initially focused on the performance of the algorithm for training sets smaller than 16 (the number of disaccharides). In this first step, training sets containing 15 disaccharides were assembled to evaluate the classification outcome in predicting the structure (HS or CS) and the presence or absence of *N*-, 2-*O*-, 4-*O*-, and 6-*O*-sulfation. This yielded a total of 80 models trained to cover the complete combinatorial space (Figure 4). To assess the impact of the training set on the model robustness, a prediction score was defined that averages the number of correct classifications across all excluded samples (see Supporting Information, Model Evaluation). The individual prediction scores were subsequently used to evaluate the overall prediction score over all training sets. The models predicted the correct structural motifs in 73% of the cases. High prediction accuracies were achieved in classification tasks for 4-*O*- and *N*-sulfation, with prediction scores of 94% and 81%, respectively. The feature with the best prediction scores (4-*O*-sulfation) classified all structures correctly, except for **6**, where the algorithm missed the spectral traits of the present 4-*O*-sulfation. Lower prediction accuracies were observed in predicting 2-*O*-sulfation (73%) and 6-*O*-sulfation (63%). In

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

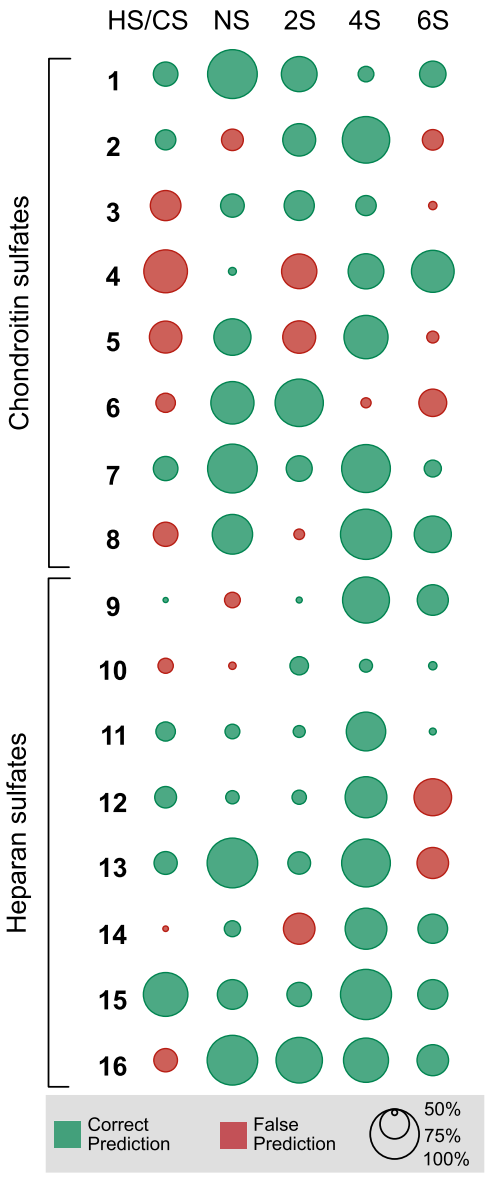


Figure 4. Prediction outcome and confidence for disaccharide classifications in the complete combinatorial space of excluding one sample from the base set. The prediction is for HS/CS, and the presence or absence of *N*-, 2-*O*-, 4-*O*-, and 6-*O*-sulfation (NS, 2S, 4S, and 6S, respectively).

the case of 6-*O*-sulfation (with the worst prediction score among sulfation motifs), the structures **2**, **5**, and **6** were incorrectly assigned as 6-*O*-sulfated molecules, while the structures **3**, **12**, and **13** were falsely labeled as non-6-*O*-sulfated molecules. Similar accuracies had already been reported for functional group prediction from FTIR spectra of small organic molecules, with the accuracy depending on the functional group in question.[46] Generally, high prediction accuracy is expected for *N*-sulfation due to the absence of the amide vibrations, thus rendering the 1200−1700 cm^{-1} range silent. In comparison, the lowest prediction accuracy is expected to be associated with HS/CS classifications, since HS and CS can potentially have varying monomer units and associated sulfation patterns, yielding nondiagnostic vibrational modes. This leads to more complex and ambiguous decision rules in the RF classifier. In the trained models, the classification for HS/CS remained inconclusive; the challenge of correctly classifying such structural differences is also reflected in the number of features selected to train the RF classifiers, with the highest feature count observed for HS/CS classifications.

The majority of features used in predicting *N*-sulfation were between 1200 and 1700 cm^{-1} (Table S8), covering the complete range of amide vibrations. The lack of amide vibrations in *N*-sulfated HS disaccharides resulted in strong decision rules in individual decision trees (i.e., intensities of ≈0 for *N*-acetylation). However, this spectral region is also populated by weak R−CH═CH−R and red-shifted carbonyl C═O stretching modes, which could give rise to false predictions. Based on Figure 4, an overall better prediction accuracy is observed for HS samples than for CS samples. This is due to the fact that in CS, sulfation at the 2-*O*, 4-*O*, and 6-*O* position must be predicted from vibrational modes in the range of 1150−1400 cm^{-1}, whereas this region is only used to predict 2-*O*- and 6-*O*-sulfation in HS, allowing for better distinction. It is important to consider that the prediction confidence for false predictions was generally lower than for correct predictions, indicating that upon increase of the training set size, classification performance can likely be improved as further information in this region is added.

To evaluate the influence of the training set size on the prediction score, IR spectra of tetrasaccharide samples **17**−**22** were systematically added to the training set. For model training, the full set of disaccharide IR spectra presented the base set of information and were always included. The tetrasaccharide samples were different not only in terms of the number of disaccharide building blocks, but also because of the presence of the aminopentyl linker. Therefore, a "label" feature was introduced to mark the difference between di- and tetrasaccharides by accounting for C−H and C−N deformations of the linker that were not yet part of the training set. By combinatorially adding tetrasaccharides to the base set, 62 unique training sets ($X_{m=17} = 6$, $X_{m=18} = 15$, $X_{m=19} = 20$, $X_{m=20} = 15$, $X_{m=21} = 6$) were assembled, leading to a total of 310 models trained for the five structural motifs. To evaluate the classification performance, the prediction score was calculated against the excluded tetrasaccharide samples and averaged over all predictions. Figure 5 shows that with increasing training set size, the prediction accuracy expectedly increased and approached a maximum prediction score of 97% for training

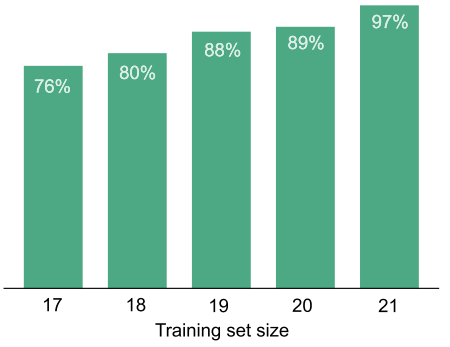


Figure 5. Prediction score as a function of the training set size. For each training set, RF classifiers were trained for the five classification tasks. The prediction outcome was averaged over all classifications and all training sets in the respective subset. The subsets were obtained by combinatorially adding tetrasaccharides to the disaccharide base set.

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

sets with m = 21 samples. The RF classifiers were trained with bootstrap aggregation, i.e., only a subset of samples in the training set was used for growing the decision tree. This led to more diverse subsets due to the increased training set size, resulting in a minimized out-of-bag error rate in testing against the out-of-bag samples.[26,47] A smaller out-of-bag error rate directly translates to higher prediction scores in evaluation runs against unknown samples,[24] which is in agreement with the observations made in our classifications. In $X_{m=21}$ training sets, HS, NS, 2S, and 4S structural motifs were predicted correctly across all excluded samples, with only a single classification for 6-*O*-sulfation failing with a confidence of 60% (Figure 6). The

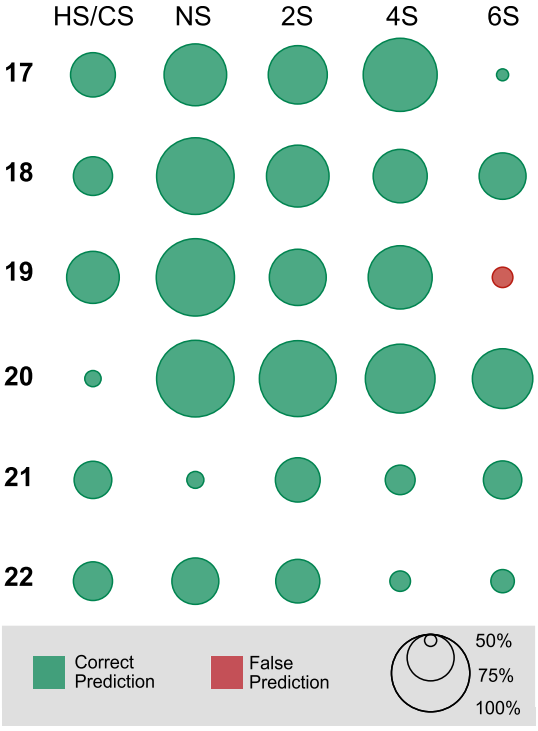


Figure 6. Prediction outcome and confidence in $X_{m=21}$ training sets tested against the excluded sample. The classification outcomes are HS/CS, and the presence or absence of *N*-, 2-*O*-, 4-*O*-, and 6-*O*-sulfation (NS, 2S, 4S, and 6S, respectively).

selected features for 6-*O*-sulfation were the "label" feature and wavenumber bins at 1122, 1212, 1467, 1587, 1632, and 1737 cm^{-1}. Sample **19** is structurally related to samples **17**, **18**, and **20**, with the sulfation at the 6-*O* position being located on a GlcNAc unit. Comparing the IR spectra of samples **17** and **18** to the IR spectrum of sample **19**, clear intensity differences above 1450 cm^{-1} are visible, which was already linked to the presence of the terminal IdoA-GlcNAc6S-L1 building block for samples **19** and **20**.[14] Therefore, by removing sample **19** from the data set, ambiguity was likely introduced into the model. In comparison, the model trained against sample **20** used the feature "charge" and wavenumber bins of 1212, 1287, and 1467 cm^{-1} for 6-*O*-sulfation, effectively avoiding the regions of intensity differences between the diastereomers, which yielded the overall correct prediction and higher prediction confidence (Table S9).

Although the "label" feature was not selected in all models (the evolutionary feature selection strategy employed here only yields local optima), it provided an unambiguous pathway to identify samples as HS. Since the "label" feature was not used in all RF classifiers, we believe that including this information may not be necessary. This is important from the standpoint of sample preparation, and for moving toward an analytical technique to predict structural motifs in truly unknown GAGs.

To assess the robustness of the method, we further extended the set of excluded samples by a hexasaccharide with the so far unknown *p*-methoxyphenyl linker. Upon validation of $X_{m=21}$ and $X_{m=22}$ training sets, all models classified the structural motifs of the hexasaccharide correctly (Figure 7). This

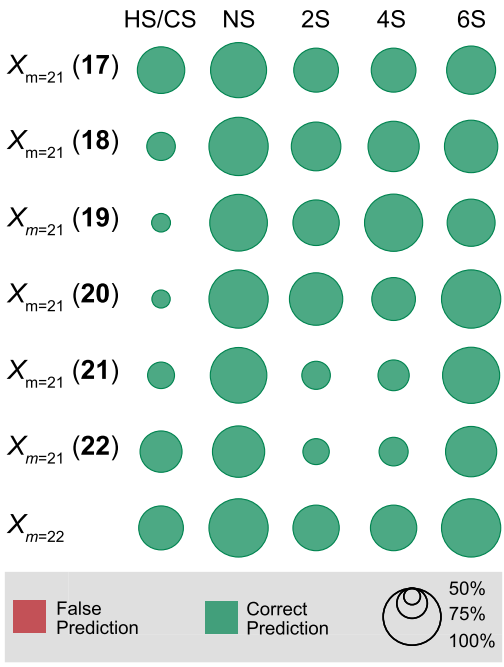


Figure 7. Prediction accuracy and probability of predicting the structural motifs of hexasaccharide 23 from $X_{m=21}$ and $X_{m=22}$ training sets. The sample in parentheses marks the tetrasaccharide that was not included in the training set. The prediction outcomes are HS/CS, and the presence or absence of *N*-, 2-*O*-, 4-*O*-, and 6-*O*-sulfation (NS, 2S, 4S, and 6S, respectively). Overall best prediction confidence was obtained for the $X_{m=22}$ training set.

indicates that training sets consisting of merely 21 high-resolution IR spectra are sufficient to make high-confidence predictions about structural motifs in unknown GAGs. Therefore, RF classification models trained on a representative sample set of different sulfation motifs can provide a computationally fast method for an initial estimate of GAG class and sulfation characteristics that provides complementary information to conventional mass spectrometry and ion mobility-based techniques. More importantly, the overall observed prediction accuracy suggests that a diverse library of smaller synthetic oligosaccharide fragments is sufficient to predict structural motifs in larger oligosaccharides, which are not yet accessible by chemical synthesis. Eventually, the outlined approach has the potential to be implemented in other cryogenic action spectroscopy approaches (e.g., tagging spectroscopy), where cryogenic temperatures increase spectral resolution and purge higher-energy conformers.

CONCLUSION

In this work, helium nanodroplet spectra were recorded and used in Random Forest classifications to predict the class (HS/CS) and sulfate positions (2-*O*-, 4-*O*-, 6-*O*-, and *N*-sulfation) of

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

GAG oligosaccharides. Even though certain structural motifs can be directly inferred from the vibrational signatures, it was expected that machine learning will lead to further improvement in feature identification and, with that, facilitate the automated identification of GAG functional groups. The implementation of an RF workflow revealed that, on surprisingly small training sets with as low as 21 samples, high confidence predictions of GAG structural motifs can be made. More importantly, the data indicate that a training set of small synthetic reference compounds such as di- and tetrasacchrides is sufficient to reliably predict structural motifs in larger structures up to hexasaccharides. This implies that structural annotations do not necessarily require an extensive set of synthetic standards to cover the full structural space. Instead, a few smaller, synthetically accessible molecules can be used to train the model for the structural prediction of larger, more complex GAG oligosaccharides, which are often not accessible by chemical synthesis. As a result, cryogenic gas-phase IR spectroscopy in combination with RF has exceptional potential to sequence larger GAG oligosaccharides to full-length GAG chains and serves as a blueprint for the analysis of other biomolecules, such as metabolites. In conjunction with classical ion-mobility and mass-spectrometry-based techniques, this has important implications for the understanding of GAG binding sequences, which can aid in the design and efficacy enhancement of novel and existing pharmaceuticals that target GAG structural motif patterns for disease treatment.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.2c12762.

Additional sample identification, spectra, software descriptions, and model information (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Gergo Peter Szekeres − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0002-4597-6264; Email: gpszekeres@fhi.mpg.de

Kevin Pagel − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0001-8054-4718; Email: kevin.pagel@fu-berlin.de

Authors

Jerome Riedel − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

Maike Lettow − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0003-4063-1977

Márkó Grabarics − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* Present Address: University of Oxford, Department of Chemistry, Oxford, OX1 3TA, United Kingdom; orcid.org/0000-0002-2550-637X

Michael Götze − *Department of Biology, Chemistry, and Pharmacy, Freie Universität Berlin, Berlin 14195, Germany; Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany*

Rebecca L. Miller − *Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen Center for Glycomics, Copenhagen N 2200, Denmark;* orcid.org/0000-0001-8574-1948

Geert-Jan Boons − *Bijvoet Center for Biomolecular Research, Utrecht University, 3584 CG Utrecht, The Netherlands; Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia 30602-0002, United States*

Gerard Meijer − *Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0001-9669-8340

Gert von Helden − *Department of Molecular Physics, Fritz-Haber-Institut der Max-Planck-Gesellschaft, 14195 Berlin, Germany;* orcid.org/0000-0001-7611-8740

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.2c12762

Author Contributions

°JR and ML contributed equally

Funding

Open access funded by Max Planck Society.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support for this research was provided by the European Union's Horizon 2020 Research and Innovation Programme grant number 899687-HS-SEQ. The authors would like to thank the HPC Service of ZEDAT, Freie Universität Berlin, for computing time.[48]

■ REFERENCES

(1) Linhardt, R. J.; Toida, T. Role of Glycosaminoglycans in Cellular Communication. *Acc. Chem. Res.* **2004**, *37*, 431−438.

(2) Bourin, M.-C.; Lindahl, U. Glycosaminoglycans and the Regulation of Blood Coagulation. *Biochem. J.* **1993**, *289*, 313.

(3) Gandhi, N. S.; Mancera, R. L. The Structure of Glycosaminoglycans and their Interactions with Proteins. *Chem. Biol. Drug Des.* **2008**, *72*, 455−482.

(4) Pepi, L. E.; Sanderson, P.; Stickney, M.; Amster, I. J. Developments in Mass Spectrometry for Glycosaminoglycan Analysis: A Review. *Mol. Cell. Proteomics* **2021**, *20*, 100025.

(5) Grabarics, M.; Lettow, M.; Kirschbaum, C.; Greis, K.; Manz, C.; Pagel, K. Mass Spectrometry-Based Techniques to Elucidate the Sugar Code. *Chem. Rev.* **2022**, *122*, 7840−7908.

(6) Zappe, A.; Miller, R. L.; Struwe, W. B.; Pagel, K. State-Of-The-Art Glycosaminoglycan Characterization. *Mass Spectrom. Rev.* **2022**, *41*, 1040−1071.

(7) Zamfir, A.; Seidler, D. G.; Kresse, H.; Peter-Katalinić, J. Structural Characterization of Chondroitin/Dermatan Sulfate Oligosaccharides from Bovine Aorta by Capillary Electrophoresis and Electrospray Ionization Quadrupole Time-of-Flight Tandem Mass Spectrometry. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 2015−2024.

(8) Gill, V. L.; Aich, U.; Rao, S.; Pohl, C.; Zaia, J. Disaccharide Analysis of Glycosaminoglycans using Hydrophilic Interaction

Chromatography and Mass Spectrometry. *Anal. Chem.* **2013**, *85*, 1138−1145.
(9) Miller, R. L.; Guimond, S. E.; Schworer, R.; Zubkova, O. V.; Tyler, P. C.; Xu, Y.; Liu, J.; Chopra, P.; Boons, G.-J.; Grabarics, M.; Manz, C.; Hofmann, J.; Karlsson, N. G.; Turnbull, J. E.; Struwe, W. B.; Pagel, K. Shotgun Ion Mobility Mass Spectrometry Sequencing of Heparan Sulfate Saccharides. *Nat. Commun.* **2020**, *11*, 1−12.
(10) Manz, C.; Pagel, K. Glycan Analysis by Ion Mobility-Mass Spectrometry and Gas-Phase Spectroscopy. *Curr. Opin. Chem. Biol.* **2018**, *42*, 16−24.
(11) Masellis, C.; Khanal, N.; Kamrath, M. Z.; Clemmer, D. E.; Rizzo, T. R. Cryogenic Vibrational Spectroscopy Provides Unique Fingerprints for Glycan Identification. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 2217−2222.
(12) Greis, K.; Kirschbaum, C.; von Helden, G.; Pagel, K. Gas-phase Infrared Spectroscopy of Glycans and Glycoconjugates. *Curr. Opin. Struct. Biol.* **2022**, *72*, 194−202.
(13) Mucha, E.; González Flórez, A. I.; Marianski, M.; Thomas, D. A.; Hoffmann, W.; Struwe, W. B.; Hahm, H. S.; Gewinner, S.; Schöllkopf, W.; Seeberger, P. H.; et al. Glycan Fingerprinting via Cold-Ion Infrared Spectroscopy. *Angew. Chem., Int. Ed.* **2017**, *56*, 11248−11251.
(14) Lettow, M.; Grabarics, M.; Greis, K.; Mucha, E.; Thomas, D. A.; Chopra, P.; Boons, G.-J.; Karlsson, R.; Turnbull, J. E.; Meijer, G.; et al. Cryogenic Infrared Spectroscopy Reveals Structural Modularity in the Vibrational Fingerprints of Heparan Sulfate Diastereomers. *Anal. Chem.* **2020**, *92*, 10228−10232.
(15) Lettow, M.; Greis, K.; Grabarics, M.; Horlebein, J.; Miller, R. L.; Meijer, G.; von Helden, G.; Pagel, K. Chondroitin Sulfate Disaccharides in the Gas Phase: Differentiation and Conformational Constraints. *J. Phys. Chem. A* **2021**, *125*, 4373−4379.
(16) Bierau, F.; Kupser, P.; Meijer, G.; von Helden, G. Catching proteins in Liquid Helium Droplets. *Phys. Rev. Lett.* **2010**, *105*, 133402.
(17) Thomas, D. A.; Chang, R.; Mucha, E.; Lettow, M.; Greis, K.; Gewinner, S.; Schöllkopf, W.; Meijer, G.; von Helden, G. Probing the Conformational Landscape and Thermochemistry of DNA Dinucleotide Anions via Helium Nanodroplet Infrared Action Spectroscopy. *Phys. Chem. Chem. Phys.* **2020**, *22*, 18400−18413.
(18) Janet, J. P.; Kulik, H. J. *Machine Learning in Chemistry*; American Chemical Society: Washington, DC, United States, 2020.
(19) Moon, S.; Chatterjee, S.; Seeberger, P. H.; Gilmore, K. Predicting Glycosylation Stereoselectivity using. *Machine Learning. Chem. Sci.* **2021**, *12*, 2931−2939.
(20) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(21) Menze, B. H.; Petrich, W.; Hamprecht, F. A. Multivariate Feature Selection and Hierarchical Classification for Infrared Spectroscopy: Serum-Based Detection of Bovine Spongiform Encephalopathy. *Anal. Bioanal. Chem.* **2007**, *387*, 1801−1807.
(22) Smith, B. R.; Ashton, K. M.; Brodbelt, A.; Dawson, T.; Jenkinson, M. D.; Hunt, N. T.; Palmer, D. S.; Baker, M. J. Combining Random Forest and 2D Correlation Analysis to Identify Serum Spectral Signatures for Neuro-Oncology. *Analyst* **2016**, *141*, 3668−3678.
(23) Zivanovic, V.; Seifert, S.; Drescher, D.; Schrade, P.; Werner, S.; Guttmann, P.; Szekeres, G. P.; Bachmann, S.; Schneider, G.; Arenz, C.; et al. Optical Nanosensing of Lipid Accumulation due to Enzyme Inhibition in Live Cells. *ACS Nano* **2019**, *13*, 9363−9375.
(24) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(25) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Routledge: Oxfordshire, England, United Kingdom, 2017.
(26) Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: Sebastopol, California, United States, 2019.
(27) Brieuc, M. S.; Waters, C. D.; Drinan, D. P.; Naish, K. A. A Practical Introduction to Random Forest for Genetic Association Studies in Ecology and Evolution. *Mol. Ecol. Resour.* **2018**, *18*, 755−766.
(28) Hua, J.; Xiong, Z.; Lowey, J.; Suh, E.; Dougherty, E. R. Optimal Number of Features as a Function of Sample Size for Various Classification Rules. *Bioinformatics* **2005**, *21*, 1509−1515.
(29) Menze, B. H.; Kelm, B. M.; Masuch, R.; Himmelreich, U.; Bachert, P.; Petrich, W.; Hamprecht, F. A. A Comparison of Random Forest and its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC Bioinform.* **2009**, *10*, 1−16.
(30) Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York City, New York, United States, 2013; Vol. *26*.
(31) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P.; Marth, J. D.; Bertozzi, C. R.; Hart, G. W.; Etzler, M. E. Symbol Nomenclature for Glycan Representation. *Proteomics* **2009**, *9*, 5398−5399.
(32) Schwörer, R.; Zubkova, O. V.; Turnbull, J. E.; Tyler, P. C. Synthesis of a Targeted Library of Heparan Sulfate Hexa-to Dodecasaccharides as Inhibitors of β-Secretase: Potential Therapeutics for Alzheimer's Disease. *Chem. - Eur. J.* **2013**, *19*, 6817−6823.
(33) Arungundram, S.; Al-Mafraji, K.; Asong, J.; Leach, F. E., III; Amster, I. J.; Venot, A.; Turnbull, J. E.; Boons, G.-J. Modular Synthesis of Heparan Sulfate Oligosaccharides for Structure- Activity Relationship Studies. *J. Am. Chem. Soc.* **2009**, *131*, 17394−17405.
(34) Gonzalez Florez, A. I.; Ahn, D.-S.; Gewinner, S.; Schollkopf, W.; von Helden, G. IR Spectroscopy of Protonated Leu-enkephalin and its 18-crown-6 complex Embedded in Helium Droplets. *Phys. Chem. Chem. Phys.* **2015**, *17*, 21902−21911.
(35) sklearn genetic opt *sklearn-genetic-opt User Guide*, 2022. https://sklearn-genetic-opt.readthedocs.io, accessed: 02/05/2022.
(36) Fortin, F.-A.; De Rainville, F.-M.; Gardner, M.-A. G.; Parizeau, M.; Gagné, C. DEAP: Evolutionary Algorithms Made Easy. *J. Mach. Learn. Res.* **2012**, *13*, 2171−2175.
(37) Eiben, A. E.; Smith, J. E. *Introduction to Evolutionary Computing*; Springer: New York City, New York, United States, 2015.
(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(39) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-Learn Documentation. 2023; https://scikit-learn.org/stable/modules/tree.html#tree-mathematical-formulation, accessed: 08/02/2023.
(40) Schindler, B.; Barnes, L.; Gray, C.; Chambert, S.; Flitsch, S.; Oomens, J.; Daniel, R.; Allouche, A.-R.; Compagnon, I. IRMPD spectroscopy Sheds New (Infrared) Light on the Sulfate Pattern of Carbohydrates. *J. Phys. Chem. A* **2017**, *121*, 2114−2120.
(41) Hansen, T.; Elferink, H.; van Hengst, J. M. A.; Houthuijs, K. J.; Remmerswaal, W. A.; Kromm, A.; Berden, G.; van der Vorm, S.; Rijs, A. M.; Overkleeft, H. S.; Filippov, D. V.; Rutjes, F. P. J. T.; van der Marel, G. A.; Martens, J.; Oomens, J.; Codee, J. D. C.; Boltje, T. J. Characterization of Glycosyl Dioxolenium Ions and their Role in Glycosylation Reactions. *Nat. Commun.* **2020**, *11*, 1−9.
(42) Bush, M. F.; Forbes, M. W.; Jockusch, R. A.; Oomens, J.; Polfer, N. C.; Saykally, R. J.; Williams, E. R. Infrared Spectroscopy of Cationized Lysine and ϵ-N-methyllysine in the Gas Phase: Effects of Alkali-Metal Ion Size and Proton Affinity on Zwitterion Stability. *J. Phys. Chem. A* **2007**, *111*, 7753−7760.
(43) Bush, M. F.; Oomens, J.; Williams, E. R. Proton Affinity and Zwitterion Stability: New Results from Infrared Spectroscopy and Theory of Cationized Lysine and Analogues in the Gas Phase. *J. Phys. Chem. A* **2009**, *113*, 431−438.
(44) Liaw, A.; Wiener, M. Classification and Regression by Random Forest. *R News* **2002**, *2*, 18−22.

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868

(45) Hastie, T.; Tibshirani, R.; Friedman, J. H.; Friedman, J. H. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*; Springer, 2009, Vol. 2.
(46) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral Deep Learning for Prediction and Prospective Validation of Functional Groups. *Chem. Sci.* **2020**, *11*, 4618−4630.
(47) Zhang, C.; Ma, Y. *Ensemble Machine Learning: Methods and Applications*; Springer: New York City, New York, United States, 2012.
(48) Bennett, L.; Melchers, B.; Proppe, B. *Curta: A General-Purpose High-Performance Computer at ZEDAT*; Freie Universität: Berlin, Germany, 2020.

Recommended by ACS

Solvent Model Benchmark for Molecular Dynamics of Glycosaminoglycans
Mateusz Marcisz and Sergey A. Samsonov
MARCH 29, 2023
JOURNAL OF CHEMICAL INFORMATION AND MODELING
READ

Substituted Oligosaccharides as Protein Mimics: Deep Learning Free Energy Landscapes
Benjamin Bouvier.
APRIL 11, 2023
JOURNAL OF CHEMICAL INFORMATION AND MODELING
READ

Glycan Stability and Flexibility: Thermodynamic and Kinetic Characterization of Nonconventional Hydrogen Bonding in Lewis Antigens
Jeahoo Kwon, Darón I. Freedberg, *et al.*
APRIL 26, 2023
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
READ

The Good, the Bad, and the Ugly Memories of Carbohydrate Fragments in Collision-Induced Dissociation
Chia Yen Liew, Chi-Kung Ni, *et al.*
SEPTEMBER 16, 2022
JOURNAL OF THE AMERICAN SOCIETY FOR MASS SPECTROMETRY
READ

Get More Suggestions >

https://doi.org/10.1021/jacs.2c12762
J. Am. Chem. Soc. 2023, 145, 7859−7868