Molecular dynamics-based descriptors of 3-O-Sulfated Heparan sulfate as contributors of protein binding specificity

Annemarie Danielsson [a], Małgorzata M. Kogut [a,b], Martyna Maszota-Zieleniak [a], Pradeep Chopra [c], Geert-Jan Boons [c,d,e], Sergey A. Samsonov [a,*]

[a] *Faculty of Chemistry, University of Gdańsk, ul. Wita Stwosza 63, 80-308 Gdańsk, Poland*
[b] *Intercollegiate Faculty of Biotechnology of University of Gdańsk and Medical University of Gdańsk, ul. Abrahama 58, 80-307 Gdańsk, Poland*
[c] *Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA*
[d] *Department of Chemistry, University of Georgia, Athens, GA 30602, USA*
[e] *Department of Chemical Biology and Drug Discovery, Utrecht Institute for Pharmaceutical Sciences and Bijvoet Center for Biomolecular Research, Utrecht University, 3584 CG Utrecht, the Netherlands*

ARTICLE INFO

ABSTRACT

Glycosaminoglycans are linear periodic and anionic polysaccharides found in the extracellular matrix, involved in a range of key biochemical processes as a result of their interactions with a variety of protein partners. Due to the template-less synthesis, high flexibility and charge of GAGs, as well as the multipose binding of GAG ligands to receptors, the specificity of GAG-protein interactions can be difficult to elucidate. In this study we propose a set of MD-based descriptors of unbound Heparan Sulfate hexasaccharides that can be used to characterize GAGs and explain their binding affinity to a set of protein receptors. With the help of experimental data on GAG-protein binding affinity, we were able to further characterize the nature of this interaction in addition to providing a basis for predictor functions of GAG-protein binding specificity.

1. Introduction

Glycosaminoglycans (GAGs) are a family of naturally-occurring linear periodic and anionic polysaccharides whose building blocks are repeating disaccharide units composed of an amino sugar and an uronic acid or hexose (Esko et al., 2009). Depending on their exact dimeric unit composition and their glycosidic linkages, several classes of GAGs can be distinguished: Heparan Sulfate (HS), Heparin (HP), Dermatan Sulfate (DS), Chondroitin Sulfate (CS), Keratan Sulfate (KS), and Hyaluronic Acid (HA). Heparan Sulfate (HS), composed of alternating *N*-actetyl-glucosamine (GlcNAc) and glucuronic acid (GlcA) residues (Esko et al., 2009), is found often covalently attached to proteins of the extracellular matrix and plasma membrane, forming proteoglycans (PGs). HS is known to be involved in key biological processes, including cell division (Ughy et al., 2019) and differentiation (Kraushaar et al., 2012; Patel et al., 2008; Yokoyama et al., 2020), angiogenesis (Zhang et al., 2014), coagulation (Ho et al., 1997), viral infection (Clausen et al., 2020; Yue

et al., 2021), neuron growth (Brickman et al., 1998; Johnson et al., 2007), as well as tumor proliferation (Hendriks et al., 2005) and metastasis (Qazi et al., 2016).

HS can undergo (selective) de-N-acetylation and N-sulfation of its GlcNAc residues, epimerization of GlcA to iduronic acid (IdoA), as well as sulfation by O-sulfotransferases on the 2-*O* position of its uronic acid (C2 of the uronic acid) and of the 6-*O* and 3-*O* positions of GlcNAc residues (Esko et al., 2009). While 3-*O*-sulfation is the rarest of all HS modifications, constituting only about 0.5% of the total sulfation (Huang et al., 2015; Pejler et al., 1987), 3-*O*-sulfotransferases (3-OSTs) represent the largest family of enzymes modifying HS polysaccharide chains (Thacker et al., 2014). This may indicate that 3-OSTs in cooperation with other HS-modifying enzymes produce HS molecules that are unique and therefore engage in highly-specific interactions with proteins (Chopra et al., 2021).

The specificity of GAG-protein binding has been studied extensively before for a variety of complexes (Gama et al., 2006; Gama and

---

Hsieh-Wilson, 2005; Joseph et al., 2015; Künze et al., 2021; Liu et al., 2002; Pichert et al., 2012; Rogers et al., 2011; Sankarayanarayanan et al., 2017; Schlorke et al., 2012). In some cases, the only parameter that affects the strength of such interactions is GAG charge, which makes them entirely electrostatically-driven and, therefore, unspecific (Hintze et al., 2014; Koehler et al., 2017; Panitz et al., 2016; Rother et al., 2016). However, a considerable number of GAG-protein complexes rely on specific interactions (Almond, 2018; Guerrini et al., 2008; Künze et al., 2021; Pomin and Mulloy, 2015; Sage et al., 2013). A number of proteins have been identified as receptors of 3-*O*-sulfated HS molecules participating in specific interactions, including Antithrombin III (Guerrini et al., 2008), Neuropilin 1 (Thacker et al., 2016), Stabilin 2 (Pempe et al., 2012), Advanced Glycosylation End-Product Receptor (Thacker et al., 2014) and growth factors such as Fibroblast Growth Factor 7 (Luo et al., 2006). Anomalies of 3-*O*-Sulfation have been implicated in tumorigenesis (Denys and Allain, 2019), renal fibrosis (Ferreras et al., 2019), and tauopathies relating to Alzheimer's Disease (Zhao et al., 2020). 3-*O*-suflated HS molecules on cell surfaces also participate in viral infections (O'Donnell and Shukla, 2008). Taken together, the characteristics of HS and their interactions with proteins, as well as involvement in essential biological pathways render them promising and interesting targets in medicine.

Although significant research has been conducted on this topic, the template-less synthesis and resulting structural complexity of HS molecules complicates studies of their interaction with proteins. While research on GAG-protein specificity concentrates mostly on the GAG-binding sites of proteins (Jokiranta et al., 2005; Li et al., 2016; Morgan et al., 2015; Mosier et al., 2012; Multhaup, 1994; Pratt and Church, 1992; Sarkar and Desai, 2015; Sun et al., 2001; Taylor et al., 1995; Witt and Lander, 1994), it has been also observed that the sequence (Irie et al., 2002; Raghuraman et al., 2010, 2006; Sankaranarayanan et al., 2015; Sankaranarayanan and Desai, 2014; Shworak et al., 1994), sulfation degree and pattern (Ashikari-Hada et al., 2009; Irie et al., 2002; Kinnunen et al., 1996; Nakato and Kimata, 2002; Patel et al., 2008; Pye et al., 2000; Stringer and Gallagher, 1997; Viviano et al., 2004), chain length (Patel et al., 2008; Viviano et al., 2004) and, to a lesser degree, the conformation (Guglier et al., 2008) of GAGs generally and HS specifically can contribute to binding specificity.

The aim of this study was to inspect and analyze the specificity of HS-protein interactions employing computational approaches. To this end, we applied molecular dynamics (MD) simulations as well as a standard Machine Learning Algorithm (MLA) to examine 27 3-, 6-, and 3,6-*O*-sulfated HS molecules characterized and studied by (Chopra et al., 2021), followed by linear regression and cluster analysis in order to assess the connection between characteristics of unbound HS molecules and the specificity of binding to their protein partners. The analyzed dataset of 27 HS molecules with clearly defined sequences and sulfation patterns presents an unique opportunity to gain insight regarding the influence of specific differences in sulfation modifications on the 3D structure and physico-chemical characteristics of these molecules during Molecular Dynamics simulations. Those, in turn, can be linked to the affinity between the analyzed HS molecules and a set of proteins known to bind HS with different specificity. The application of a simple and well-known unsupervised MLA, Principal Component Analysis (PCA), enabled us to characterize the unbound HS molecules individually, but also the interplay between the proposed molecular descriptors in the context of GAG-protein binding. The investigation of the behavior of unbound HS molecules during MD simulations at atomistic level, that is not accessible using most experimental approaches, was supported and complemented by information gained from experiments (Chopra et al., 2021). Taken together, both approaches helped to overcome the limitations of either of them when applied alone and allowed us to identify characteristics of unbound HS molecules that could be linked with confidence to differences in binding affinity for a set of HS-protein complexes. This knowledge may prove useful in drug design of HS molecules exhibiting traits in line with our findings. In particular it

could suggest modifications of the GAG sequence that would lead to the appropriate conformations of the unbound HS with a higher propensity of binding their protein partners. Hence, our data assist in enabling the design and synthesis of specific HS molecules for potential use in regenerative medicine, cancer treatment and the prevention and treatment of viral infections.

2. Materials and methods

2.1. Heparan sulfate and protein dataset

In our study, we considered the library of 27 HS hexasaccharides previously synthesized and analyzed by (Chopra et al., 2021). The synthetic HS molecules, constructed from 9 different backbone templates, differed in the modification of their central GlcNAc residue (3-*O*-Sulfation, 6-*O*-Sulfation, 3,6-*O*-Sulfation) (sequences detailed in Table 1). In order to conduct simulations via computational methods, the 27 HS molecules were built in the xLeap module of AMBER16 (Case et al., 2016) and described with the GLYCAM06 force field parameters (Kirschner et al., 2008). The two building blocks of the analyzed HS hexasaccharides are shown in Fig. 1. The exact placement of the sulfate groups along the GAG influences the strength of electrostatic interaction in GAG-protein complexes and represents a "sulfation code" assumed to regulate the specificity of interactions with proteins (Gama and Hsieh-Wilson, 2005).

The analyzed HS molecules were named 1A, 1B, 1C to 9A, 9B, 9C in accordance with the naming scheme in (Chopra et al., 2021). The names, detailed in Table 1, reflect the position of the sulfate group on the GlcNAc residue ("A": 3,6-*O*-sulfation, "B": 3-*O*-sulfation, "C": 6-*O*-Sulfation) and the sequence of the 3 residues varying between the 27 oligosaccharides.

Binding affinity data between the 27 HS hexasaccharides and a set of nine functionally diverse proteins known to recognize 3-*O*-sulfation to varying degrees was taken from (Chopra et al., 2021). The nine proteins considered are: Antithrombin III (ATIII), Heparin Cofactor 2 (HC-II), Fibroblast Growth Factor 7 and 9 (FGF-7, FGF-9), Fibroblast Growth Factor Receptor I (FGFR-I), Neuropilin 1 (Nrp-1), Bone Morphogenic Protein 2 (BMP-2), Stabilin 2 (Stab-2), Advanced Glycosylation End-Product Receptor (RAGE). The interaction study between HS molecules and the nine proteins presented in (Chopra et al., 2021) is the first such study examining the interaction selectivity for the set of HS hexasaccharides of clearly defined sequences and specific sulfation patterns along the GAG. All nine proteins are known to interact with HS and are involved in key biochemical processes connected to cell-cell signalling, e.g. angiogenesis, immunity, axon guidance, embryonic development. The study by (Chopra et al., 2021) is thus a particularly exclusive dataset that can be studied in order to elucidate the variability in binding affinity and selectivity of specific HS molecules and a wide range of their protein partners.

2.2. Molecular dynamics

Molecular dynamics (MD) simulations of the 27 unbound HS hexaseccharides were carried out in AMBER16 (Case et al., 2016; Götz et al., 2012) with the GLYCAM06 (Kirschner et al., 2008) force field in order to obtain descriptors of the oligosacchrides with possible predictive power in relation to their binding affinity and specificity.

Each of the 27 molecules was solvated in a TIP3P octahedral periodic box with minimum distance between solute and box edge of 15.0 Å and neutralized with counterions (Na$^+$). For every hexasaccharide, two energy minimization steps were carried out (first 1.5×10^3 steepest descent cycles and 10^3 conjugate gradient cycles with harmonic force restraints on solute atoms, followed by 6×10^3 steepest descent cycles and 3×10^3 conjugate gradient cycles without restraints). Subsequently, the system was heated up to 300 K for 10 ps with harmonic force restraints of 100 kcal/mol Å^{-2} on solute atoms, and equilibration for 50 ps

Table 1
Sequences of the 27 Heparan Sulfate molecules; differences in sequence between the HS types are marked in bold.

|		
1A: GlcA–GlcNS6S–**GlcA–GlcNS3S6S–GlcA**–GlcNS6S	4A: GlcA–GlcNS6S–**GlcA–GlcNS3S6S–IdoA**–GlcNS6S	7A: GlcA–GlcNS6S–**GlcA–GlcNS3S6S–IdoA2S**–GlcNS6S
1B: GlcA–GlcNS6S–**GlcA–GlcNS3S–GlcA**–GlcNS6S	4B: GlcA–GlcNS6S–**GlcA–GlcNS3S–IdoA**–GlcNS6S	7B: GlcA–GlcNS6S–**GlcA–GlcNS3S–IdoA2S**–GlcNS6S
1C: GlcA–GlcNS6S–**GlcA–GlcNS6S–GlcA**–GlcNS6S	4C: GlcA–GlcNS6S–**GlcA–GlcNS6S–IdoA**–GlcNS6S	7C: GlcA–GlcNS6S–**GlcA–GlcNS6S–IdoA2S**–GlcNS6S
2A: GlcA–GlcNS6S–**IdoA–GlcNS3S6S– GlcA**–GlcNS6S	5A: GlcA–GlcNS6S–**IdoA–GlcNS3S6S–IdoA**–GlcNS6S	8A: GlcA–GlcNS6S–**IdoA–GlcNS3S6S–IdoA2S**–GlcNS6S
2B: GlcA–GlcNS6S–**IdoA–GlcNS3S–GlcA**–GlcNS6S	5B: GlcA–GlcNS6S–**IdoA–GlcNS3S–IdoA**–GlcNS6S	8B: GlcA–GlcNS6S–**IdoA–GlcNS3S–IdoA2S**–GlcNS6S
2C: GlcA–GlcNS6S–**IdoA–GlcNS6S–GlcA**–GlcNS6S	5C: GlcA–GlcNS6S–**IdoA–GlcNS6S–IdoA**–GlcNS6S	8C: GlcA–GlcNS6S–**IdoA–GlcNS6S–IdoA2S**–GlcNS6S
3A: GlcA–GlcNS6S–**IdoA2S–GlcNS3S6S–GlcA**–GlcNS6S	6A: GlcA–GlcNS6S–**IdoA2S–GlcNS3S6S–IdoA**–GlcNS6S	9A: GlcA–GlcNS6S–**IdoA2S–GlcNS3S6S–IdoA2S**–GlcNS6S
3B: GlcA–GlcNS6S–**IdoA2S–GlcNS3S–GlcA**–GlcNS6S	6B: GlcA–GlcNS6S–**IdoA2S–GlcNS3S–IdoA**–GlcNS6S	9B: GlcA–GlcNS6S–**IdoA2S–GlcNS3S–IdoA2S**–GlcNS6S
3C: GlcA–GlcNS6S–**IdoA2S–GlcNS6S–GlcA**–GlcNS6S	6C: GlcA–GlcNS6S–**IdoA2S–GlcNS6S–IdoA**–GlcNS6S	9C: GlcA–GlcNS6S–**IdoA2S–GlcNS6S–IdoA2S**–GlcNS6S
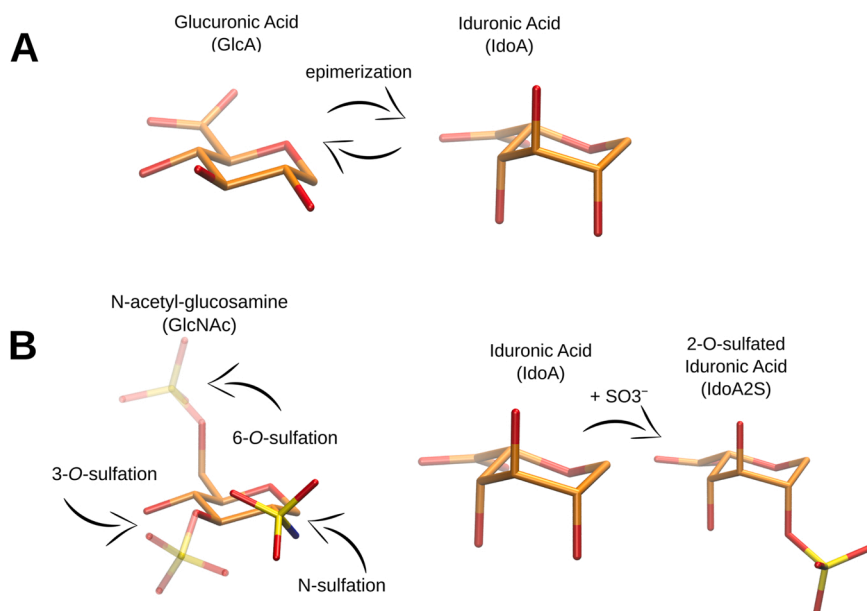
Fig. 1. The possible modifications of the HS building blocks in this study. All the HS molecules are made up of repetitive disaccharide units containing an uronic acid (glucuronic acid, GlcA or iduronic acid, IdoA) and N-acetyl-glucosamine, GlcNAc derivative. A) Glucuronic acid can undergo reversible epimerization to iduronic acid. B) Possible sulfation modifications of N-acetyl-glucosamine on its 3-*O* and 6-*O* positions (semi-transparent licorice representation) as well as the 2-*O* sulfation position of iduronic acid.

at 300 K and 10^5 Pa in isothermal isobaric ensemble (NPT). A 100 ns productive MD run was carried out in an NPT ensemble. The SHAKE algorithm (Ryckaert et al., 1977), 2 fs time integration step, 8 Å cutoff for non-bonded interactions, and the Particle Mesh Ewald method (Darden et al., 1993) were used.

2.3. Descriptors of unbound HS molecules

Based on the MD trajectories of each unbound HS, a set of physico-chemical properties ("descriptors") was analyzed in the PTRAJ (Roe and Cheatham, 2013) module of AMBER16 using default parameters (Table 2). Restricting the descriptors to only those relating to intra-molecular characteristics of the HS was done with the goal to discover what properties of the unbound molecules are putatively linked to the binding affinity. In case those descriptors are identified, one can argue that while the synthesis of HS molecules in vivo is template-less, the cellular machinery is required to ensure certain characteristics and modifications of HS molecules in vivo in order for the molecules to be able to interact specifically with their protein partners. The inclusion of other descriptors to be obtained from MD simulation of HS complexes with proteins as well as descriptors of the HS-binding sites on the proteins should also be considered, as these would likely enhance the information gained from computational studies of these unbound HS oligosaccharides. Nevertheless, it can be assumed that the characteristics of unbound HS molecules are likely to affect the binding affinity to proteins in the extracellular matrix, as the long GAGs naturally occurring in cells would possibly have substantially limited flexibility in vivo to ensure correct complex formation with proteins (Sattelle et al., 2015; Spencer et al., 2010). Therefore, it is likely that certain characteristics of HS have to be already encoded and ensured in their unbound state, thus

likely to be observed in the preference of certain values of the physico-chemical descriptors analyzed in this study.

To detect groups of highly correlated descriptors, Pearson product-moment correlation coefficients were computed for the descriptor dataset and plotted as a heatmap. The correlation analysis was used as a guide in feature selection.

2.4. Machine learning and correlation analysis

The descriptor dataset was standardized prior to applying MLAs. The feature extraction and selection approaches were used to reduce the dimensionality of the original dataset and to uncover patterns in the descriptor dataset that could later be linked to the binding affinity of HS-protein complexes.

PCA of the descriptor dataset of unbound HS molecules was carried out in python 3.8.5 using the scikit-learn library (ver. 0.24.2) (Pedregosa et al., 2011). PCA is a type of unsupervised machine learning technique used in a variety of fields to reduce the dimensionality of large datasets as well as to reduce noise and extract patterns from the data. The approach works by transforming a matrix of correlated variables into a new coordinate system of uncorrelated variables called Principal Components (PCs) which are linear combinations of the initial ones and capture the essential information from the original dataset. We applied PCA to the descriptor dataset to create a lower-dimensional dataset of PCs that explain most of the variability in the MD-derived descriptors. Each PC can be seen as a summary of a set of variables that captures the physico-chemical character shared by the descriptors that contribute the most to the given PC. Because the construction of the PCs did not include any information on the HS-protein binding affinity, the obtained PCs only describe the character of the unbound HS oligosaccharides

Table 2
MD-derived descriptors of unbound HS molecules used in the study.

No.	Descriptor	Symbol (Units)	Explanation
1.–8.	Fraction of formed H-bonds	Hbond[0/1/2/3/4/5/6/7] (no unit)	Sum of fraction of MD-simulation frames where 0/1/2/3/4/5/6/7 intramolecular H-bonds were formed by atoms of the HS molecule
9.	Radius of gyration	R_gyration (Å)	A measure of elastic stability (resistance to deformation of shape and conformation) of the HS molecule; compactness of the molecule
10.	Dipole moment	Dipole (e * Å)	The dipole moment: a measure of the polarity of a molecule
11.	Molecule length	Length (Å)	Length of the molecule in Å defined as the distance between two terminal atoms
12.–19.	Dihedral angles of glycosidic linkages	glycosidic_[1min1/1min2/3min1/ 3min2/5min1/5min1/2 min/4 min] (no unit)	The distribution of dihedral angles of each glycosidic linkages defined as $O5_{n+1}$-$C1_{n+1}$-$O4_n$-$C4_n$ and $C1_{n+1}$-$O4_n$-$C4_n$-$C5_n$, where n is the sequential number of the sugar monomeric unit; Specifically, the percentage of points in the distribution of the dihedral angles that belong to a minimum
20.		glycosidic_percent (%)	Total percentage of points of the dihedral angle distribution that belong to minima
21.–26.	Fluctuation of HS monosaccharide units	fluct_[r1/r2/r3/r4/r5/r5] (Å)	Root mean square fluctuation analysis of atoms of HS residues
27.–28.	Ring pucker	pucker_[1C4/4C1] (no unit)	Fraction of $^1C_4$ and $^4C_1$ ring pucker conformations of the sugar monomeric units; The conformations are defined as: $^1C_4$: $\gamma = 48° \pm 30°$, $\delta = -64° \pm 30°$, $^4C_1$: $\gamma = -54° \pm 30°$, $\delta = 62° \pm 30°$, where angle γ is defined by atoms C1–C2–C3–C4, and angle δ by atoms C1–O5–C5–C4
29.	Free energy components from the MM-GBSA model	VDW (kcal/mol)	Van der Waals energy
30.	Electrostatic energy	EEL (kcal/mol)	Electrostatic energy in vacuo
31.	EGB	EGB (kcal/mol)	Generalized Born electrostatic solvation energy
32.	ESURF	ESURF (kcal/mol)	Non-polar solvation energy
33.	Total free energy	TOTAL (kcal/mol)	Total free energy
34.	QH Entropy	S_APPROX (kcal/mol)	Quasi-Harmonic configurational entropy
35.	NM Entropy	S_nm (kcal/mol)	Normal mode entropy

simulated using MD.

As an alternative approach, Pearson correlation was used to assess the relationship between each individual descriptor and the HS-protein binding affinity. The values of each descriptor measured during the MD simulation were correlated with the binding affinity values for each complex individually to investigate which descriptors can be assumed to be most strongly linked with HS-protein binding strength and whether differences can be observed between complexes that can be attributed to the character and specificity of the binding. A descriptor was considered to be significantly associated with the binding affinity data if the p-value of the *t*-test for the correlation was below the $\alpha = 0.05$ significance level.

2.5. Linear regression analysis

Linear regression analysis was carried out in R (*stats* base package) with the purpose of determining the strength of the link between the computed PCs and HS-protein binding affinity data as given in (Chopra et al., 2021). To evaluate the usefulness of the PCA-based approach, for each HS-binding protein, the first four PCs of the dataset were used in a matching linear regression analysis.

Statistical significance and comparisons of regression models were conducted using the *t*-test and F-test. The cutoff for statistical significance used equals $\alpha = 0.05$; p-values of the *t*-test for the independent variables (PCs) below the α level were considered statistically significant. The F-tests were used to assess how well the regression models fit the data. R^2 (coefficient of determination) values were reported for each linear regression model to describe the amount of variance in the target variable (HS-binding affinity for each protein) accounted for by the independent variables.

2.6. Statistical and cluster analysis

Statistical analyses were carried out in R version 3.6.3 (R Core Team, 2020) (descriptive statistics of the descriptor dataset, Pearson correlation calculations using the "psych" package version 2.1.9 (Revelle, 2021)) as well as python 3.8.5 using numpy 1.19.5 (Harris et al., 2020) and pandas 1.2.2 (McKinney et al., 2010). Cluster analysis via

Hierarchical Clustering was carried out for the Pearson correlation coefficients between PCs ("scoring function") and the binding affinity data in python using seaborn 0.11.1 (Waskom, 2021).

2.7. Visualization

Visualization was carried out in python using matplotlib 3.3.4 (Hunter, 2007) and seaborn 0.11.1 (Waskom, 2021).

3. Results

The descriptors of unbound HS hexasaccharides analyzed during the MD simulation are summarized in Supplementary Table 1.

3.1. MD simulation analysis

The properties of the 27 HS molecules throughout the MD simulation were analyzed and summarized in terms of the major descriptors. The evolution over time for Root Mean Square Deviation (RMSD) with respect to the starting conformation, radius of gyration and molecule length (end-to-end distance) are shown in Fig. 2 for molecules 1A, 1B, 1C, which were chosen as representatives for the analysis of these descriptors for two reasons. First, all the three different analyzed modes of sulfation (3,6-*O*-, 3-*O*-, and 6-*O*-sulfation) can be found in 1A, 1B and 1C, respectively. Moreover, the majority of the analyzed hexasaccharides behaved throughout the MD simulation in a manner qualitatively indistinguishable from these three HS molecules. Therefore, the evaluation whether the MD simulation reached convergence and the analysis of the descriptors over time is described in greater detail for 1A, 1B and 1C, as well as for any of the other molecules if their behavior over time during the MD was clearly different than for the majority of analyzed HS hexasaccharides. Supplementary Figs. 1–3 show the behavior over time for RMSD, radius of gyration, and end-to-end distance for all 27 simulated HS hexasaccharides. Upon inspecting the changes of descriptors over time for the 27 HS (Fig. 2 and Supplementary Figs. 1–3), it can be seen that the MD simulation converged, i.e. the simulated system reached its equilibrium state.
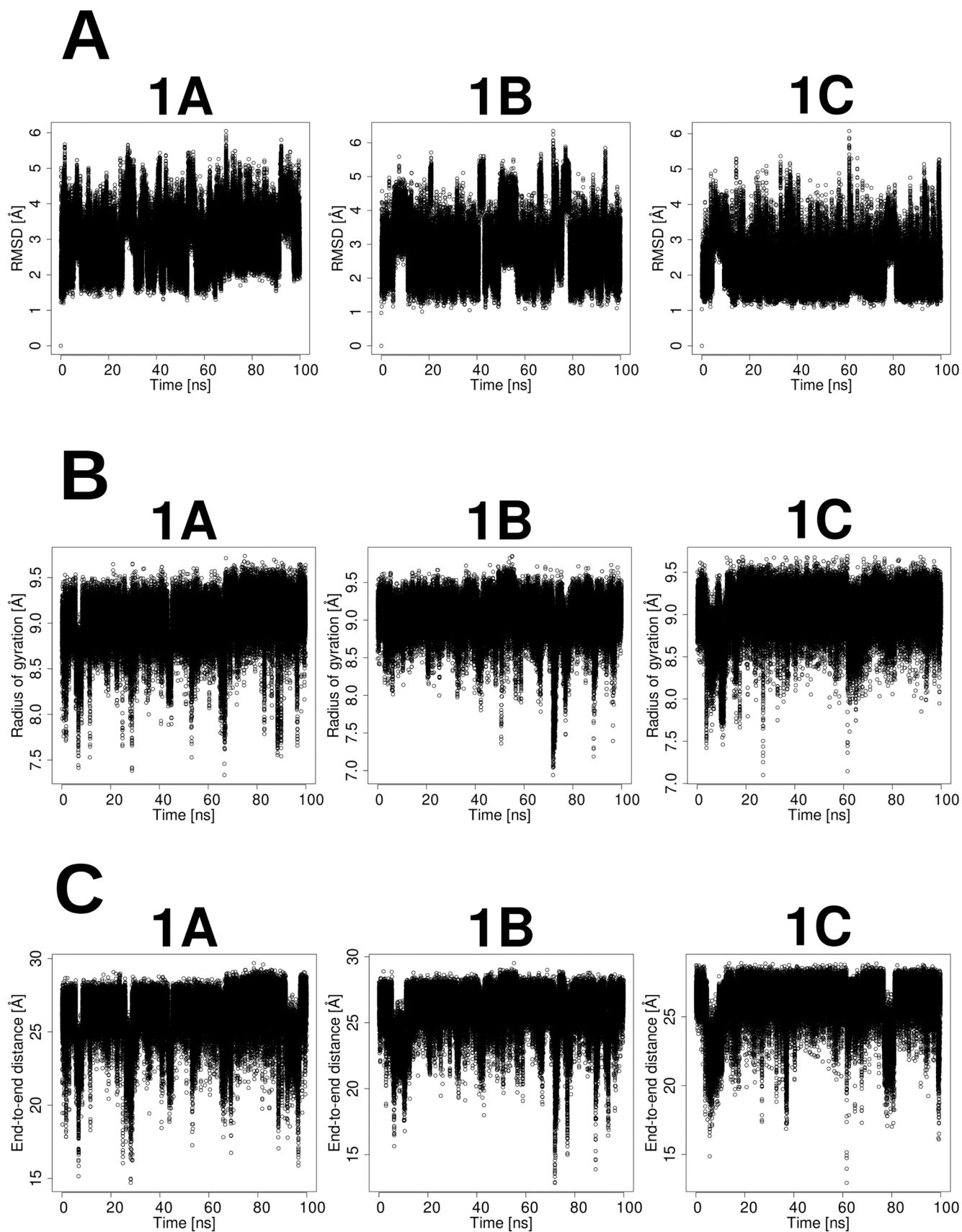
Fig. 2. A) Root Mean Square Deviation (RMSD) [Å] of HS molecules 1A, 1B, 1C calculated in reference to the starting position of the simulation. B) Radius of gyration [Å] of HS molecules 1A, 1B, 1C. C) Length of the simulated hexasaccharide, calculated as the end-to-end distance [Å], of HS molecules 1A, 1B, 1C. All the data are obtained from 100 ns of the MD simulation.

3.1.1. Root mean square deviation

RMSD analysis indicated that most of the HS molecules exhibit large fluctuations in the RMSD value across the MD simulation, as can be seen for molecules 1A, 1B, 1C in Fig. 2, chosen as representatives for the HS dataset. This indicates a high mobility of the unbound molecules. Exceptions can be observed for molecules 7B and 9A (Supplementary Fig. 1), which reach a stable state at a relatively early point of the MD simulation, as well as for 7C, for which RMSD reaches convergence towards the end of the simulation.

3.1.2. Radius of gyration

An analysis of the radius of gyration (R_gyration) of the 27 HS molecules shows trends in line with those observed for RMSD. As can be seen in the example of HS 1A, 1B, 1C (Fig. 2), the hexasaccharides are very flexible showing a high degree of mobility. As with RMSD analysis, molecule 9A, and to a lesser degree 7B and 7C (Supplementary Fig. 2) behave differently than the other molecules by exhibiting a decrease in R_gyration and thereby achieving a more stable and compact conformation towards the end of the MD simulation. 6C also exhibited a low radius of gyration for a significant portion of the MD simulation, but towards the end the values of the radius of gyration increased again, signifying a switching from an extended conformation in the beginning of the MD, to a more compact conformation in the middle of the simulation, and back to a more extended and flexible conformation towards the end.

3.1.3. HS length

The hexasaccharide molecules exhibit rapid changes of molecule length (measured as end-to-end distance) throughout the MD simulation, with most molecules not converging to either an extended nor a compact state at the end of the simulation, as can be seen in Fig. 2 for the representative molecules 1A, 1B, 1C. However, HS 9A is an exception in the dataset, converging to a clearly compact conformation, with a significant decrease in the end-to-end distance seen throughout the simulation (Supplementary Fig. 3). Similarly, molecules 7C and 9C are seen to shorten towards the very end of the MD simulation.

3.1.4. Intramolecular H-bonds

Fig. 3 shows the distribution of frequencies of intramolecular hydrogen bonds (H-bonds) for the 27 HS molecules. For the vast majority of the analyzed molecules, the median value of H-bonds formed during the MD simulation equals 3 H-bonds, with some exceptions (median of 2 H-bonds for 7A, 8A, 9A). Conformations in which either no intramolecular H-bonds are formed during the MD simulation as well as those with the maximum possible amount of intramolecular H-bonds are unlikely to observe for all of the analyzed molecules. Nevertheless, the 3,6-sulfated 9A, 8A, and 7A hexasaccharides seem to have a preference of conformations with less intramolecular H-bonds compared to the other HS molecules.

3.1.5. Dipole moment

All of the molecules seem to be similar in terms of polarity, with values of the dipole moment measured across the MD simulation ranging from 420.6 e*Å (for 4B) to 615.6 e*Å (for 9C) (Supplementary Table 1). The most polar molecules are 9C (615.8 e*Å) and 9A (600.0 e*Å), substantially separated from the third-most polar HS hexasaccharide (3A with 562.4 e*Å) by over 30 e*Å units.
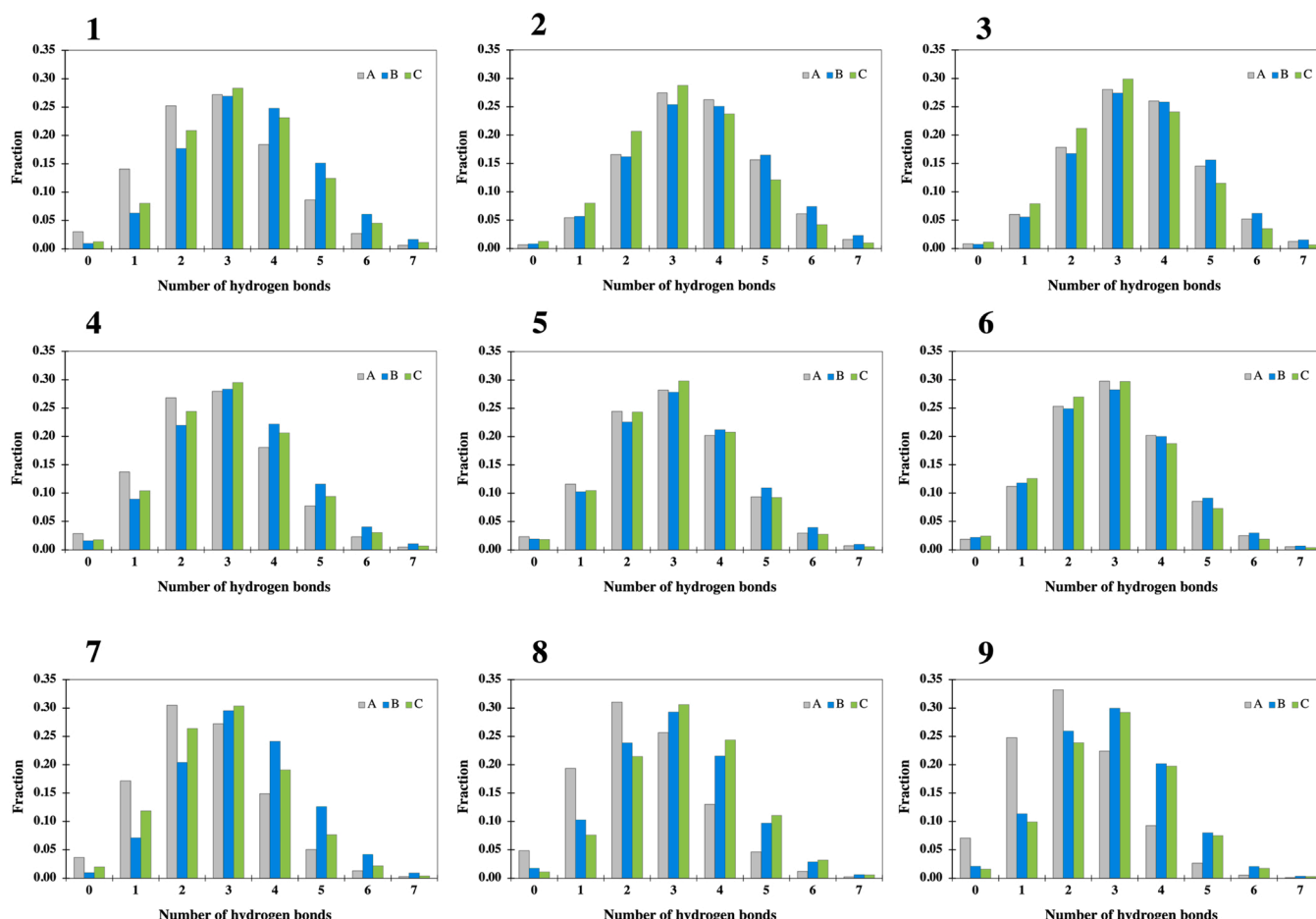

Fig. 3. Distribution of frequencies of intramolecular H-bonds formed by the HS molecules during MD simulations.

3.1.6. Glycosidic linkage dihedral angles

In terms of dihedral angles of the glycosidic linkages (Supplementary Table 1), the HS molecules do not seem to display significant differences between each other. The percentage of observations belonging to minima of the glycosidic linkage dihedral angles are comparable between the analyzed hexasaccharides.

3.1.7. Sugar ring fluctuation

For all of the analyzed HS molecules, the highest values of sugar ring fluctuations throughout the MD simulation were observed for the sugar ring of the first residue (Supplementary Table 1) and to a much lesser degree for the sugar ring of the last residue. Hexasaccharides 6C, 7C and 9A exhibit the highest fluctuation of their first sugar ring (8.8 Å, 8.3 Å and 8.8 Å, respectively, versus a range of 4.7–6.8 Å for the remaining HS molecules of the dataset). The same trend can be observed for the last sugar residue (fluctuation values of 4.6 Å for 6C, 3.8 Å for 7C and 5.5 Å for 9A, compared to a range of 2.0–3.6 Å for the rest of the hexasaccharides), but not for the 2nd, 3rd, 4th and 5th sugar rings.

3.1.8. Ring pucker conformation

A strong preference of the $^4C_1$ sugar ring pucker conformation can be seen for all 27 HS hexasaccharides (Supplementary Table 1). Among the 27 molecules, sugar rings of 1A, 6A and 7B show the highest propensity for the $^1C_4$ conformation, with a fraction of MD time in which such a conformation could be observed equal to 0.1, 0.19 and 0.11, respectively.

3.1.9. Enthalpy and entropy

In terms of enthalpy and entropy components of the unbound HS molecules throughout the MD simulation, most of the analyzed hexasaccharides have comparable values (Supplementary Table 1). A stark difference can be only observed for the electrostatic energy in vacuo for sugar 9A in comparison to the other HS molecules (1568 kcal/mol, versus a range of 356 kcal/mol to 1065 kcal/mol for the remaining sugars), indicating highly unfavorable electrostatic energy values of the unbound 9A hexasaccharide in solvent environment.

In summary, the 3,6-*O*-Sulfated HS hexasaccharide 9A, where the central sugar residue is flanked by IdoA2S residues on both sides, shows behavior clearly different than the other sugar hexasaccharides throughout the MD simulation, especially in terms of its RMSD calculated in reference to the starting conformation, frequency of intramolecular H-bonds, radius of gyration, fluctuations of the first and sixth sugar rings and electrostatic energy in vacuo. Noteworthy, sugar 9A was found to bind with high affinity to all proteins except for ATIII (Chopra et al., 2021).

3.2. Descriptor analysis

Pairwise Pearson correlation coefficients were calculated for the 35 descriptors summarized in Table 2, in order to examine the relationships between the descriptors. The correlation matrix visualized as a heatmap (Fig. 4) shows that some descriptors form clusters of strong positive and negative correlation, which may point to mutual relationships/dependencies and therefore possible redundancy in the primary descriptor dataset. A strong positive correlation (high positive Pearson correlation coefficient) is depicted by a red color, a strong negative correlation (negative Pearson correlation coefficient) – by violet, while no correlation corresponds to green colors on Fig. 4.

H-bond–related descriptors form three clusters based on the correlation coefficients: low amount of or no intramolecular H-bonds (Hbond0, Hbond1, Hbond2), high amount of intramolecular H-bonds (Hbond4, Hbond5, Hbond6, Hbond7) and exactly 3 H-bonds formed during the MD simulation (Hbond3). These descriptors correspond to the fraction of MD trajectory frames in which the simulated HS molecule formed the given amount of H-bonds. Therefore, the distribution of those frequencies, shown in Fig. 3, hints at preferences of certain HS

types regarding the formation of intramolecular H-bonds. While the most frequent amount of H-bonds for the majority of HS molecules was 3 H-bonds, hexasaccharides 7A, 8A and 9A are characterized by right-skewed distributions with the mode equal to 2 intramolecular H-bonds. The differences in sequence and sulfation pattern between the HS molecules dictates the preference for H-bond formation within the molecule. This information can in turn shed light on the formation of intermolecular H-bonds between the given HS and proteins, because the preference for forming many intramolecular H-bonds may indicate also a high likelihood of forming intermolecular H-bonds upon binding to proteins.

The closely related descriptors R_gyration and Length, corresponding to the mean radius of gyration of the HS molecule and its end-to-end distance, respectively, show a strong positive correlation to each other. The more compact a molecule is and thus the smaller its radius of gyration, the smaller is its end-to-end distance. Conversely, an extended molecule will display a higher radius of gyration and a greater length. Interestingly, among all other descriptors, both R_gyration and Length are strongly negatively correlated to glycosidic_3min2. A stiff hexasaccharide structure, i.e. only a narrow range of values possible for glycosidic bond dihedral angles, would influence the level of compactness of a molecule. The negative correlation coefficient value therefore links a preference of lower dihedral angle values with a less compact structure. Among the 27 HS hexasaccharides, 6C, 7B and 9A display a more compact structure during the MD simulations (lower radius of gyration and shorter end-to-end distance) compared to the other oligosaccharides.

All descriptors of sugar ring fluctuation are positively correlated to each other, however the fluctuation of the second sugar ring (fluct_r2) only shows a strong correlation to the first sugar ring and a weak correlation to the other fluctuation descriptors. For all studied HS molecules, the first sugar ring shows the highest fluctuations, with the 2nd, 3rd, 4th, 5th and 6th rings being considerably less mobile throughout the MD simulation. Molecules 6C, 7B, 7C and 9A have the most fluctuating 1st sugar residue among the HS hexasaccharides and this trend, although weaker, is also seen for residues 3, 4, 5 and 6. However, the 2nd sugar residue for 6C, 7B, 7C, 9A shows fluctuation levels comparable to the other HS molecules. Therefore, the distinction between fluct_r2 and the remaining fluctuation descriptors can be attributed to the behavior of those four HS hexasaccharides. At the same time, all fluctuation parameters are strongly correlated to R_gyration and Length (negative correlation) as well as to the second dihedral angle minimum of the third glycosidic bond (glycosidic_3min2, positive correlation). The fluctuation of a molecule is strongly linked to the molecule's compactness as well as the rigidity of its structure, with less fluctuation corresponding to an extended structure with a higher radius of gyration.

All of the 27 HS show a strong preference for the $^4C_1$ geometry. Ring pucker descriptors (pucker_1C4 and pucker_4C1) are strongly anti-correlated to each other and similarly correlated to most other descriptors. However, while pucker_1C4 is positively correlated to the first dihedral angle minimum of the first glycosidic linkage (glycosidic_1min1), pucker_4C1 shows a negative correlation to this descriptor.

The descriptors of the glycosidic bond dihedral angle minima do not form clusters as distinct and clear as the other descriptors. Nevertheless, some of the glycosidic linkage descriptors can be seen to be strongly correlated to certain other descriptors in the dataset, e.g. the strong negative correlation between the second minimum of the 3rd linkage and R_gyration and Length, as well as the aforementioned positive correlation to fluctuation descriptors.

The enthalpy and entropy terms, as well as the dipole moment descriptor establish two separate clusters. Descriptors relating to the dipole moment, electrostatic energy, the non-polar solvation energy, the total free energy and the normal mode-calculated entropy are positively correlated to each other (correlation coefficient above 0.78). At the same time, the van der Waals energy term, the polar solvation energy term,
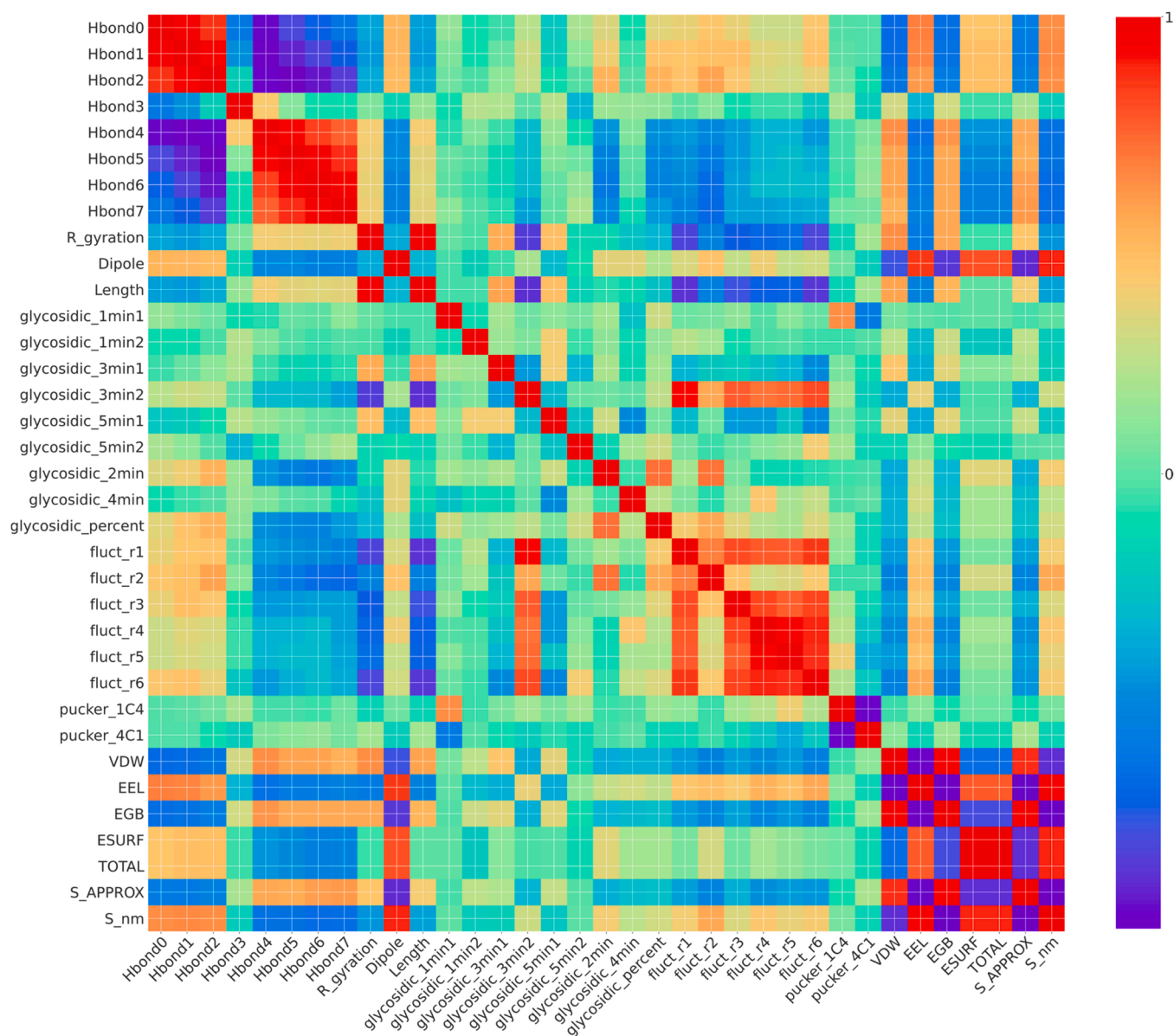
Fig. 4. Heatmap of Pearson correlation coefficients between the analyzed HS descriptors obtained from MD simulations.

and the quasi-harmonic entropy form another group of highly-correlated descriptors (correlation coefficients above 0.87). The negative correlation of the two entropy descriptors can seem surprising. What is important to consider for these descriptions of entropy, in this study they correspond to unbound ligand molecules and not, as may usually be the case when calculating entropy, to the protein-ligand complex formation. This negative correlation can be rooted in the difference in approaches used to calculate the entropy of a system. Quasi-harmonic entropy, corresponding to descriptor S_approx, is better at describing the translational and rotational changes in a molecule's conformation, closely related to the lengths, angles and torsions of bonds within the molecule. This approach also implicitly includes the effect of solvent on the entropy of the analyzed system. Normal mode entropy calculations, corresponding to S_nm in the descriptor data set, is linked to vibrational entropy, i.e. the bending and stretching of the molecule. Hence, both approaches describe different characteristics of the HS hexasaccharides and do not have to necessarily be positively correlated. The two groups formed by descriptors of enthalpy and entropy can be roughly classified as a group of "electrostatics" descriptors and a group of "shape" or "conformation" descriptors. The electrostatic

energy part of the total energy of the unbound HS molecules is the highest for 6A, 8A and 9A. The same oligosaccharides also exhibit very low van der Waals interaction energy, polar energy and quasi-harmonic entropy values. This could mean that the relatively high (i.e. unfavorable) electrostatic energy caused by the conformation of those molecules is balanced out by the favorable shape adapted by the molecules during the MD simulation.

The MD analysis of the 27 HS molecules provides invaluable information about the conformational landscape of those GAG molecules, uncovering preferences among the different HS types that may ultimately influence the affinity and specificity of binding to proteins.

3.3. Correlation analysis using Pearson product-moment correlation

The correlation between each of the 35 descriptors and the binding affinity data for 9 HS-protein complexes was assessed using Pearson product-moment correlation. Supplementary Table 2 contains the correlation coefficient values for all the descriptors. One immediate observation is that for different proteins, particular descriptors are correlated with binding affinity of respective HS-protein complexes.

ATIII is the most distinct from all other considered proteins – only a small number of descriptors can be found to be linked to the binding affinity: the fluctuation of the second sugar ring and the dihedral angle descriptor for the 4th glycosidic linkage.

3.3.1. Intramolecular H-bonds

H-bond descriptors do not seem to be strongly correlated with the binding affinity of the remaining proteins, except for BMP2 (strong correlation for Hbond0 and Hbond3), FGFR1 (strong correlation for Hbond0, Hbond1, Hbond3) and Stab-2 (strong correlation for Hbond0, Hbond1, Hbond3).

3.3.2. Radius of gyration

The radius of gyration and length of the HS molecules are similarly linked only to the binding affinity for complexes containing BMP2, FGFR1 and Stab-2. The dipole moment seems to be strongly correlated for all proteins except for ATIII.

3.3.3. Glycosidic linkage dihedral angles

In the case of glycosidic linkage dihedral angle descriptors, the descriptor relating to the second minimum of the first glycosidic linkage is strongly negatively correlated only with HCII and FGF-7. The descriptor of the first minimum of the 3rd glycosidic linkage as well as the first minimum of the 5th glycosidic linkage exhibit strong anti-correlation with the binding affinity of all protein complexes except for ATIII. The descriptor of the dihedral angle of the 4th glycosidic linkage positively correlated with complexes containing HCII, BMP2, Nrp1, and negatively correlated with ATIII.

3.3.4. Sugar ring fluctuation

The fluctuation of the sugar residues is linked to the binding affinities only in the cases of: the second sugar ring and ATIII, the fourth sugar ring and BMP2, FGFR1, Stab-2, Nrp1, the fifth sugar ring and FGFR1, Stab-2, and the sixth sugar ring and BMP2, FGFR1, Stab-2, Nrp1.

3.3.5. Enthalpy and entropy

The enthalpy and entropy strongly correlated for all proteins except for ATIII as well as in the case of the solvation descriptor (ESURF) and the total energy descriptor (TOTAL) and proteins FGFR1 and Stab-2.

Taken together, this analysis indicates that there is no clear unique pattern of correlation across the complexes, with each exhibiting differences in correlation sign and value. Therefore, no one elegant analytical function, linking the descriptors of intramolecular physico-chemical properties of the HS molecules with the binding affinity of complexes, could be constructed even for groups of HS-protein complexes. Instead, any further analysis incorporating these correlation values would necessarily have to be tailored to each HS-protein complex.

3.4. Principal component analysis

Based on the conclusion drawn in the previous section, PCA seems, therefore, to be a appropriate next step for the investigation of the descriptors set. This alternative approach to the calculation of correlation coefficients for each pair of descriptor and HS-protein complex binding affinity is feature extraction, by which initial properties (here: the physico-chemical descriptors of HS hexasaccharides) are transformed into a set of new, independent descriptors (Principal Components, PCs). The constructed PCs are linear combinations of the original descriptors and as such combine information from the whole unreduced initial dataset.

The scree plot (Fig. 5) was generated in order to assess the optimal amount of PCs needed to explain as much of the variance in the descriptor dataset as possible while still offering a reduction in dimensionality. Since the goal of applying PCA was not to reduce the descriptor dataset for visualization in a 2D- nor 3D-space but rather to

find an overarching model describing the relationship between characteristics of the unbound HS hexasaccharides and their protein binding affinity, the amount of PCs for subsequent analysis was reduced to 4 PCs. It can be seen that the first 4 PCs combined explain 74% of the variance of the initial descriptor dataset, whereas adding the 5th PC would not increase that cumulative variance in a meaningful way.

To decipher the chemical nature of the constructed PCs, the contribution of each descriptor to the PCs was calculated and visualized by means of a heatmap (Fig. 6). The main contributors to the first PC (PC1) are H-bond descriptors (with the exception of Hbond3) and terms relating to enthalpy and entropy. PC2 is mostly influenced by R_gyration, Length, ring fluctuation (except fluct_r2), ESURF, enthalpy/entropy terms, and most of the glycosidic linkage dihedral angle descriptors. The dihedral angle descriptors are the main driving force behind PC3, whereas enthalpy, entropy and H-bond terms contribute less to this PC. The dominant HS properties governing PC4 are the ring pucker descriptors, the first minimum of the first glycosidic linkage dihedral angle and, to a lesser extent, Hbond3. The colors in Fig. 6 indicate the interplay between different descriptors and strength of contribution to each PC.

Some of the observed relationships are to be expected, e.g. the opposition of VDW to EEL – in cases where structural complementarity plays a greater role, the van der Waals interaction energy matters more than the electrostatic energy and vice versa. The antagonistic behavior of the low- and high-H-bond descriptors is also understandable and in accordance with previous observations. However, Hbond3 shows a surprising pattern: while for PC1 and PC2 it is not strongly tied to either of the two other groups of H-bond descriptors, in PC3 it follows the same trend as low-H-bond terms, while in PC4 it behaves more similar to high-H-bond descriptors. Another intriguing detail is the complete absence of fluct_r2 from PC2, despite the strong presence of all other fluctuation-related descriptors in that PC. Moreover, the two descriptors R_gyration and Length operate in the same direction in all descriptors except PC1, where they act in an opposing manner.

Additional information on the cooperation between different descriptors and PCs can be gained from inspecting the loading plots (Supplementary Fig. 4). In loading plots, the length of the vectors indicate the importance of the given descriptor for the PC. The orientation in space of the vectors depicts whether the descriptor has an influence on both, only one or none of the PCs considered in the plot.

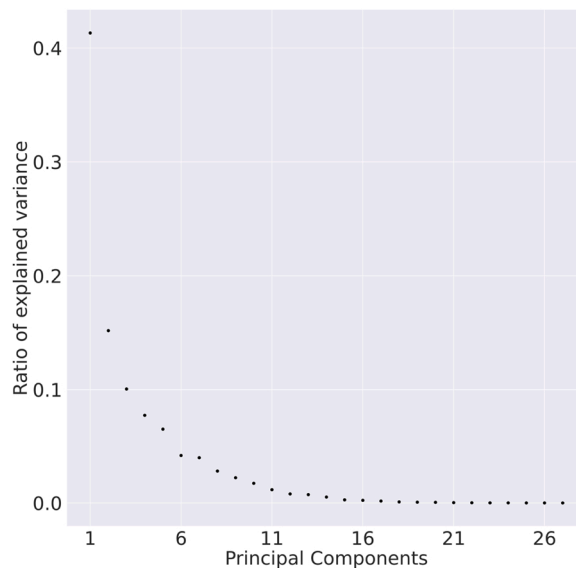Analysis of the interactions between descriptors in the PC-space

Fig. 5. Scree plot detailing the fraction of variance explained by the Principal Components.
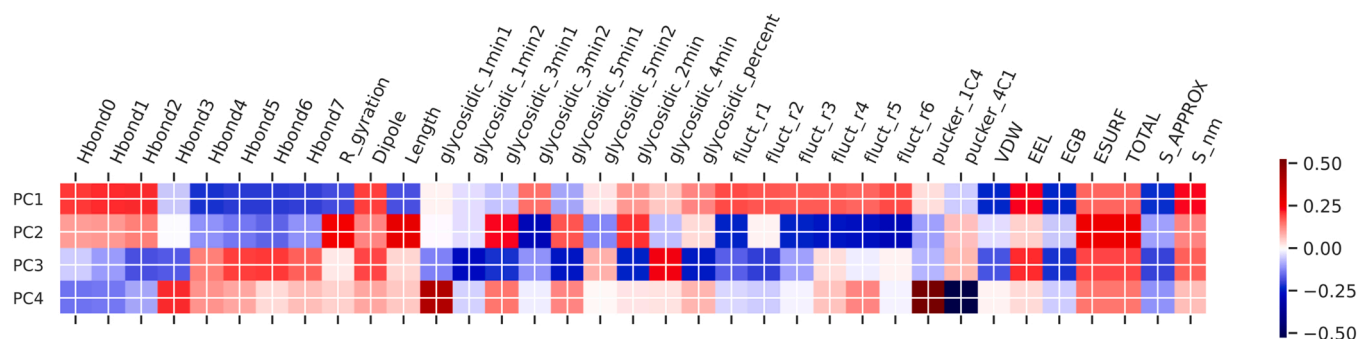
Fig. 6. Heatmap showing the contribution of descriptors of unbound HS molecules to the Principal Components.

illustrates a variety of coexisting conformations characterized by distinct values adopted by the analyzed descriptors. A low amount of intramolecular H-bonds formed, together with a less compact molecule (high radius of gyration and length of the HS molecule) coincides with high ring fluctuations, a larger dipole moment, high (unfavorable) electrostatic energy values, a greater entropy stemming from vibration and stretching of the molecule, as well as less entropy linked to the bending of the HS molecule. This corresponds to an extended, highly polar HS molecule; by analyzing the biplots of Supplementary Fig. 4, it can be seen that oligosaccharides likely to be found in such conformations are e.g. 9A, 9C, 7A, 7C, 8A, 8B, 8C, 1A, 4A, 9B, 5A. A contrasting conformation would be characterized by a more compact structure, a higher amount of intramolecular H-bonds, less fluctuation of the sugar rings, low solvation energy, more entropy contributed by bending of the molecule and low electrostatic energy. The HS hexasaccharides that seem to correspond to this combination of characteristics are 1B, 1C, 2A, 2B, 2 C, 3B, 4B, 4C, 5B, 5C.

A different look on the conformation would consider the glycosidic linkages. A preference of the dihedral angles of the 4th glycosidic linkage for specific dihedral values (i.e. a high concentration of observations belonging to the dihedral angle minimum, described by glycosidic_4min) co-occurs with dihedral angle values spread out more for the 1st, 2nd, 3rd, 5th glycosidic linkages (less observations found in the minima during the MD simulations). This situation would be accompanied by very low fluctuations of the 1st and 2nd sugar rings, a high electrostatic energy, higher solvation energy, and a higher vibrational entropy (S_nm). HS hexasaccharides that could be linked to this conformation include 2A, 3A, 3B, 3C, 5A, 6A, 9A, 9B, 9C.

While ring pucker descriptors contribute mostly to PC4, which explains the least variation in the initial dataset compared to the other PCs, the information they carry is still valuable for describing the characteristics of some of the HS oligosaccharides. A strong preference for the $^1C_4$ sugar ring puckering is linked to a high concentration of dihedral bonds of the 1st glycosidic linkage that belong to the minimum dihedral angle values, as well as a high fraction of structures throughout the MD trajectory that display 3 intramolecular H-bonds. This arrangement of characteristics can be linked to oligosaccharides 1A, 1B, 6A, 7B. The opposite situation, corresponding to a preference for $^4C_1$ sugar ring puckering, can be seen to be characteristic especially for HS 9A.

When taking into account the HS-binding affinities determined by (Chopra et al., 2021), additional information can be gained about the preference of the HS-binding proteins regarding the characteristics of the hexasaccharides. As detailed in Fig. 1 in (Chopra et al., 2021), ATIII shows a high affinity for 7A, 7B, a weaker affinity for 4A, 4B, and only minimal affinity for the other hexasaccharides. In the PC-space, unbound 7A and 7B exhibit low intramolecular H-bonding and high ring fluctuations, while differing in the ESURF and length of the molecules (7A is more extended than 7B) and ring puckering conformations (7A prefers $^4C_1$, while 7B prefers $^1C_4$). Unbound 4B forms more intramolecular H-bonds than 7A and 7B. Both 4A and 4B are characterized by a less defined conformation of the glycosidic linkages (limited

preference of the minimum conformation) and less fluctuation of the sugar rings.

The other analyzed proteins show a clear preference for hexasaccharides 9A and 3A. According to our MD analysis, in its unbound form 9A is compact, forms little to no intramolecular H-bonds, exhibits high fluctuation of sugar ring atoms and shows a strong preference for $^4C_1$ ring puckering. At the same time, 3A is not characterized by considerable fluctuations of its sugar rings. Both unbound 9A and 3A show little preference of glycosidic linkage dihedral bond minima.

The obtained PCs provide a sort of model explaining the physico-chemical properties of the 27 unbound 3-O-sulfated HS hexasaccharides. Each of the HS molecules differs to some extent from the others in the PC-space which is significantly reduced in dimensionality relative to the original descriptor-space yet contains all relevant information on the chemical and structural characteristics of the studied GAGs. The observed differences may contribute to the understanding of HS-protein binding specificity and of the sulfation code.

3.5. Linear regression analysis

Linear regression was used to evaluate the extent to which descriptors of the unbound HS hexasaccharides can explain HS-protein binding affinity. The experimental binding affinity data between the 27 analyzed HS molecules and a set of protein targets known to specifically recognize 3-O-sulfation was taken from (Chopra et al., 2021). The independent variables of the linear regression model were the four PCs obtained from PCA, with the binding affinity value expressed in Relative Fluorescent Unit (RFU) as response variable. The results of this analysis are detailed in Supplementary Table 3.

For all proteins except ATIII, PC1 and PC3 are consistently statistically significant (p < α = 0.05). These PCs corresponded mostly to, in the case of PC1, H-bonds, enthalpy and entropy, as well as glycosidic linkage dihedral angles for PC3. Hence, it could be assumed that those sets of descriptors of unbound HS hexasaccharides are unlikely to have any substantial influence on the HS/ATIII binding affinity. While the other two PCs, PC2 and PC4, do not exhibit statistically significant links to the response variable, it cannot be ruled out that the descriptors these PCs summarize contribute to the binding affinity of the GAG-protein complexes. For all proteins except ATIII, the PC-based linear regression model offers a high degree of explanation of the target variable, with R^2 values ranging between 0.68 and 0.8 (adjusted R^2 between 0.62 and 0.77). Thus, the MD-derived descriptors of unbound 3-O-sulfated HS hexasaccharides, simulated by MD, are alone capable of explaining up to around 70% of the binding affinity data in 3-O-sulfated HS-protein interactions. In the case of ATIII, the R^2 equals 0.11 (adjusted $R^2 = -0.06$), which indicates that the PC-based linear regression models fails at explaining the variance in HS/ATIII binding affinity. Consequently, it can be deduced that the descriptors serving as main signals in the four PCs considered are not the driving force behind the binding affinity between ATIII and 3-O-sulfated HS.

3.6. Cluster analysis

In order to discover any patterns and subgroups in the binding affinity data in relation to the descriptor-based PC space, cluster analysis by means of Hierarchical Clustering (HC) with average linkage was conducted. The clusterization of the PC versus binding affinity correlation matrix is depicted as a heatmap in Fig. 7.

Three main groups of proteins could be identified by cluster analysis – "group 0″ composed only of ATIII, "group 1" comprising BMP-2, FGFR-I and Stab-2, and "group 2" corresponding to RAGE, Nrp-1, HC-II, FGF-7, FGF-9. All proteins are positively-correlated with PC1 (the H-bond and free energy PC), however for ATIII this correlation is the weakest. In the case of PC3, ATIII shows negative correlation with all other proteins and a strong positive correlation with the values of PC3. The same pattern, although reversed in direction, can be observed for PC4, however here groups 1 and 2 differ visibly in the strength of correlation. In the case of PC2, only group 1 follows a clear trend as a whole, with all group 1 proteins negatively correlated to PC2. ATIII shows only a very weak positive correlation, while group 2 is more heterogeneous in its correlation to PC2 across its protein members, with some showing a weak positive and other a weak negative correlation.

Cluster analysis by HC revealed differences between the proteins in PC-space that may point to interesting information about the character of their binding with 3-O-sulfated HS. Among the descriptors considered, the binding between group 2-proteins and HS seems to be thus mostly linked to H-bond formation, enthalpy and entropy, as well as the glycosidic bond dihedral angle. Based on the HC analysis, HS/ATIII binding is influenced by, in order of importance, glycosidic bond dihedral angle descriptors, H-bonds and free energy components, ring puckering, and the least by sugar ring fluctuations. Meanwhile, the binding between group 1-proteins and 3-O-sulfated HS can be said to be influenced by all four PCs considered.

Taken together with the information about contribution of each descriptor to the PCs, it is possible to further speculate on the exact character of the HS-protein binding. PC3 is strongly linked to descriptors of glycosidic linkage dihedral angle minima and for all those descriptors except glycosidic_4min the correlation to PC3 is negative. A positive correlation between the HS-binding affinity of a protein and PC3 consequently also means a negative correlation to the glycosidic linkage descriptors and a more spread out distribution of dihedral angle values sampled from the MD trajectories. Conversely, a negative correlation with PC3, as is the case for ATIII, means a positive correlation with the glycosidic bond descriptors, thus pointing towards a preference of a high fraction of data points in the minima and a specificity of dihedral angle values of the HS molecules sampled from the MD simulations.

Similar conclusions can be drawn about proteins of group 1 and sugar ring fluctuation and puckering. Both sugar ring puckering and fluctuation describe the flexibility of the HS molecule. The strong correlation of group 1-binding affinity with PC2 and PC4 indicates a clear preference of certain conformational states. A negative correlation with PC2 corresponds to a positive correlation with fluctuation descriptors, thus a greater link between binding affinity and high fluctuation and consequently a greater entropy of the HS molecule. At the same time, the negative correlation with PC4 entails a link between binding affinity and a tendency for the $^4C_1$ ring pucker conformation, thus less flexibility in terms of ring puckering. Conversely, no strong correlation with PC4, as is the case e.g. for some proteins of group 2, would indicate no strong tendency towards either pucker conformation and thus perhaps an increased flexibility of the HS molecules bound with high affinity. Interestingly, ATIII is known to have a preference of certain ring pucker conformations of its ligands (Demeter et al., 2018), while in our analysis only a weak correlation with PC4 could be observed (Fig. 7). However,
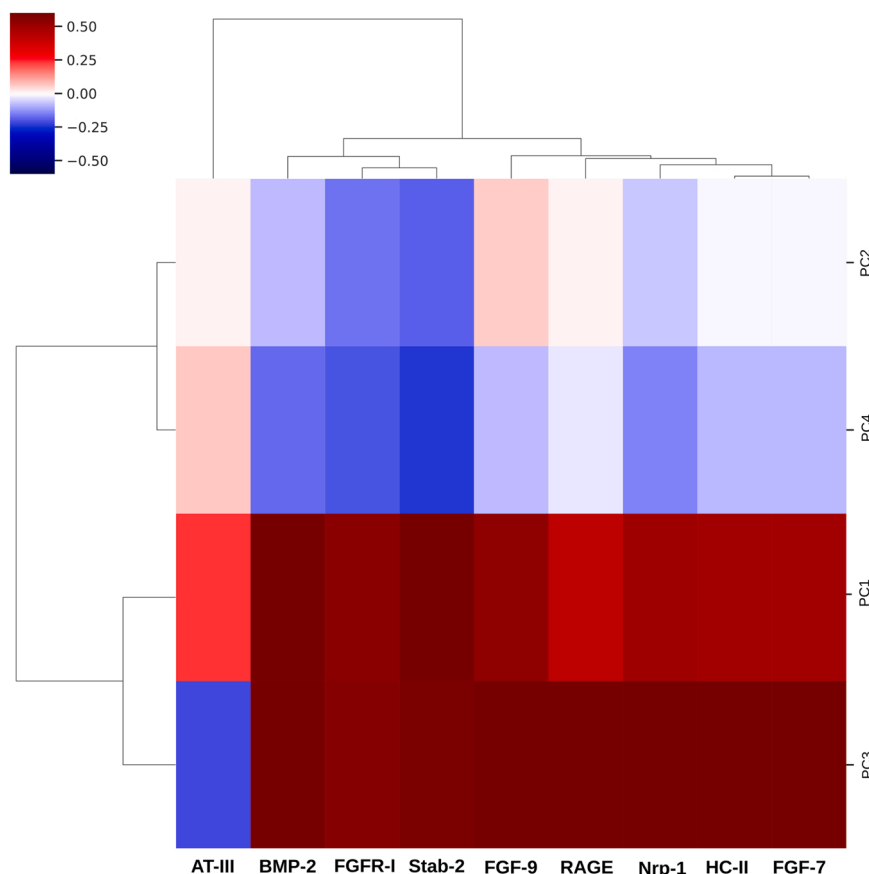

Fig. 7. Heatmap and hierarchical clustering of the correlation coefficients between PCs and HS-protein binding affinity values.

(Stancanelli et al., 2018) found that the ring pucker conformation of unbound HP is not obligatory for high-affinity binding, as the correct conformation is adopted by HP upon binding to ATIII. Therefore, a weak correlation with PC4 does not necessarily correspond to a lack of preference of sugar ring conformation.

3.7. Assessment of binding affinity based on the MD-derived Principal Components

In order to use the obtained PCs as a scoring function, adding descriptors of either the protein or specifically the HS-binding site would be necessary. However, knowledge of structure or function similarity, evolutionary relationships between proteins or co-localization of proteins in different sub-cellular locations, tissues or organs can also be used to form hypotheses on the interaction of proteins with ligands. We used the experimental binding affinity between Fibroblast Growth Factor 2 (FGF-2) and the 27 3-*O*-sulfated HS hexasaccharides (Chopra et al., 2021) to examine the usefulness of our PCA-based approach in predicting the approximate physico-chemical nature of the HS/FGF-2 binding. Due to being a member of the FGF family and interacting with Nrp-1 (Uniewicz and Fernig, 2008), we hypothesized that the binding of FGF-2 to the 27 HS molecules will be similar in character to the binding affinity of FGF-7, FGF-9 and Nrp-1. Therefore, FGF-2 would be assigned to group 2" of HS-binding proteins by HC.

We calculated the mean binding affinity for each sub-group identified by cluster analysis (i.e. group 0, group 1, group 2) and calculated the Pearson product-moment correlation between the FGF-2 binding affinity and the mean binding affinities. The highest correlation was between FGF-2 and the mean binding affinity of group 2 (correlation coefficient equal to 0.82), followed by group 1 (correlation coefficient 0.78) and ATIII (group 0, correlation coefficient = −0.006). After including the FGF-2 data into the HC analysis, the clustering algorithm thus correctly sorted FGF-2 to group 2. As can be seen in Supplementary Fig. 5, the binding profile of FGF-2 is indeed similar to the binding profile of the other proteins of group 2: RAGE, Nrp-1, HC-II, FGF-7, FGF-9. Proteins of this group seem to lack the slightly higher selectivity of HS binding that can be observed for proteins of group 1. Therefore, the PCA approach, trained on data not including the FGF-2 binding affinity, was able to distill the information contained in the unbound HS molecules in such a way that enabled us to correctly link the experimental binding affinity with principal components containing the most relevant physico-chemical information on the unbound HS structures.

4. Conclusions

The analysis of the physico-chemical descriptors of unbound HS hexasaccharides using computational approaches gave insight into the specificity and nature of binding between HS and a set of diverse proteins as well as into understanding of the sulfation code. Differences in sequence dictated preferences of unbound HS molecules regarding shape and flexibility. The conformational landscape of unbound HS that arose from the cooperation of physico-chemical descriptors is likely an important factor in the determination of binding affinity and specificity in HS/protein binding.

The elucidation of the links between properties of the HS molecules to the HS/protein binding affinity determined by (Chopra et al., 2021) as well as removing redundancy from the available data was achieved by applying PCA to the initial descriptor dataset and identifying four principal components corresponding to the main groups of characteristics that explained most of the variance of the HS descriptors. The calculation of correlation between the obtained PCs and the HS-protein binding affinity values followed by hierarchical clustering of the correlation coefficients, as well as analysis of the association between PCs and binding affinity data clearly showed that information obtained from MD simulations of unbound HS molecules can offer insight into HS/protein binding. All analyses underlined the difference in

HS-binding specificity between ATIII and the other proteins, which is in accordance with multiple experimental and theoretical studies (e.g. Mosier et al., 2012; Raghuraman et al., 2010). Additionally, we were able to further divide the remaining proteins into two subsets depending on the correlation of their HS-binding affinity with PCs describing the unbound HS GAGs.

The linear regression analysis on the PCA and HS-binding affinity data showed that, except for ATIII, the binding of HS by the proteins could be explained to a high degree by the information summarized in the PCs. Therefore, PCA not only reduced the dimension of the initial descriptor dataset but also allowed us to identify the main characteristics of HS that can putatively contribute to protein binding affinity. However, depending on the HS/protein complex considered, a part of variance in binding affinity was left unaccounted for by the HS descriptors and PCs. This unexplained variance may be caused by the lack of descriptors of either the proteins or the GAG-binding sites located on those proteins in our study. While GAG-binding sites on proteins are usually composed of clusters of positively-charged amino-acid residues, (Sarkar and Desai, 2015) describe the importance of other types of residues in the binding site and its vicinity for the specificity of protein/GAG binding. (Joseph et al., 2015) identified residues of IL–8 outside of the GAG-binding site that do not directly bind the GAG but contribute to GAG-binding by enhancing the plasticity of the binding site in order to accommodate GAG molecules. Therefore, the differences between the proteins that can be reflected by other descriptors than those included in our study, not only restricted to the binding site size and amino-acid sequence, is likely to further improve our model. Additional factors known to influence GAG/protein binding affinity and specificity, which could contribute to our analysis in future research, is the presence of ions (Kogut et al., 2021; Multhaup, 1994), the change in protein conformation upon GAG-binding (Rosenberg et al., 1997; Sage et al., 2013) or considering the presence of water molecules in the GAG-binding site of proteins (Mosier et al., 2012).

The use of *in silico* methods allowed us to study in detail the properties of 3-*O*-sulfated HS hexasaccharides in relation to their protein binding specificity. This shed light on relationships between structural properties of the HS molecules and protein binding affinity. A deeper understanding of HS/protein-binding specificity attributed to the sulfation code and of the physico-chemical characteristics of HS molecules that assist in the identification and design of specific GAG molecules and their mimetics with applications in medicine and pharmacology.

CRediT authorship contribution statement

Annemarie Danielsson: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Małgorzata M. Kogut:** Formal analysis, Investigation, Visualization. **Martyna Maszota-Zieleniak:** Formal analysis, Investigation, Visualization. **Pradeep Chopra:** Data curation, Investigation. **Geert-Jan Boons:** Conceptualization, Supervision. **Sergey A. Samsonov:** Conceptualization, Investigation, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

All software used in this study is available for download free of charge from the websites of the respective developers: AMBER (https://ambermd.org/), Python (https://www.python.org/), R (https://www.r-project.org/). The trajectories obtained using MD

simulations as well as the results of PCA are available upon request. The binding affinity data was obtained from (Chopra et al., 2021).

Acknowledgments

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2022.107716.

References

Almond, A., 2018. Multiscale modeling of glycosaminoglycan structure and dynamics: current methods and challenges. Curr. Opin. Struct. Biol. 50, 58–64. https://doi.org/10.1016/j.sbi.2017.11.008.

Ashikari-Hada, S., Habuchi, H., Sugaya, N., Kobayashi, T., Kimata, K., 2009. Specific inhibition of FGF-2 signaling with 2-O-sulfated octasaccharides of heparan sulfate. Glycobiology 19, 644–654. https://doi.org/10.1093/glycob/cwp031.

Brickman, Y.G., Ford, M.D., Gallagher, J.T., Nurcombe, V., Bartlett, P.F., Turnbull, J.E., 1998. Structural modification of fibroblast growth factor-binding heparan sulfate at a determinative stage of neural development. J. Biol. Chem. 273, 4350–4359. https://doi.org/10.1074/jbc.273.8.4350.

Case, D.A., Betz, R.M., Cerutti, D.S., Cheatham III, T.E., Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., 2016. AMBER. University of California,, San Francisco, p. 2016.

Chopra, P., Joshi, A., Wu, J., Lu, W., Yadavalli, T., Wolfert, M.A., Shukla, D., Zaia, J., Boons, G.-J., 2021. The 3-O-sulfation of heparan sulfate modulates protein binding and lyase degradation. Proc. Natl. Acad. Sci. U. S. A 118, e2012935118. https://doi.org/10.1073/pnas.2012935118.

Clausen, T.M., Sandoval, D.R., Spliid, C.B., Pihl, J., Perrett, H.R., Painter, C.D., Narayanan, A., Majowicz, S.A., Kwong, E.M., McVicar, R.N., Thacker, B.E., Glass, C. A., Yang, Z., Torres, J.L., Golden, G.J., Bartels, P.L., Porell, R.N., Garretson, A.F., Laubach, L., Feldman, J., Yin, X., Pu, Y., Hauser, B.M., Caradonna, T.M., Kellman, B. P., Martino, C., Gordts, P.L.S.M., Chanda, S.K., Schmidt, A.G., Godula, K., Leibel, S. L., Jose, J., Corbett, K.D., Ward, A.B., Carlin, A.F., Esko, J.D., 2020. SARS-CoV-2 Infection Depends on Cellular Heparan Sulfate and ACE2. e15 Cell 183, 1043–1057. https://doi.org/10.1016/j.cell.2020.09.033.

Darden, T., York, D., Pedersen, L., 1993. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. J. Chem. Phys. 98, 10089–10092. https://doi.org/10.1063/1.464397.

Demeter, F., Gyöngyösi, T., Bereczky, Z., Kövér, K.E., Herczeg, M., Borbás, A., 2018. Replacement of the L-iduronic acid unit of the anticoagulant pentasaccharide idraparinux by a 6-deoxy-L-talopyranose - Synthesis and conformational analysis. Sci. Rep. 8, 13736. https://doi.org/10.1038/s41598-018-31854-z.

Denys, A., Allain, F., 2019. The Emerging Roles of Heparan Sulfate 3-O-Sulfotransferases in Cancer. Front. Oncol. 9, 507. https://doi.org/10.3389/fonc.2019.00507.

Esko, J.D., Kimata, K., Lindahl, U., 2009. Proteoglycans and Sulfated Glycosaminoglycans. In: Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W., Etzler, M.E. (Eds.), Essentials of Glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY).

Ferreras, L., Moles, A., Situmorang, G.R., El Masri, R., Wilson, I.L., Cooke, K., Thompson, L., Kusche-Gullberg, M., Vivès, R.R., Sheerin, N.S., Ali, S., 2019. Heparan sulfate in chronic kidney diseases: Exploring the role of 3-O-sulfation. Biochim. Biophys. Acta Gen. Subj. 1863, 839–848. https://doi.org/10.1016/j.bbagen.2019.02.009.

Gama, C.I., Hsieh-Wilson, L.C., 2005. Chemical approaches to deciphering the glycosaminoglycan code. Curr. Opin. Chem. Biol. 9, 609–619. https://doi.org/10.1016/j.cbpa.2005.10.003.

Gama, C.I., Tully, S.E., Sotogaku, N., Clark, P.M., Rawat, M., Vaidehi, N., Goddard, W.A., Nishi, A., Hsieh-Wilson, L.C., 2006. Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. Nat. Chem. Biol. 2, 467–473. https://doi.org/10.1038/nchembio810.

Götz, A.W., Williamson, M.J., Xu, D., Poole, D., Le Grand, S., Walker, R.C., 2012. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput. 8, 1542–1555. https://doi.org/10.1021/ct200909j.

Guerrini, M., Guglieri, S., Casu, B., Torri, G., Mourier, P., Boudier, C., Viskov, C., 2008. Antithrombin-binding octasaccharides and role of extensions of the active pentasaccharide sequence in the specificity and strength of interaction. Evidence for very high affinity induced by an unusual glucuronic acid residue. J. Biol. Chem. 283, 26662–26675. https://doi.org/10.1074/jbc.M801102200.

Guglier, S., Hricovíni, M., Raman, R., Polito, L., Torri, G., Casu, B., Sasisekharan, R., Guerrini, M., 2008. Minimum FGF2 binding structural requirements of heparin and heparan sulfate oligosaccharides as determined by NMR spectroscopy. Biochemistry 47, 13862–13869. https://doi.org/10.1021/bi801007p.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M.,

Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

Hendriks, J., Planelles, L., de Jong-Odding, J., Hardenberg, G., Pals, S.T., Hahne, M., Spaargaren, M., Medema, J.P., 2005. Heparan sulfate proteoglycan binding promotes APRIL-induced tumor cell proliferation. Cell Death Differ. 12, 637–648. https://doi.org/10.1038/sj.cdd.4401647.

Hintze, V., Samsonov, S.A., Anselmi, M., Moeller, S., Becher, J., Schnabelrauch, M., Scharnweber, D., Pisabarro, M.T., 2014. Sulfated glycosaminoglycans exploit the conformational plasticity of bone morphogenetic protein-2 (BMP-2) and alter the interaction profile with its receptor. Biomacromolecules 15, 3083–3092. https://doi.org/10.1021/bm5006855.

Ho, G., Broze, G.J., Schwartz, A.L., 1997. Role of heparan sulfate proteoglycans in the uptake and degradation of tissue factor pathway inhibitor-coagulation factor Xa complexes. J. Biol. Chem. 272, 16838–16844. https://doi.org/10.1074/jbc.272.27.16838.

Huang, Y., Mao, Y., Zong, C., Lin, C., Boons, G.-J., Zaia, J., 2015. Discovery of a Heparan Sulfate 3-O-Sulfation Specific Peeling Reaction. Anal. Chem. 87, 592–600. https://doi.org/10.1021/ac503248k.

Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Comput. Sci. Eng. 9, 90–95.

Irie, A., Yates, E.A., Turnbull, J.E., Holt, C.E., 2002. Specific heparan sulfate structures involved in retinal axon targeting. Dev. Camb. Engl. 129, 61–70.

Johnson, C.E., Crawford, B.E., Stavridis, M., Ten Dam, G., Wat, A.L., Rushton, G., Ward, C.M., Wilson, V., van Kuppevelt, T.H., Esko, J.D., Smith, A., Gallagher, J.T., Merry, C.L.R., 2007. Essential alterations of heparan sulfate during the differentiation of embryonic stem cells to Sox1-enhanced green fluorescent protein-expressing neural progenitor cells. Stem Cells Dayt. Ohio 25, 1913–1923. https://doi.org/10.1634/stemcells.2006-0445.

Jokiranta, T.S., Cheng, Z.-Z., Seeberger, H., Jòzsi, M., Heinen, S., Noris, M., Remuzzi, G., Ormsby, R., Gordon, D.L., Meri, S., Hellwage, J., Zipfel, P.F., 2005. Binding of complement factor H to endothelial cells is mediated by the carboxy-terminal glycosaminoglycan binding site. Am. J. Pathol. 167, 1173–1181. https://doi.org/10.1016/S0002-9440(10)61205-9.

Joseph, P.R.B., Mosier, P.D., Desai, U.R., Rajarathnam, K., 2015. Solution NMR characterization of chemokine CXCL8/IL-8 monomer and dimer binding to glycosaminoglycans: structural plasticity mediates differential binding interactions. Biochem. J. 472, 121–133. https://doi.org/10.1042/BJ20150059.

Kinnunen, T., Raulo, E., Nolo, R., Maccarana, M., Lindahl, U., Rauvala, H., 1996. Neurite outgrowth in brain neurons induced by heparin-binding growth-associated molecule (HB-GAM) depends on the specific interaction of HB-GAM with heparan sulfate at the cell surface. J. Biol. Chem. 271, 2243–2248. https://doi.org/10.1074/jbc.271.4.2243.

Kirschner, K.N., Yongye, A.B., Tschampel, S.M., González-Outeiriño, J., Daniels, C.R., Foley, B.L., Woods, R.J., 2008. GLYCAM06: a generalizable biomolecular force field. Carbohydr. J. Comput. Chem. 29, 622–655. https://doi.org/10.1002/jcc.20820.

Koehler, L., Samsonov, S., Rother, S., Vogel, S., Köhling, S., Moeller, S., Schnabelrauch, M., Rademann, J., Hempel, U., Pisabarro, M.T., Scharnweber, D., Hintze, V., 2017. Sulfated Hyaluronan Derivatives Modulate TGF-β1:Receptor Complex Formation: Possible Consequences for TGF-β1 Signaling. Sci. Rep. 7, 1210. https://doi.org/10.1038/s41598-017-01264-8.

Kogut, M.M., Maszota-Zieleniak, M., Marcisz, M., Samsonov, S.A., 2021. Computational insights into the role of calcium ions in protein-glycosaminoglycan systems. Phys. Chem. Chem. Phys. PCCP 23, 3519–3530. https://doi.org/10.1039/d0cp05438k.

Kraushaar, D.C., Rai, S., Condac, E., Nairn, A., Zhang, S., Yamaguchi, Y., Moremen, K., Dalton, S., Wang, L., 2012. Heparan sulfate facilitates FGF and BMP signaling to drive mesoderm differentiation of mouse embryonic stem cells. J. Biol. Chem. 287, 22691–22700. https://doi.org/10.1074/jbc.M112.368241.

Künze, G., Huster, D., Samsonov, S.A., 2021. Investigation of the structure of regulatory proteins interacting with glycosaminoglycans by combining NMR spectroscopy and molecular modeling - the beginning of a wonderful friendship. Biol. Chem. https://doi.org/10.1515/hsz-2021-0119.

Li, Y., Sun, C., Yates, E.A., Jiang, C., Wilkinson, M.C., Fernig, D.G., 2016. Heparin binding preference and structures in the fibroblast growth factor family parallel their evolutionary diversification. Open Biol. 6, 150275 https://doi.org/10.1098/rsob.150275.

Liu, J., Shriver, Z., Pope, R.M., Thorp, S.C., Duncan, M.B., Copeland, R.J., Raska, C.S., Yoshida, K., Eisenberg, R.J., Cohen, G., Linhardt, R.J., Sasisekharan, R., 2002. Characterization of a heparan sulfate octasaccharide that binds to herpes simplex virus type 1 glycoprotein D. J. Biol. Chem. 277, 33456–33467. https://doi.org/10.1074/jbc.M202034200.

Luo, Y., Ye, S., Kan, M., McKeehan, W.L., 2006. Structural specificity in a FGF7-affinity purified heparin octasaccharide required for formation of a complex with FGF7 and FGFR2IIIb. J. Cell. Biochem. 97, 1241–1258. https://doi.org/10.1002/jcb.20724.

McKinney, W., others, 2010. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. p. 51–6.

Morgan, A., Sepuru, K.M., Feng, W., Rajarathnam, K., Wang, X., 2015. Flexible Linker Modulates Glycosaminoglycan Affinity of Decorin Binding Protein A. Biochemistry 54, 5113–5119. https://doi.org/10.1021/acs.biochem.5b00253.

Mosier, P.D., Krishnasamy, C., Kellogg, G.E., Desai, U.R., 2012. On the specificity of heparin/heparan sulfate binding to proteins. Anion-binding sites on antithrombin and thrombin are fundamentally different. PloS One 7, e48632. https://doi.org/10.1371/journal.pone.0048632.

Multhaup, G., 1994. Identification and regulation of the high affinity binding site of the Alzheimer's disease amyloid protein precursor (APP) to glycosaminoglycans. Biochimie 76, 304–311. https://doi.org/10.1016/0300-9084(94)90163-5.

Nakato, H., Kimata, K., 2002. Heparan sulfate fine structure and specificity of proteoglycan functions. Biochim. Biophys. Acta 1573, 312–318. https://doi.org/10.1016/s0304-4165(02)00398-7.

O'Donnell, C.D., Shukla, D., 2008. The Importance of Heparan Sulfate in Herpesvirus Infection. Virol. Sin. 23, 383–393. https://doi.org/10.1007/s12250-008-2992-1.

Panitz, N., Theisgen, S., Samsonov, S.A., Gehrcke, J.-P., Baumann, L., Bellmann-Sickert, K., Köhling, S., Pisabarro, M.T., Rademann, J., Huster, D., Beck-Sickinger, A. G., 2016. The structural investigation of glycosaminoglycan binding to CXCL12 displays distinct interaction sites. Glycobiology 26, 1209–1221. https://doi.org/10.1093/glycob/cww059.

Patel, V.N., Likar, K.M., Zisman-Rozen, S., Cowherd, S.N., Lassiter, K.S., Sher, I., Yates, E. A., Turnbull, J.E., Ron, D., Hoffman, M.P., 2008. Specific heparan sulfate structures modulate FGF10-mediated submandibular gland epithelial morphogenesis and differentiation. J. Biol. Chem. 283, 9308–9317. https://doi.org/10.1074/jbc.M709995200.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pejler, G., Danielsson, A., Björk, I., Lindahl, U., Nader, H.B., Dietrich, C.P., 1987. Structure and antithrombin-binding properties of heparin isolated from the clams Anomalocardia brasiliana and Tivela mactroides. J. Biol. Chem. 262, 11413–11421. https://doi.org/10.1016/S0021-9258(18)60822-1.

Pempe, E.H., Xu, Y., Gopalakrishnan, S., Liu, J., Harris, E.N., 2012. Probing structural selectivity of synthetic heparin binding to Stabilin protein receptors. J. Biol. Chem. 287, 20774–20783. https://doi.org/10.1074/jbc.M111.320069.

Pichert, A., Samsonov, S.A., Theisgen, S., Thomas, L., Baumann, L., Schiller, J., Beck-Sickinger, A.G., Huster, D., Pisabarro, M.T., 2012. Characterization of the interaction of interleukin-8 with hyaluronan, chondroitin sulfate, dermatan sulfate and their sulfated derivatives by spectroscopy and molecular modeling. Glycobiology 22, 134–145. https://doi.org/10.1093/glycob/cwr120.

Pomin, V.H., Mulloy, B., 2015. Current structural biology of the heparin interactome. Curr. Opin. Struct. Biol. 34, 17–25. https://doi.org/10.1016/j.sbi.2015.05.007.

Pratt, C.W., Church, F.C., 1992. Heparin binding to protein C inhibitor. J. Biol. Chem. 267, 8789–8794.

Pye, D.A., Vivès, R.R., Hyde, P., Gallagher, J.T., 2000. Regulation of FGF-1 mitogenic activity by heparan sulfate oligosaccharides is dependent on specific structural features: differential requirements for the modulation of FGF-1 and FGF-2. Glycobiology 10, 1183–1192. https://doi.org/10.1093/glycob/10.11.1183.

Qazi, H., Shi, Z.-D., Song, J.W., Cancel, L.M., Huang, P., Zeng, Y., Roberge, S., Munn, L. L., Tarbell, J.M., 2016. Heparan sulfate proteoglycans mediate renal carcinoma metastasis. Int. J. Cancer 139, 2791–2801. https://doi.org/10.1002/ijc.30397.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Raghuraman, A., Mosier, P.D., Desai, U.R., 2010. Understanding Dermatan Sulfate-Heparin Cofactor II Interaction through Virtual Library Screening. ACS Med. Chem. Lett. 1, 281–285. https://doi.org/10.1021/ml100048y.

Raghuraman, A., Mosier, P.D., Desai, U.R., 2006. Finding a needle in a haystack: development of a combinatorial virtual screening approach for identifying high specificity heparin/heparan sulfate sequence(s). J. Med. Chem. 49, 3553–3562. https://doi.org/10.1021/jm060092o.

Revelle, William, 2021. psych: Procedures for Personality and Psychological Research [WWW Document]. URL ⟨https://CRAN.R-project.org/package=psych⟩.

Roe, D.R., Cheatham, T.E., 2013. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. 9, 3084–3095. https://doi.org/10.1021/ct400341p.

Rogers, C.J., Clark, P.M., Tully, S.E., Abrol, R., Garcia, K.C., Goddard, W.A., Hsieh-Wilson, L.C., 2011. Elucidating glycosaminoglycan-protein-protein interactions using carbohydrate microarray and computational approaches. Proc. Natl. Acad. Sci. U. S. A 108, 9747–9752. https://doi.org/10.1073/pnas.1102962108.

Rosenberg, R.D., Shworak, N.W., Liu, J., Schwartz, J.J., Zhang, L., 1997. Heparan sulfate proteoglycans of the cardiovascular system. Specific structures emerge but how is synthesis regulated. J. Clin. Invest. 99, 2062–2070. https://doi.org/10.1172/JCI119377.

Rother, S., Samsonov, S.A., Hempel, U., Vogel, S., Moeller, S., Blaszkiewicz, J., Köhling, S., Schnabelrauch, M., Rademann, J., Pisabarro, M.T., Hintze, V., Scharnweber, D., 2016. Sulfated Hyaluronan Alters the Interaction Profile of TIMP-3 with the Endocytic Receptor LRP-1 Clusters II and IV and Increases the Extracellular TIMP-3 Level of Human Bone Marrow Stromal Cells. Biomacromolecules 17, 3252–3261. https://doi.org/10.1021/acs.biomac.6b00980.

Ryckaert, J.-P., Ciccotti, G., Berendsen, H.J.C., 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. J. Comput. Phys. 23, 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

Sage, J., Mallèvre, F., Barbarin-Costes, F., Samsonov, S.A., Gehrcke, J.-P., Pisabarro, M. T., Perrier, E., Schnebert, S., Roget, A., Livache, T., Nizard, C., Lalmanach, G., Lecaille, F., 2013. Binding of chondroitin 4-sulfate to cathepsin S regulates its enzymatic activity. Biochemistry 52, 6487–6498. https://doi.org/10.1021/bi400925g.

Sankaranarayanan, N.V., Desai, U.R., 2014. Toward a robust computational screening strategy for identifying glycosaminoglycan sequences that display high specificity for target proteins. Glycobiology 24, 1323–1333. https://doi.org/10.1093/glycob/cwu077.

Sankaranarayanan, N.V., Sarkar, A., Desai, U.R., Mosier, P.D., 2015. Designing "high-affinity, high-specificity" glycosaminoglycan sequences through computerized modeling. Methods Mol. Biol. Clifton NJ 1229, 289–314. https://doi.org/10.1007/978-1-4939-1714-3_24.

Sankarayanarayanan, N.V., Strebel, T.R., Boothello, R.S., Sheerin, K., Raghuraman, A., Sallas, F., Mosier, P.D., Watermeyer, N.D., Oscarson, S., Desai, U.R., 2017. A Hexasaccharide Containing Rare 2-O-Sulfate-Glucuronic Acid Residues Selectively Activates Heparin Cofactor II. Angew. Chem. Int. Ed. Engl. 56, 2312–2317. https://doi.org/10.1002/anie.201609541.

Sarkar, A., Desai, U.R., 2015. A Simple Method for Discovering Druggable, Specific Glycosaminoglycan-Protein Systems. Elucidation of Key Principles from Heparin/Heparin Sulfate-Binding Proteins. PloS One 10, e0141127. https://doi.org/10.1371/journal.pone.0141127.

Sattelle, B.M., Shakeri, J., Cliff, M.J., Almond, A., 2015. Proteoglycans and Their Heterogeneous Glycosaminoglycans at the Atomic Scale. Biomacromolecules 16, 951–961. https://doi.org/10.1021/bm5018386.

Schlorke, D., Thomas, L., Samsonov, S.A., Huster, D., Arnhold, J., Pichert, A., 2012. The influence of glycosaminoglycans on IL-8-mediated functions of neutrophils. Carbohydr. Res. 356, 196–203. https://doi.org/10.1016/j.carres.2012.02.025.

Shworak, N.W., Shirakawa, M., Colliec-Jouault, S., Liu, J., Mulligan, R.C., Birinyi, L.K., Rosenberg, R.D., 1994. Pathway-specific regulation of the synthesis of anticoagulantly active heparan sulfate. J. Biol. Chem. 269, 24941–24952.

Spencer, J.L., Bernanke, J.A., Buczek-Thomas, J.A., Nugent, M.A., 2010. A Computational Approach for Deciphering the Organization of Glycosaminoglycans. PLOS ONE 5, e9389. https://doi.org/10.1371/journal.pone.0009389.

Stancanelli, E., Elli, S., Hsieh, P.-H., Liu, J., Guerrini, M., 2018. Recognition and Conformational Properties of an Alternative Antithrombin Binding Sequence Obtained by Chemoenzymatic Synthesis. Chembiochem Eur. J. Chem. Biol. https://doi.org/10.1002/cbic.201800095.

Stringer, S.E., Gallagher, J.T., 1997. Specific binding of the chemokine platelet factor 4 to heparan sulfate. J. Biol. Chem. 272, 20508–20514. https://doi.org/10.1074/jbc.272.33.20508.

Sun, Y.J., Chang, N.C., Hung, S.I., Chang, A.C., Chou, C.C., Hsiao, C.D., 2001. The crystal structure of a novel mammalian lectin, Ym1, suggests a saccharide binding site. J. Biol. Chem. 276, 17507–17514. https://doi.org/10.1074/jbc.M010416200.

Taylor, G.J., Yorke, S.C., Harding, D.R., 1995. Glycosaminoglycan specificity of a heparin-binding peptide. Pept. Res. 8, 286–293.

Thacker, B.E., Seamen, E., Lawrence, R., Parker, M.W., Xu, Y., Liu, J., Vander Kooi, C.W., Esko, J.D., 2016. Expanding the 3-O-Sulfate Proteome–Enhanced Binding of Neuropilin-1 to 3-O-Sulfated Heparan Sulfate Modulates Its Activity. ACS Chem. Biol. 11, 971–980. https://doi.org/10.1021/acschembio.5b00897.

Thacker, B.E., Xu, D., Lawrence, R., Esko, J.D., 2014. Heparan sulfate 3-O-sulfation: A rare modification in search of a function. Matrix Biol., Prote Biol. 35, 60–72. https://doi.org/10.1016/j.matbio.2013.12.001.

Ughy, B., Schmidthoffer, I., Szilak, L., 2019. Heparan sulfate proteoglycan (HSPG) can take part in cell division: inside and outside. Cell. Mol. Life Sci. CMLS 76, 865–871. https://doi.org/10.1007/s00018-018-2964-z.

Uniewicz, K.A., Fernig, D.G., 2008. Neuropilins: a versatile partner of extracellular molecules that regulate development and disease. Front. Biosci. J. Virtual Libr. 13, 4339–4360. https://doi.org/10.2741/3008.

Viviano, B.L., Paine-Saunders, S., Gasiunas, N., Gallagher, J., Saunders, S., 2004. Domain-specific modification of heparan sulfate by Qsulf1 modulates the binding of the bone morphogenetic protein antagonist Noggin. J. Biol. Chem. 279, 5604–5611. https://doi.org/10.1074/jbc.M310691200.

Waskom, M.L., 2021. seaborn: statistical data visualization. J. Open Source Softw. 6, 3021. https://doi.org/10.21105/joss.03021.

Witt, D.P., Lander, A.D., 1994. Differential binding of chemokines to glycosaminoglycan subpopulations. Curr. Biol. CB 4, 394–400. https://doi.org/10.1016/s0960-9822(00)00088-9.

Yokoyama, M., Matsuzawa, T., Yoshikawa, T., Nunomiya, A., Yamaguchi, Y., Yanai, K., 2020. Heparan sulfate controls skeletal muscle differentiation and motor functions. Biochim. Biophys. Acta Gen. Subj. 1864, 129707 https://doi.org/10.1016/j.bbagen.2020.129707.

Yue, J., Jin, W., Yang, H., Faulkner, J., Song, X., Qiu, H., Teng, M., Azadi, P., Zhang, F., Linhardt, R.J., Wang, L., 2021. Heparan Sulfate Facilitates Spike Protein-Mediated SARS-CoV-2 Host Cell Invasion and Contributes to Increased Infection of SARS-CoV-2 G614 Mutant and in Lung Cancer. Front. Mol. Biosci. 8, 649575 https://doi.org/10.3389/fmolb.2021.649575.

Zhang, B., Xiao, W., Qiu, H., Zhang, F., Moniz, H.A., Jaworski, A., Condac, E., Gutierrez-Sanchez, G., Heiss, C., Clugston, R.D., Azadi, P., Greer, J.J., Bergmann, C., Moremen, K.W., Li, D., Linhardt, R.J., Esko, J.D., Wang, L., 2014. Heparan sulfate deficiency disrupts developmental angiogenesis and causes congenital diaphragmatic hernia. J. Clin. Invest. 124, 209–221. https://doi.org/10.1172/JCI71090.

Zhao, J., Zhu, Y., Song, X., Yuanyuan, Xiao Su, G., Liu, X., Wang, Z., Xu, Y., Liu, J., Eliezer, D., Ramlall, T.F., Lippens, G., Gibson, J., Zhang, F., Linhardt, R.J., Wang, L., Wang, C., 2020. 3-O-Sulfation of Heparan Sulfate Enhances Tau Interaction and Cellular Uptake. Angew. Chem. Int. Ed. Engl. 59, 1818–1827. https://doi.org/10.1002/anie.201913029.