GlycoTorch Vina: Improved Docking of Sulfated Sugars Using QM-derived Scoring Functions

Eric D. Boittier[†], *Jed M. Burns*[†], *Neha S. Gandhi*[*,‡], *Vito Ferro*[*,†,¶]

[†]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland 4072, Australia

[‡]School of Mathematical Sciences and Institute of Health and Biomedical Innovation, Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Queensland 4000, Australia.

[¶]Australian Infectious Diseases Research Centre, The University of Queensland, Brisbane, Queensland 4072, Australia

RECEIVED DATE (to be automatically inserted after your manuscript is accepted if required according to the journal that you are submitting your paper to)

CORRESPONDING AUTHOR FOOTNOTE

*Correspondence

Vito Ferro, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland 4072, Australia. Email: v.ferro@uq.edu.au

Neha S. Gandhi, School of Mathematical Sciences and Institute for Health and Biomedical Innovation, Faculty of Science and Engineering, Queensland University of Technology, Brisbane, Queensland 4000, Australia. Email: neha.gandhi@qut.edu.au

Abstract

Glycosaminoglycans (GAGs) are a family of anionic carbohydrates that play an essential role in the physiology and pathology of all eukaryotic life. Experimental determination of GAG-protein complexes remains difficult due to the considerable diversity in both carbohydrate linkage, and sulfation patterns. To complement existing methods of structural determination, we present our molecular docking tool, GlycoTorchVina (GTV), which demonstrates a substantial improvement at reproducing low energy conformations of GAGs compared with traditional docking programs. Based on the carbohydrate specific docking program VinaCarb (VC), GTV utilizes rotational energy functions, calculated using density functional theory (DFT), specifically designed for glycosidic linkages found in GAGs. The redocking accuracy of four programs (GTV, VC, AutoDock Vina and Glide) was tested over a set of 10 high-quality crystal structures containing co-crystallized GAGs (tetrasaccharides or longer). GTV outperformed other programs and was able to reproduce the native pose of eight structures and produced top-scoring docked poses that were on average only 1.8 Å RMSD away from the crystal structure. Although imitation of crystal structures is a standard test used for assessing the accuracy of docking programs, we illustrate how the underlying quality of the crystal structure, which is often overlooked during benchmarking, affects conclusions drawn from this approach. Statistical and theoretical investigations into charge-charge ("salt-bridge") interactions are also presented. Again, DFT calculations were used to derive non-bonded potentials describing salt-bridges, and solvent-mediated charge-charge ("water-bridge") interactions. These data suggest that water-bridges play an important, yet poorly understood, role in the structures of GAG-protein complexes.

Keywords: Glycotorch, Glycosaminoglycans, VinaCarb, DFT, Molecular docking

Introduction

On the surface of all eukaryotic cells, highly sulfated complex carbohydrates known as glycosaminoglycans (GAGs) facilitate the reception and modulation of a diverse range of proteins.[1-3] These sugars are essential for normal physiological functions, such as blood coagulation and neuronal development,[4] and are also involved in severe pathophysiological disorders, such as cancer[5-6] and Alzheimer's disease.[7-8] GAGs are comprised of repeating disaccharide subunits consisting of an aminosugar bound to a uronic acid or D-galactose, and are broadly classified into four main groups depending on disaccharide composition: heparan sulfate (HS)/heparin, chondroitin sulfate/dermatan sulfate, keratan sulfate and hyaluronic acid. HS/heparin GAGs (or HSGAGs) are the most structurally diverse GAGs and contain a uronic acid residue (L-iduronic acid (IdoA) or its C5 epimer, D-glucuronic acid (GlcA)) linked to a D-glucosamine residue (GlcN) (See Figure 1a).[1-3] The GlcN residues in HSGAGs are commonly sulfated or acetylated at the C2 nitrogen substituent. HSGAGs exhibit a variety of sulfation patterns, and degree of sulfation (DoS, which refers to the number of sulfate/sulfamate residues per disaccharide pair). This natural heterogeneity suggests that the amount of information encoded in the structure of HSGAGs potentially exceeds that of DNA.[1] In addition to the C2 nitrogen, GlcN may be sulfated at C4, C6 or C3 (rare).[1-3] The uronic acid residue may be sulfated at the C2 or C3 position. For any given HSGAG octasaccharide there are over 1,000,000 possible sulfation patterns.[1] The relationship between sulfation pattern and the strength of interaction for a GAG-protein complex has been coined "the sulfation code".[9-11] It is hoped that cracking this code will shed light on the biology behind serious diseases and open up new avenues for drug discovery.

HSGAGs exhibit fascinating conformational flexibility.[12] When IdoA is sulfated, usually at C2, steric hindrance resulting from the C5 carboxylate and the additional sulfate being on the same face of the hexose ring can cause the sugar to distort and adopt unnatural ring conformations.[2, 13] IdoA (and GlcA, to a lesser extend) exists in equilibrium between the skew-boat ($^{2}S_{O}$) and ($^{1}C_{4}$) chair conformations (see Figure 1c for a description of these ring conformations), with the proportion of each conformer dependent on the nature of the surrounding monosaccharide residues.[14] The $^{2}S_{O}$ conformation is believed to be the

dominant conformation, based on crystallographic evidence and intra-ring proton–proton vicinal coupling constants seen in the ^{1}H NMR spectra.[15-17] While the exact proportion of the $^{2}S_{O}$ conformer in solution is contested in the literature, this unusual ring conformation has biological relevance.[17-19] The role of IdoA ring flexibility is emphasized in the interaction between heparin and the serpin antithrombin (AT). Heparin is essential for proper regulation of the blood coagulation cascade in humans.[20] Biological heparin, a high molecular weight polysaccharide, has a long history of use as a clinical anticoagulant, which continues to this day.[21-22] Through over a decade of research, a pentasaccharide representing the minimum active heparin sequence was characterized, and later a synthetic version was commercialized as the low molecular weight anticoagulant, fondaparinux (See Figure 1b).[23] Importantly, it was shown that sulfated IdoA was required for binding to AT. Crystal structures characterizing the heparin-AT complex model this specific sulfated IdoA in the $^{2}S_{O}$ conformation (See Figure 1d).[17] This unusual flexibility allows the molecule to adopt a low energy conformation, which may contribute to the high affinity for this specific sequence.

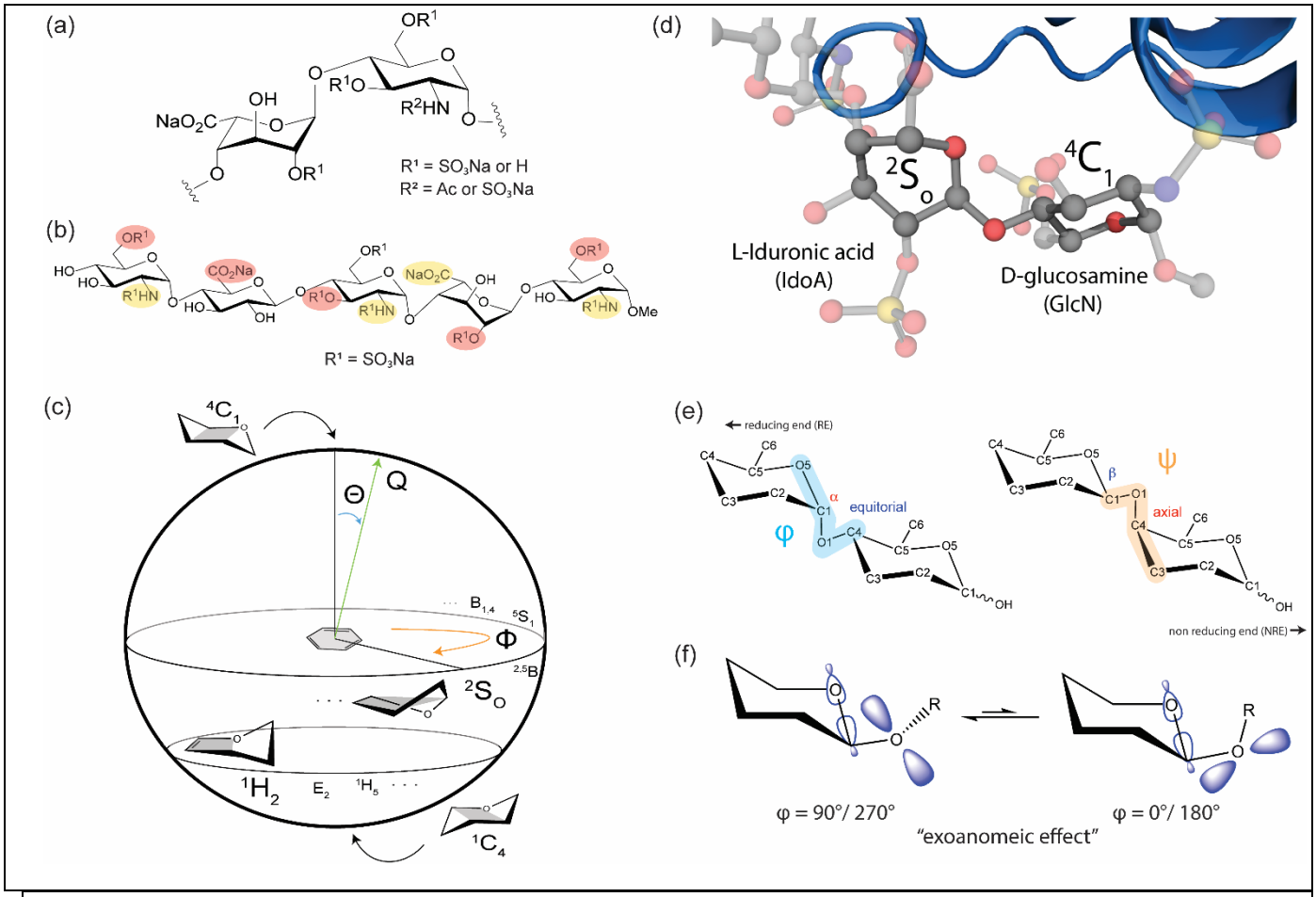

Figure 1. (a) Generic structure of the repeating GAG disaccharide building blocks that are common in HS/heparin. (b) Chemical structure of fondaparinux. Extensive SAR studies identified groups in red as essential for binding. Groups in yellow contribute to, but are not essential for, anticoagulant activity. (c) Ring conformations associated with GAGs and their relation to Cremer-Pople parameters (Φ and Θ). (d) Close up model from a crystal structure of fondaparinux bound to antithrombin (PDB code: 1NQ9), highlighting the prevalence of the $^{2}S_{O}$ IdoA conformation. (e) Carbohydrate nomenclature detailing the difference between α vs β anomeric configurations and axial vs equatorial hydroxyl groups. Atoms that define the ϕ (O5 – C1– Ox – Cx) and ψ (C1– Ox – Cx– Cx-1) torsions are labelled (x = 4 or 3 for 1 → 4 and 1 → 3 linked glycans, respectively). (f) Orbital diagram illustrating the exoanomeric effect. Preference for ϕ angles of approximately 90°/270° is determined by the favorable overlap of the anomeric oxygen and the adjacent carbon-oxygen anti-bonding orbital.

Experimental determination of protein-GAG complexes, by NMR and X-ray crystallography for example, is difficult due to high levels of natural heterogeneity and negative charge making purification from biological sources challenging. Although many crystal structures are available in the Protein Data Bank, the flexibility of GAGs can cause artifacts in the structural refinement process, and the quality of the models varies considerably.[24] Due to the immense diversity of these molecules, computational techniques (such as molecular docking) to complement experimental data, are highly sought after. A substantial limitation when using docking to predict the conformations of protein-bound GAGs is that most programs were designed for 'drug-like' molecules (i.e., docking scoring functions are parameterized to reproduce binding coefficients of drug-like molecules, and the exhaustiveness of search algorithms are often assessed with small molecule ligands).[25] For this reason, many docking programs perform best when docking small molecules with few rotatable bonds into clearly defined pockets, usually buried within the receptor. GAGs, however, defy several common assumptions that underpin standard protein-ligand docking. Firstly, as polysaccharides, they often contain tens of flexible dihedral torsions. Secondly, due to their highly anionic nature, they are often found interacting with the hydrophilic surface of proteins (i.e. solvent-exposed sites), usually in patches with high numbers of basic amino acids. Previous docking programs to include features designed for GAGs include Cluspro,[26] which offers a rigid docking method aimed at predicting GAG binding sites, and GAGDock,[27] an extension of an earlier program (DarwinDock) which uses a customized search algorithm to artificially restrict the search space to 'flat' regions of the protein surface, where GAGs are often bound. Initial validation results reported for GAGDock were promising (< 1 Å RMSD for all complexes), although the number of structures used to validate this method was considerably small (four structures from two protein families) and possibly do not represent the performance of the program for a more diverse range of proteins. Recently, a fragment-based approach was developed by Samsonov et al. which facilitates docking of larger GAGs (degree of polymerization (d.p.) > 5) by "stitching" together docking results of smaller (trisaccharides) and refining the final pose.[28]

An important step forward in carbohydrate-protein docking was taken by Nivedha and coworkers who introduced a series of energy scoring functions that describe low energy conformations of carbohydrates around the glycosidic linkage, referred to as CarboHydrate Intrinsic (CHI) energy functions.[29] These CHI-energy functions were developed for α/β disaccharides in $^4C_1$ and $^1C_4$ ring conformations, and describe the preference of ϕ, ψ and ω torsion angles in sugars. The angles ϕ (O5 – C1– Ox – Cx) and ψ (C1– Ox – Cx– Cx-1) (see Figure 1e) describe rotations around the glycosidic linkage in GAGs and are the main contributor to the flexibility of these molecules. Although carbohydrates are often considered flexible in comparison to the stable, secondary structures seen in proteins, experimental evidence indicates that deviations in values for these angles are often rather modest: approximately ± 30° from their equilibrium values.[13] The preference for the ϕ angle observed experimentally has been attributed to the exoanomeric effect (i.e. the stabilizing orbital overlap between the lone pair of the anomeric oxygen (Ox) with the antibonding σ* orbital of the bond between the ring oxygen and the carbon at the anomeric position (O5-C1), (see Figure 1d-e)).[29] For the ψ angle, there are no equivalent orbital interactions, so preferences are determined by steric interactions.

During the development of the ϕ and ψ CHI energy functions, density functional theory (DFT) energy calculations were performed at the B3LYP/6-31G++(2d,2p)//HF/6-31G++(2d,2p) level of theory, using minimal disaccharide models where all hydroxyl groups were replaced with hydrogens. Previous benchmarking studies recommend against the use of the B3LYP functional for conformational studies of carbohydrates.[30-31] Regardless, low energy torsions predicted by the CHI energy model showed good agreement with frequently observed glycosidic torsions from a survey of over 13,000 glycosidic linkages in the Protein Data Bank (PDB).[29] The developers of the CHI energy functions introduced a modified version of the docking program AutoDock Vina[32] which included these CHI energy functions, christened "Vina-Carb" (VC). [33] VC demonstrated improved ability to reproduce crystal structure poses of (unsulfated) carbohydrates in complex with proteins.[33]

Inspired by the development of VC, we decided to address several underlying deficits in order to improve performance in the docking of GAG structures specifically. The CHI energy functions were extended to support disaccharides containing sugars in the $^{2}S_{O}$ conformation, which are prevalent in GAGs. The CHI energy functions relevant to GAGs were recalculated using more accurate DFT methods to assess the functional/basis set dependence on the energies used to parameterize VC. Finally, an investigation into additional non-bonded potentials to model "salt-bridge" (cation-anion) and "water-bridge" (cation-water-anion) interactions was made, which may inform better predictions of the locations of charged groups in docking experiments (given GAGs contain many anionic sulfate groups). Previous modelling of GAG-protein complexes have highlighted the importance of solvent mediated interactions.[34] This was motivated by the improved outcomes of VC, which are often attributed to the inclusion of additional "pose" scoring functions. Several modifications to this program were made, which resulted in our new program, GlycoTorch Vina (GTV). To assess the effectiveness of our new parameterization, the docking performance of GTV was compared with AutoDock Vina, VC, and the commercial docking program Glide.[35] Crucially, this was assessed against a curated test set of high-quality GAG-protein crystallographic structures, with the implications of ambiguous reference data discussed below.

Methods

Analysis of experimentally determined structures

The structures of 86 complexes containing all families of GAGs, resolved using X-ray crystallography, were collected from the PDB. Structural analyses, including calculations of glycosidic torsions and Cremer-Pople coordinates[36] were undertaken using *GlycoTorch*, a python library for structural glycoinformatics written by the author (E.D.B.). A list of the PDB acquisition codes of the structures used in this analysis, as well as links to the *GlycoTorch* source code are made available in the Supporting Information. Seven disaccharide sequences (Figure 2) with ring conformations and anomeric configurations relevant to GAGs were identified, based on these analyses, and were used to inform the models explored in subsequent DFT calculations.

Structures of protein-ligand complexes containing charged groups typically observed in GAGs (i.e. sulfates, sulfamates and carboxylate moieties) within 7 Å of a basic amino acid were identified in the *BindingMoad*[37] database using the *Protein Ligand Interaction Profiler* (PLIP) python library.[38] The cartesian coordinates of the groups and amino acid sidechains were used to calculate interaction angles and distances. The following atomic centers were used to calculate distances: sulfur (sulfates/sulfamates), carbonyl carbon (carboxylate), C^{Z} (arginine), N^{Z} (lysine) and $C^{\varepsilon 1}$ (histidine) atoms (Figure 3). For interactions between carboxylate groups and arginine sidechains, the angle between the vector connecting the two carboxylate oxygens and the mean plane of the guanidinium moiety was measured. The absolute value of the sine of this angle was taken to normalize the data with respect to 3D space, an approach taken previously when statistical calculations involving angles were required.[39] Appropriate bin sizes for histograms of these distances were determined using the Freedman-Diaconis rule.[40]

Quantum mechanical calculations

Seven minimal disaccharide models were created, starting from representative crystal structures. In line with previous work by Nivedha et al., all hydroxyl groups were replaced with hydrogen atoms.[29] Initial geometry optimizations for these structures were performed at the M06-2X/6-31+G(d) level of theory,[41] including solvation energy corrections using the SMD model (solvent=water).[42] This functional was chosen based on previous benchmarking studies by Csonka, et al., which demonstrated that M06-2X/6-31+G(d) yielded geometries and energies in good agreement with CCSD(T) energies extrapolated to the complete basis set (CBS) limit for a range of carbohydrate structures.[30-31] The dihedral angles around the C1-O4 and O4-C4 bonds in the minimal models I - VII (corresponding to the ϕ and ψ glycosidic torsions, respectively) were scanned in increments of 15° over 24 steps, using a relaxed optimization approach. These minimized geometries were then optimized using the M06-2X/6-311++G(2d,2p)//SMD(water) method, with energies calculated at the same level of theory. During this second optimization step, the glycosidic torsional angles were kept fixed while other degrees of freedom were relaxed. All calculations described in this section were performed using *Gaussian16*.[43]

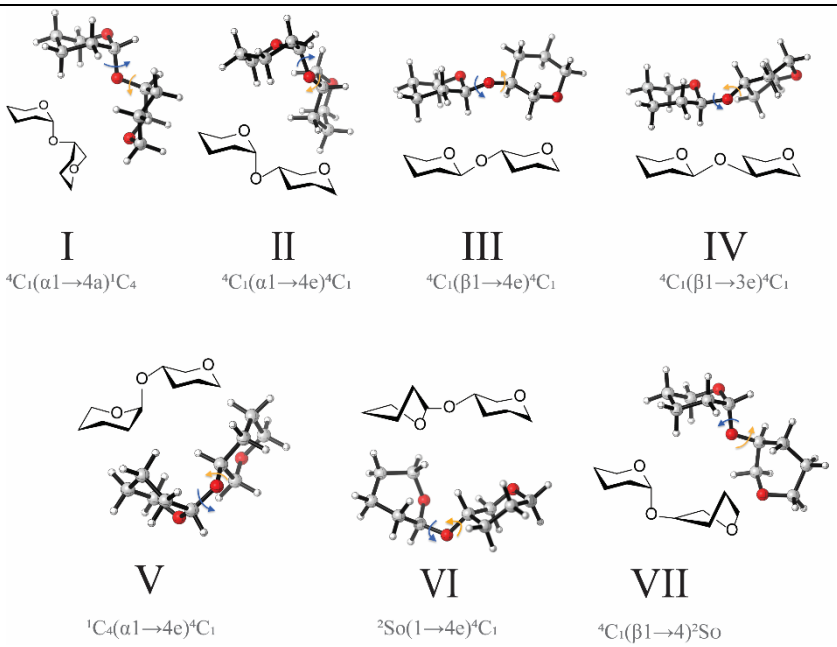

Figure 2. The seven minimal disaccharide models (I-VII) used to parameterize the energy scoring functions for the ϕ and ψ torsions, represented using the blue and orange arrows, respectively. Each model represents a common disaccharide linkage which occurs in natural GAGs.

To evaluate the importance of the anomeric interactions in stabilizing certain conformations of the glycosidic torsion,[44] we performed Natural Bonding Orbital (NBO) calculations (NBO version 3.1),[44-45] on the minimized disaccharide structures throughout the scan, also at the M06-2X/6-311++G(2d,2p) level of theory. The sum of the stabilization energies for the relevant $n_p(O5) \rightarrow \sigma^*(C1O5)$ and $n_p(O1) \rightarrow \sigma^*(C1O1)$ interactions are reported.

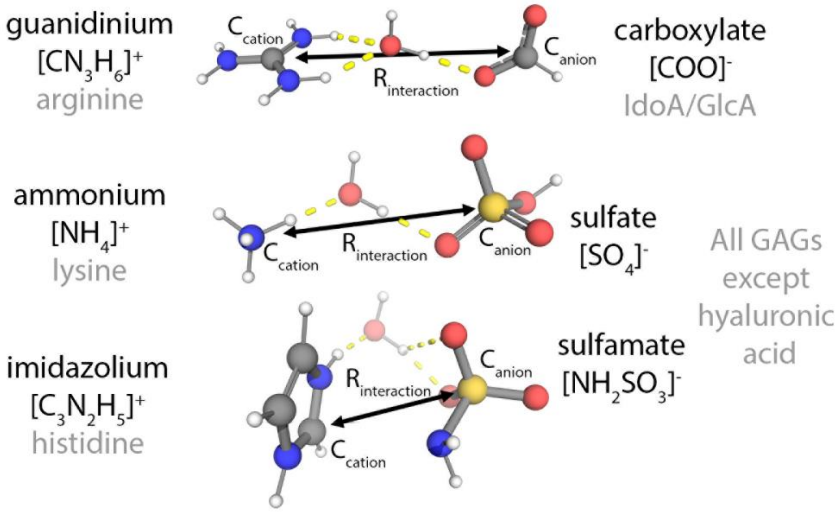

Figure 3. Examples of the systems used to create non-bonded potentials for ionic groups found to form "salt-bridges" in GAG-protein complexes. The interaction distance, $R_{interaction}$, is defined as the distance between the anion and cation centers labelled. Specifically, these were the sulfur (sulfates/sulfamates), carbonyl carbon (carboxylate), C^{Z} (arginine), N^{Z} (lysine) and $C^{\varepsilon 1}$ (histidine) atoms. Calculations with an explicit water molecule were also performed to generate non-bonded potentials for "water-bridge" systems (hydrogen bonds are shown in yellow). The relevance of each model to a biological system is also given.

To develop non-bonded interaction terms that better model GAG-protein salt-bridges, minimal models of the basic amino acids (arginine, lysine and histidine) were constructed using guanidinium, ammonium and imidazolium, respectively, and, for the anionic functional groups found in GAGs, carboxylate, sulfamate and sulfate were also used. Non-bonded interactions for each of the nine (3 × 3) cation-anion pairs were investigated. Since bridging water molecules are thought to be common in these systems, another nine complexes were constructed with a single water molecule placed between the two ions, forming a bridging hydrogen bond network. Initial structures were optimized, firstly with the MMFF94 molecular mechanics forcefield in *Avogadro*,[46] then subsequently minimized at the M06/6-31+G(d) level of theory. Here, the M06 functional was chosen due to its improved performance at predicting hydrogen bond geometries/energies, compared to M06-2X and other comparable hybrid functionals.[41] Starting from the equilibrium distance, the distance between centers ($C_{anion} - C_{cation}$) was scanned in the forward and reverse directions, with a step size of 0.1 Å. The energy of interaction between groups was calculated with the formula below:

$$\boldsymbol{E_{interaction} = E_{complex} - E_{anion} - E_{cation} - nE_{water}} \qquad (\mathbf{1})$$

where *n* is the number of water molecules.

Implementation of additional scoring functions

A series of six Gaussian curves were fit to electronic energy calculations determined in the glycosidic torsion rotational energy scans. Equation **2** was fit to the data using the least squares optimization method available in the Scipy library in the Python programming language.

$$f(x) = \sum_{i=1}^{6} a_i e^{-\frac{(x-b_i)^2}{c_i}} \quad 0 \leq x \leq 360 \qquad (2)$$

These additional energy scoring functions were added to the source code of VC within model.cpp and are evaluated using the *eval_chi* function introduced by Nivheda et al. To prevent over-counting of torsional contributions to intramolecular energies, the original CHI energy functions of Nivheda et al. were removed. Other changes to the VC source code include removal of the Best-Fit-Four-Membered-Plane (BFMP) method,[47] which has been replaced with the popular Cremer-Pople method for classifying ring conformations,[36] as well as other minor, mostly aesthetic changes in error reporting and handling. Although the BFMP method has its strengths, especially for identifying non-IUPAC ring conformations, its reliance on using idealized four-membered planes to classify structures occasionally leads it to produce non-intuitive assignments for ring systems with internal strain.[47] Indeed, many GAG crystal structures contain residues with slight deviations to non-canonical ring conformations (due to distortions caused by environmental factors) which leads to erroneous assignment. We therefore opted to use the Cremer-Pople parameters (Φ and Θ, see Figure 1c) to bin structures as either $^4C_1$, $^1C_4$ or ^2S_O based on appropriate cut-offs ($(\Theta < 40°) \rightarrow {}^4C_1$; $(\Theta > 140°) \rightarrow {}^1C_4$; $(112.5 > \Theta > 67.5) \wedge (170 > \Phi > 130) \rightarrow {}^2S_O$). Unlike VC, our program can be run directly from the executable file, without including the source code.

A major change to the VinaCarb source code has been the introduction of additional scoring functions that can be added by the user at run time. New scoring functions can be either distance based (two atoms) potentials or torsional based (four atoms) potentials, specified with the key words "FFDIST" and "FFTORS", respectively. In this way, new scoring functions can be rapidly implemented without any programming knowledge or code compilation required. A more detailed explanation of these keywords and their usage is described the Supporting Information.

System selection

Ten GAG-Protein complexes, ranging in length from tetra- to octasaccharide, were chosen for the validation set based on the following criteria: (1) the complex contained a GAG with >4 d.p., (2) the protein and ligand had comparable B-factors and, perhaps most importantly, (3) electron density maps

were available and unambiguously supported the assignment of ring conformations/orientations. Many GAG crystal structures contain high levels of disorder, especially in sections of the polysaccharide with high flexibility such as terminal residues, or residues that make few contacts with the protein. During the selection process we observed, in some cases, GAG ligands which have been extensively modelled, likely to match experimental conditions, with little supporting evidence found in the appropriate electron density maps. Our motivation for applying criteria (2) and (3) was to prevent any unnecessary bias introduced during the crystallographic refinement stage. Since our aim was to generate poses of GAGs with realistic glycosidic torsions we chose to validate our program against structures that could be unambiguously supported by crystallographic evidence. Disaccharides were excluded from the validation set. Since disaccharides are small and contain few rotational degrees of freedom, many docking programs can already dock them convincingly, as evidenced by previous studies.[48] Also, disaccharide GAGs are rarely biologically relevant in isolation from the larger polysaccharide sequence. By excluding these disaccharides, we aimed to prevent any artificial inflation of the success rate for our program.

Docking protocol

To test the efficacy of our program, we performed a series of redocking experiments using GTV. These results were also compared to the output of AutoDock Vina (ADV) and the commercial docking program, Glide. For each program, the center of the search space was defined by the geometric center of the ligand. The size of the search space was set by finding the distance between the minimum and maximum atomic positions of the ligand across each of the x, y and z axes and extending this distance by 5 Å in either direction. For ADV, VC and GTV, the "exhaustiveness" (number of independent internal docking trials), and "energy range" (cut-off to ignore poses higher in energy than the best pose) parameters were set to 32 and 12 kcal/mol, respectively. In contrast to the previous benchmarking studies involving ADV and VC,[32-33] we used a more generous value for "exhaustiveness" (32 compared to 8) in order to account for the larger conformational spaces of GAGs and to avoid any bias introduced by sampling issues. Receptor files were generated using the python script provided in AutoDock Tools. When preparing ligands with AutoDock Tools, atom types for sulfate/sulfamate oxygens are assigned incorrectly. Also, GTV and VC

requires each atom in the pyranose ring to have the correct type, based on the standard carbohydrate nomenclature (i.e. C1, C2 …), to recognize which portions of the model to apply the CHI energy functions. A script (Carbohydrate_to_PDBQT.py, download link provided in the Supporting Information) was written and used to generate ligand input files for GAGs, with correct atom typing and appropriate rotatable groups. In Glide, complexes were prepared using the Protein Preparation Wizard and receptor grid files were created using the Receptor Grid Generation Tool. The size of the grid (specifically values for the "ligand diameter midpoint box") was kept consistent with the grid sizes used in ADV and GTV. For the ligand, all anionic groups were assigned a formal charge of -1. OPLS3e atom types and charges were selected. Final scoring was done using GlideScore version SP5. Docking results were analyzed by calculating the heavy atom RMSD (Å) between the predicted (docked) pose and the native crystal structure conformation.

Results and Discussion

Analysis of empirical GAG-protein complexes with QM derived pose scoring function

Initially, we performed a meta-analysis of experimentally determined GAG crystal structures by counting the number of discrete disaccharide pairs, based on ring conformation and anomeric configuration (i.e. the orientation of the anomeric oxygen, O1, with respect to neighboring carbon atoms, for example α/β or axial (abbreviated as a)/equatorial (abbreviated as e), this nomenclature is described in Figure 1e). Although there are several hundred possible unique arrangements of the disaccharides, in our data set, seven distinct classes accounted for over 80% of the data. Importantly, approximately 15% of the disaccharide pairs contained a sugar in the $^{2}S_{O}$ conformation, which is characteristic of heparin, heparan sulfate and dermatan sulfate GAGs. Visualizing the distribution of ϕ and ψ glycosidic torsions, with respect to the distinct disaccharide pairs suggested that the distribution of angles for the $^{2}S_{O}$ model VI was most similar to systems III and V (see Figure 4a). Furthermore, these data showed good agreement with the predicted lowest-energy conformers of model GAG disaccharides I-VII (see Figure S1 – S7), which provides confidence that DFT-derived energies are suitable for the generation of the dihedral

scoring functions. Our observations are consistent with previous attempts at defining the bounds of accessible conformational space available to GAGs, such as Clerc et al., although we extend this analysis to a larger number of experimentally determined crystal structures.[49]

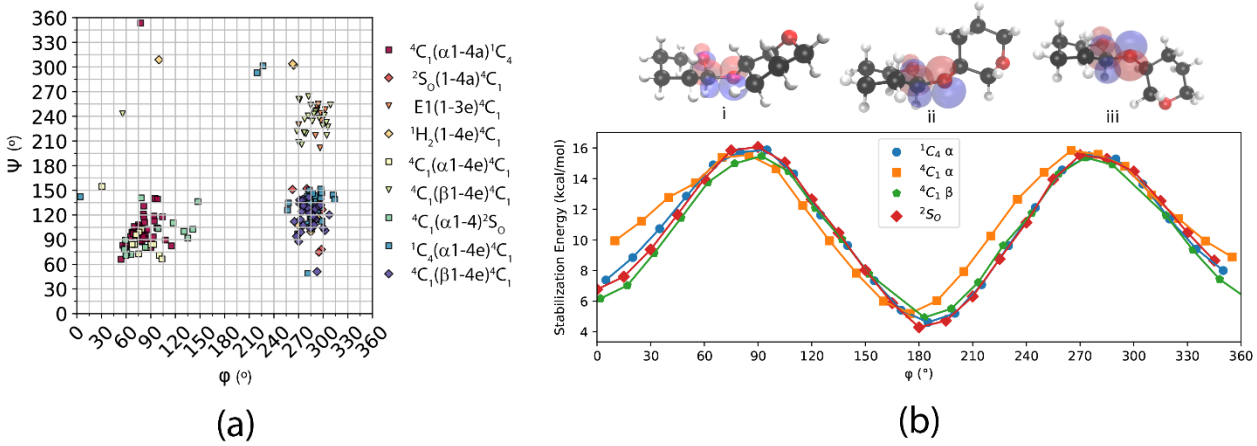

Figure 4. Common GAG conformations around the glycosidic linkage can be rationalized on the grounds of electronic and steric interactions. (a) Empirical distribution of ϕ and ψ glycosidic torsions taken from 86 GAG crystal structures surveyed from the PDB. Points are colored by each unique ring conformation/linkage geometry observed (square = α1→4 linkages, diamond = β1→4 linkages, triangle = 1→3 linkages). (b) NBO stabilization energies associated with the $n_p(O5) \rightarrow \sigma^*(C1O5)$ and $n_p(O1) \rightarrow \sigma^*(C1O1)$ orbital interactions for a variety of model disaccharides. Representative structures (i, ii, iii), with idealized natural bond orbitals, are shown for ϕ = 90°, 180°, 270°, respectively.

On consideration of the data (Figure 4a), two main populations are observed with regards to the ϕ torsional angle, centered at 90° and 270°. This bimodal distribution is due to the exoanomeric effect, a stabilizing interaction whereby the OX lone pair delocalizes into the antibonding orbital of the adjacent CX-OX bond. In order to assess this hypothesis, the contribution of the $n_p(O5) \rightarrow \sigma^*(C1O5)$ and $n_p(O1) \rightarrow \sigma^*(C1O1)$ orbital interactions were quantified using NBO analysis for each of the model systems. The stabilization energy of four representative systems (Figure 4b), where the sugar at the non-reducing end (NRE) of the disaccharide is either α/β $^4C_1$, α $^1C_4$ or ^2S_O indicate that the exoanomeric effect significantly stabilizes conformers with ϕ = ~90° or ~270° (16 kcal/mol using the second-order

perturbation theory, SOPT, method), and importantly, is largely independent of the anomeric configuration of the non-reducing sugar. On the other hand, the rotational energy barrier around the ϕ dihedral angle (which is the result of stabilizing and destabilizing interactions) is calculated to be ~6 kcal/mol for β anomers (I, II, V) and ~8 kcal/mol for α anomers (III, IV) using the M06-2X/6-311++G(2d,2p) method, which is in reasonable agreement to the barriers calculated by Nivedha, et al. For the model disaccharide VI, the sugar at the NRE is in the $^{2}S_{O}$ conformation and so the glycosidic bond is neither axial nor equatorial. For this system, the rotational energy barrier was 6.9 kcal/mol.

The preference for the ϕ angle to fall between either 90° or 270° is due to steric clashes, namely between H1 and neighboring groups on the reducing end sugar. Axial (α) and equatorial (β) anomers in the $^{4}C_{1}$ conformation prefer angles of 90° and 270°, respectively. When the conformation is flipped to the opposite chair (i.e., $^{1}C_{4}$), the preference for ϕ angle for the respective α/β anomers is also reversed. For hexoses in the $^{2}S_{O}$ conformation, since C1 is in the mean plane of the ring, again the glycosidic bond is neither axial nor equatorial, unlike for sugars in either chair conformations. The C1-O1 bond is at approximately 0° to the mean plane of the ring, closer to the ~30° angle made by β anomers in the $^{4}C_{1}$ conformation than α $^{4}C_{1}$ anomers (~90°). This agrees with empirical observations (see Figure 4a) where the distribution of ϕ angles for $^{2}S_{O}$ sugars are most comparable to β $^{4}C_{1}$ anomers and/or α $^{1}C_{4}$ anomers.

While DFT calculations were initially utilized to provide accurate ϕ and ψ torsional energy profiles for subsequent use in docking scoring functions, the data can also be used to identify possible anomalous GAG structures in the Protein Data Bank (PDB). For example, in the empirical distribution of $^{4}C_{1}$(a1-4a)$^{1}C_{4}$ disaccharides, there is one clear outlier (highlighted in Figure 5) – a disaccharide fragment from a larger structure of a heparin pentasaccharide bound to a serine protease inhibitor (Protein C inhibitor (PCI)).[50] This interaction is mediated primarily through basic residues at the surface of the protein. Concerningly though, the experimental electron density corresponding to these sidechains is inconsistent with the fitted model. In the pentasaccharide, both terminal sugars display poor electron density, which is consistent with the thermal disorder and high flexibility. The ψ angle of the disaccharide shown in Figure

5 deviates from the consensus structures and is predicted to be approximately 6 kcal/mol higher in energy, due to poor steric interactions. For this reason, the pose may be considered physically unreasonable. It is important to stress that we infer no malpractice or malicious intent on behalf of the original author, but rather use this to demonstrate the limitations of existing (force-field based) fitting models. Therefore, the CHI energy scoring functions developed by our group, and Nivedha, et. al,[29] can aid in the identification of unrealistic conformations, which will assist in the development of more accurate modelling methods (structural refinement, molecular dynamics, etc).

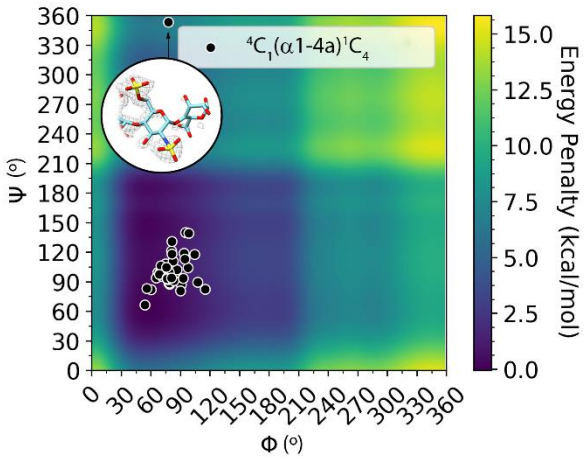

Figure 5. Empirical distribution of glycosidic torsions of GAG disaccharides with the $^4C_1(a1\text{-}4a)^1C_4$ linkage geometries taken from the crystallographic dataset. An idealized PES is also shown, colored by energy penalty applied by GlycoTorch Vina to conformations that deviate from low energy ϕ and ψ angles for the specific linkage. There is good agreement between low energy torsions and torsions observed experimentally, however, there is low specificity. An apparent outlier was identified (observed in PDB code: 3DY0) and was predicted to be a high energy conformation. The corresponding $2F_o\text{-}F_c$ electron density map, shown contoured at 1 σ, does not support the model in this region.

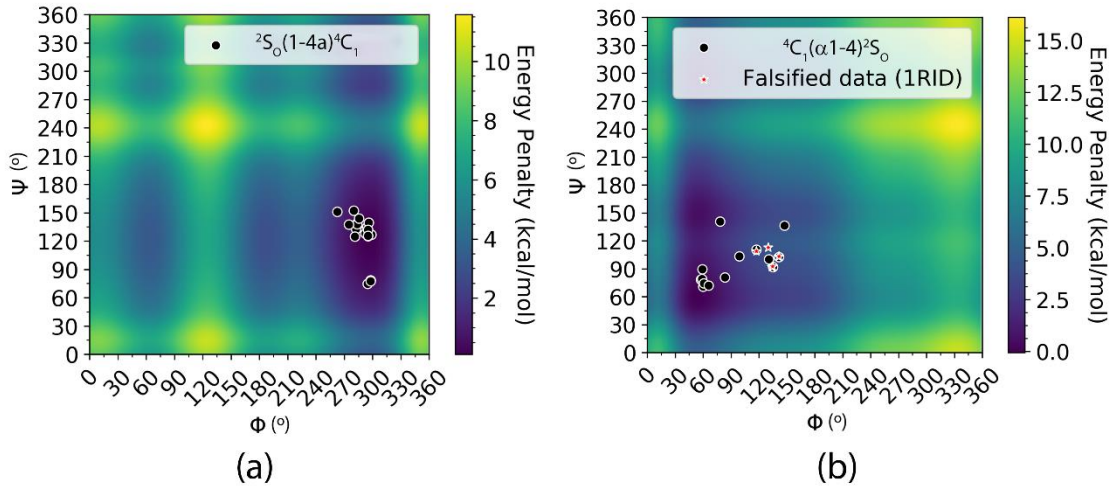

Figure 6. Empirical distribution of glycosidic torsions of GAG disaccharides with (a) $^2S_O(1\text{-}4a)^4C_1$ and (b) $^4C_1(a1\text{-}4)^2S_O$ linkage geometries taken from the crystallographic dataset. An idealized PES is also shown, colored by energy penalty applied by GlycoTorch Vina, derived from DFT calculations, applied to conformations that deviate from low energy ϕ and ψ angles for the specific linkage.

Given the importance of IdoA flexibility in determining the specificity of GAG-protein interactions, we applied the same approach in modelling low energy glycosidic torsions for sugars containing residues in the ^2S_O conformation and compared our model to experimental observations. When the sugar in the ^2S_O conformation is at the first sugar in the disaccharide sequence, from the non-reducing end, the rotational energy barriers are comparable to equivalent β–sugars in the $^4C_1$ chair conformation (approximately 4.5 kcal/mol, refer to Figures S2b and S7b). There was good agreement between torsions observed in crystallographic data and low energy torsions, as predicted by our model (See Figure 6a). A structural motif in GAG polysaccharides observed frequently in the PDB (7.1% of disaccharide fragments in our dataset) was the $^4C_1(a1\text{-}4)^2S_O$ linkage, which occurs in heparin and heparan sulfate.

The majority of glycosidic torsions observed in these disaccharides agreed with our theoretical predictions (Figure 6b). However, six disaccharide fragments (originating from PDB structures 1E0O and 1XT3) deviated significantly from suspected low energy conformations (between 3 – 6 kcal/mol higher in energy, according to our minimal models) and were also identified as outliers by the program Mogul (a knowledge-based library of molecular geometry derived from the Cambridge Structural Database). It was

also observed that a since retracted crystal structure, (PDB code 1RID), also contained this type of linkage.[51] This structure, in addition to eleven others, are suspected to be fraudulent.[52] Many of these structures, including 1RID, were used to implicate certain proteins as potential drug targets. Regrettably, two of these structures were included in a large scale virtual screening effort by IBM.[52] These specific glycosidic linkages found in 1RID were predicted to be higher in energy than the minimum energy torsions by approximately 5 kcal/mol (indicated in Figure 6b). Unfortunately, neither of the structures predicted to contain strained glycosidic torsions released electron density maps, so these models could not be compared to such primary data. This may serve as a healthy reminder that crystal structures are ultimately 'models' that should be critically assessed before being used for molecular modelling.

DFT modelling of salt-bridge interactions

In contrast to the data for ϕ and ψ angles, there was a limited agreement between theoretical low-energy distances between charged groups investigated in this report (shown in Figure 3) and the most frequent empirical distances between these groups observed in the PDB (Figure 7). While it is possible this may be due to a limitation of the method, previous work comparing QM-derived minimum energy separation distances to experimental data, such as the study of π-cation interactions by Kumar, et al.[53] and a selection of salt-bridge complexes by Kurzab, et al.,[54] have also encountered similar discrepancies. Importantly, we found that including explicit water-bridging molecules improved the agreement between experiment and theory in some systems, suggesting that implicit solvation methods alone (such as polarizable continuum model, PCM, used previously) are not sufficient to accurately model the interaction, and thereby rationalize empirical distributions of salt-bridge distances in the PDB. Theoretical and empirical data describing the systems used to model arginine-carboxylate salt-bridge, and water-bridge interactions were particularly interesting and may be used to inform common binding modes relevant to predicting GAG-protein interactions. The distribution of interaction distances between carboxylate groups and arginine side chains in the data set appeared to be bimodal, revealing two main populations (see Figure 7a), one close-range cohort of structures (centered around ~ 4.1 Å) and a second long-range cohort (centered around ~ 4.6 Å). These data were compared alongside the

angle of interaction between the guanidinium and carboxylate moieties (θ, defined in Figure 8). To normalize the distribution of these angles in relation to 3D space, the absolute value of the sine of the angle was taken. Performing this transformation, in the case of an isotropic distribution of angles, would produce an evenly distributed density of angles. This analysis of the empirical data showed that the short-range interactions favored, front-on, bidentate hydrogen bonds, with both groups being approximately coplanar ($0 < |\sin(\theta)| < 0.6$ or $0 < \theta < 40°$, depicted as I in Figure 8). In the case of the longer-range interactions, the most frequent angles of interaction appeared to be closer to 90° (depicted as II in Figure 8), which is consistent with salt-bridge interactions mediated by bridging waters.

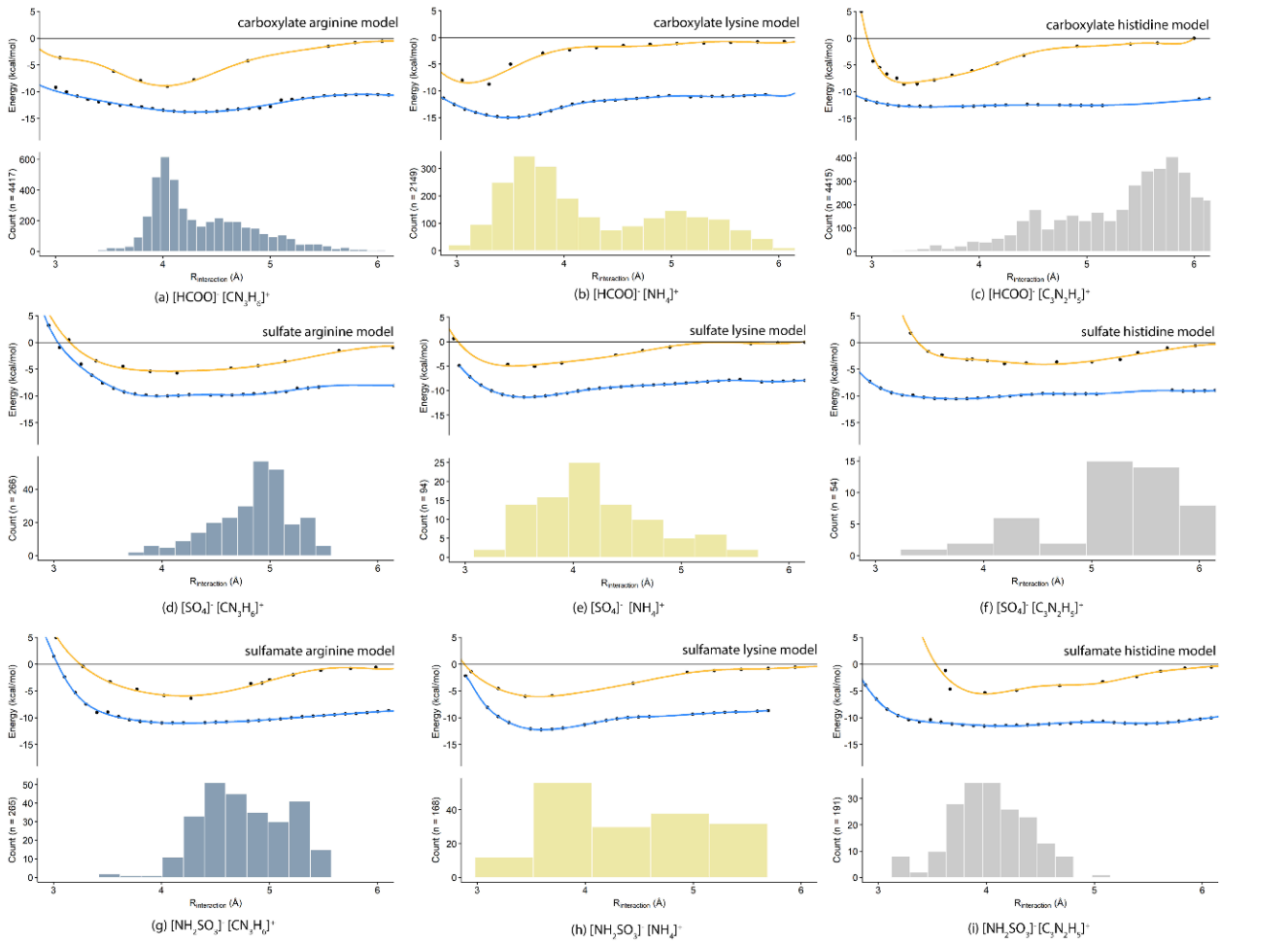

Figure 7. Interaction energies between the model salt-bridge complexes compared with the corresponding experimental distributions of interaction distances surveyed from the *BindingMoad* database. A polynomial fit to single point energies calculated at the M06-2X/6-311++G(2d,2p) level of theory are shown for the salt-bridge (orange) and water-bridge (blue) systems.

The systems used to model lysine-anion salt-bridge interactions tended to underpredict the cation-anion distance compared to the most frequently observed experimental separation distance. This was also observed by Kurzab, et. al., in a model lysine-carboxylate system where the minima was 2.9 Å vs the most common separation distance of 2.7 Å (note: these distances reported are measured between the ammonium nitrogen to the closest carboxylate oxygen atom). Interestingly, the theoretical minimum energy separation distance for a lysine – benzene model system in calculated in the study of π-cation interactions by Kumar, et al. also drastically underpredicted the most common experimentally observed distance (approximately 3.0 Å compared to 4.0 Å, note: this distance is measured between the lysine nitrogen atom and the center of the benzene ring). In some of these cases, the water-bridging model was more consistent with statistical observations made from empirical data, though was unable to replicate the bimodal distribution observed with the carboxylate anion. The maxima of the first and most frequent population was centered at ~3.6 Å, while the calculated minimum energy separation distance was ~3.3 Å. The water-bridge model for this system showed slightly better agreement, placing the minimum energy separation at 3.5 Å. Neither model was able to explain the second maximum observed, centered at 5.1 Å. The theoretical minimum energy separation for the salt-bridge and the water-bridge models of the sulfate-lysine interactions (3.6 Å and 3.5 Å, respectively) under predicted the most frequently observed empirical separation distance (~ 4.1 Å). However, conclusions drawn from this system may be limited by the low number of empirical structures taken (n = 94). Surprisingly, theoretical calculations for the equivalent sulfamate system showed better agreement with statistical observations. The most frequent separation distance was between 3.5 Å – 4.0 Å. Again, the pure salt-bridge model under predicted this distance, as the theoretical minimum energy separation for the salt-bridge and the water-bridge models were 3.4 Å and 3.6 Å, respectively.

None of the calculations involving model systems used to rationalize histidine-anion interactions were able to explain experimental distributions of atom-atom distances. This may have been caused by using the $C^{\varepsilon 1}$ of the histidine sidechain as a center to define the reaction coordinate. Although the model histidine side chain used in our calculations had two axes of symmetry (one in, and one perpendicular to, the plane

of the ring), in a biological context the histidine sidechain has less axes of symmetry (only one, in the plane of the ring). By ignoring this lack of symmetry, we do not account for anisotropic effects which may play a role in the experimental distances observed. Furthermore, histidine has two biologically relevant protonation states (doubly protonated, pK_a = 6.0, singly protonated, pK_a = 9.1).[55] The models used in this calculation assumed the model histidine to be doubly protonated, with a formal charge of +1. Since the pK_a of the cationic state is close to physiological pH, the charge of histidine is influenced substantially by neighboring amino acids and the polarity of the environment, affected in part by the location of the side chain (i.e. buried or exposed to solvent). We suspect that these ambiguities may be responsible, in part, for the poor agreement with empirical data.

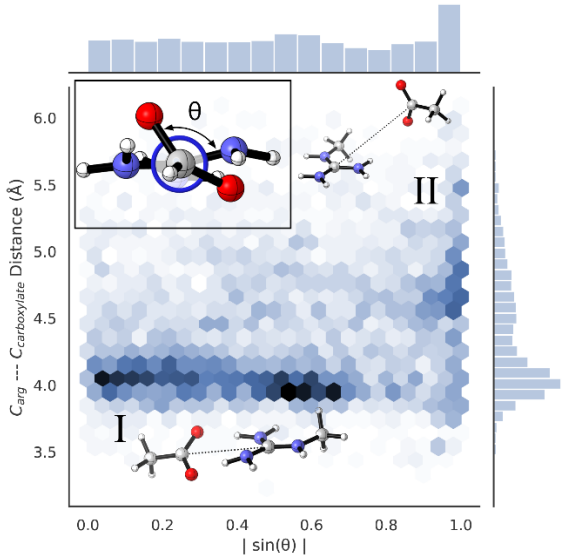

Figure 8. Analysis of empirical structures of carboxylate-arginine salt bridges reveals two common binding modes, (I) close-range, co-planar interactions and (II) long-range interactions. A 2D histogram is presented, showing the distribution of interaction distances versus the normalized distribution of angles (θ), where θ represents the angle between the mean plane of the guanidinium group and the vector that passes through each of the carboxylate oxygen atoms.

Validation of GlycoTorch Vina

In comparison to ADV and Glide, GTV demonstrated a sharp improvement when redocking ten high-quality crystal structures of GAGs (d.p. > 3) in the test set. GTV was able to successfully (i.e. average RMSD < 2 Å) reproduce eight of the structures, compared to VC (seven), ADV (six) and Glide (one) (Figure 9a). When comparing GTV to VC, over the entire test set, GTV produced poses within an average of 1.8 Å RMSD from the crystal structure, a slightly better result than VC (1.9 Å RMSD). These results are within the expected variance of the docking method, and so we are unable to assert whether GTV or VC is more accurate. A notable exception to this is that our program is parameterized to model GAG sugars in the $^{2}S_{O}$ conformation (whereas VC is not). Ultimately, we are limited by the number of available models of sufficient quality in the benchmarking set (especially those containing sugars in the $^{2}S_{O}$ conformation). In general, ADV and Glide (to a lesser extent) produced poses with some strained glycosidic torsions. ADV does not take into consideration intramolecular interactions when scoring poses generated by the search algorithm. The search algorithm implemented in Glide involves several additional steps in comparison to ADV. Unlike ADV, Glide considers intramolecular interactions in certain stages of the search algorithm, firstly, when the initial trial poses are generated and, later during the final optimization. Both programs appear to prioritize favorable intermolecular interactions at the expense of producing strained poses of GAGs. An example which illustrates this point are the results of redocking the heparin pentasaccharide to a polypeptide growth factor (PDB code: 1GMN) (Figure 9b). In the native crystal structure, the GAG sits in a shallow cleft in the surface of the protein, making key interactions with basic residues. Lysine 62 extends out from this cleft and is positioned to participate in three hydrogen bonds with sulfate moieties on sugars one, three and five (counting from the NRE). The distances between heavy atoms in these hydrogen bonds are 4.0 Å, 3.6 Å and 4.1 Å, respectively. The length of these hydrogen bonds indicate that these interactions may be mediated by solvent. The pose produced by ADV also contains multiple hydrogen bonds between sulfates (this time on sugars one, two and five) and lysine 62. In this structure, however, several glycosidic torsions are uncharacteristically strained to accommodate these interactions. Applying the pose scoring functions implemented in GTV would penalize this structure

sufficiently (> 15 kcal/mol) for it to be discarded by the search algorithm. ADV appeared to underpredict these sulfate-lysine hydrogen bond distances, which were 3.6 Å, 3.4 Å and 3.8 Å, respectively. These short distances are more typical of salt-bridge interactions. For this reason, the inclusion of softer potentials between ionic interactions might provide more realistic separations between charged groups. The accuracy of redocking also suffered in Glide due to the strength of intermolecular interactions outweighing the penalty of strained intramolecular torsions. Glide also under predicted the lysine-sulfate hydrogen bond distances, which were as short as 2.8 Å (characteristic of a strong hydrogen bond). The orientation of the third residue, IdoA2S, is highly strained to facilitate an interaction between the carboxylate of the uronic acid with lysine 64. This forces the sugar into a strained conformation, which would be heavily penalized by the GTV pose scoring functions and therefore rejected.. In the native crystal structure, this carboxylate group forms a strong, close-range salt-bridge interaction with the sidechain of arginine 73, with both groups being approximately co-planar. The length and orientation of this native interaction (4.2 Å) is comparable to that of the theoretical energy minimum calculated for the model carboxylate-arginine salt-bridge system (4.0 Å, shown in Figure 7a).

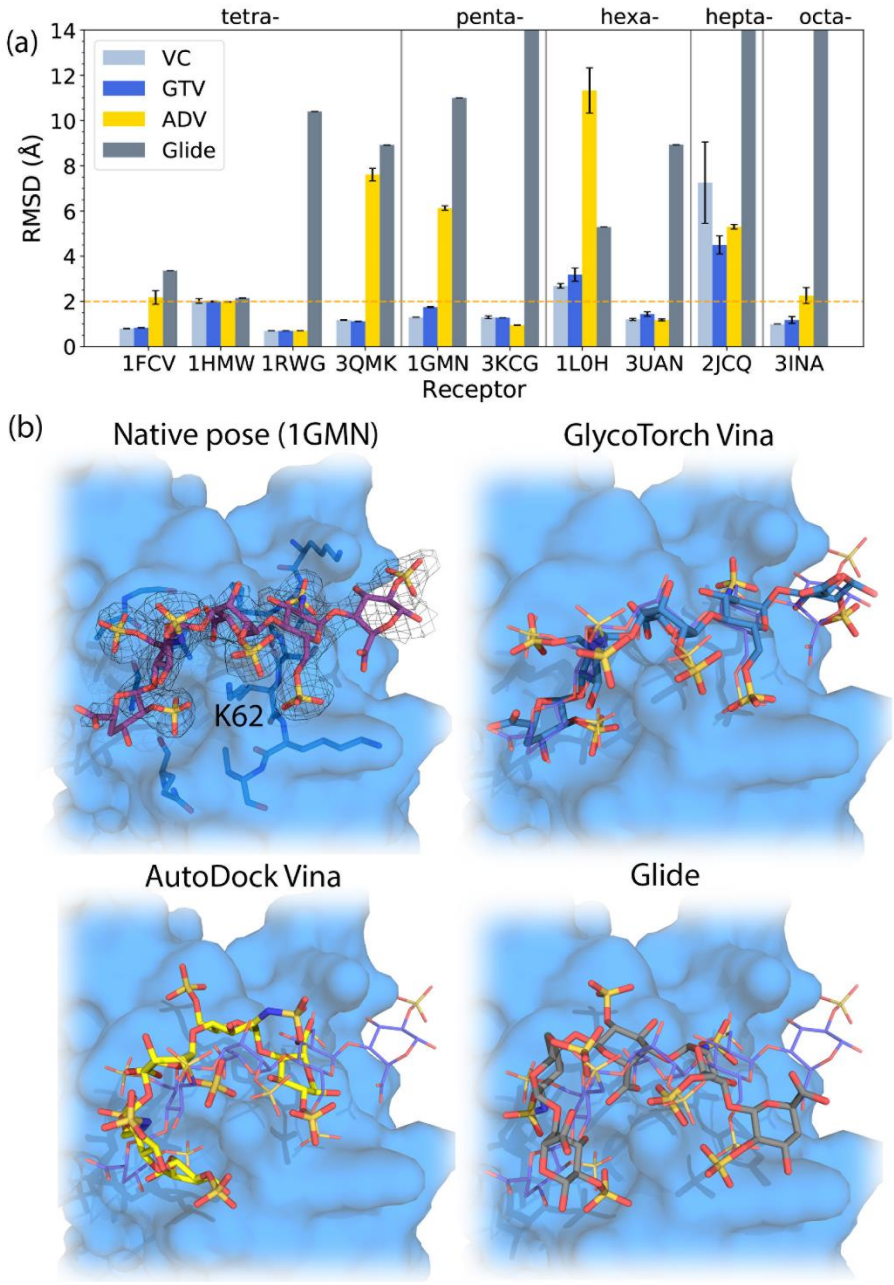

Figure 9. GlycoTorch Vina outperformed AutoDock Vina and Glide across a test set of ten well-resolved GAG-protein complexes containing large glycans (d.p. ≥ 4). (a) Comparison of docking results for the test set. The dashed orange line indicates the arbitrary 2 Å RMSD cutoff for redocking "accuracy". (b) An example comparing the top-scored pose for each of the programs to the native pose of the heparin pentasaccharide bound to a polypeptide growth factor (PDB code: 1GMN). Amino acids within 6 Å of the native ligand have been visualized. The portion of the $2F_o$-F_c electron density map, shown contoured at 1 σ, corresponding to the native ligand is also shown.

Model quality affects the reliability of docking benchmarks

The local quality of the model, inferred from relative B factors and/or fit to electron density, plays a role in the observed accuracy of docking methods. Previous efforts to benchmark the accuracy of docking programs for predicting interactions between carbohydrates and proteins have employed large test sets.[33, 48, 56] However, due to practical limitations, the local quality of these crystallographic models around the binding site are rarely assessed.[57] Instead, global metrics for model quality, such as resolution, are often used, instead, as a cut-off.[58] There is an intrinsic relationship between the amount of disorder of the ligand in the binding site and the quality of the fit between observed and modelled electron density.[59] The problems with assessing the 'accuracy' of a docking program, based on pose RMSD from the native crystal structure is two-fold, given the frequent uncertainties in crystallographic models. Firstly, areas of poor quality in the model are often due to the ligand accessing multiple, similarly low energy conformations. In these regions, the 'true' pose of the ligand might be better represented as an ensemble of states, so reporting the RMSD between the predicted and original model may not be productive. Secondly, some crystal structures, especially those containing ligands with low occupancy, are refined computationally, by optimizing the fit between the model and observed electron density. These approaches often incorporate force-field based methods which make similar assumptions to physical and knowledge-based scoring functions used in docking programs. Redocking structures that have been refined by these methods might be successful, but this success may be attributed to over-fitting the model refinement stage, rather than the accuracy of the docking program.

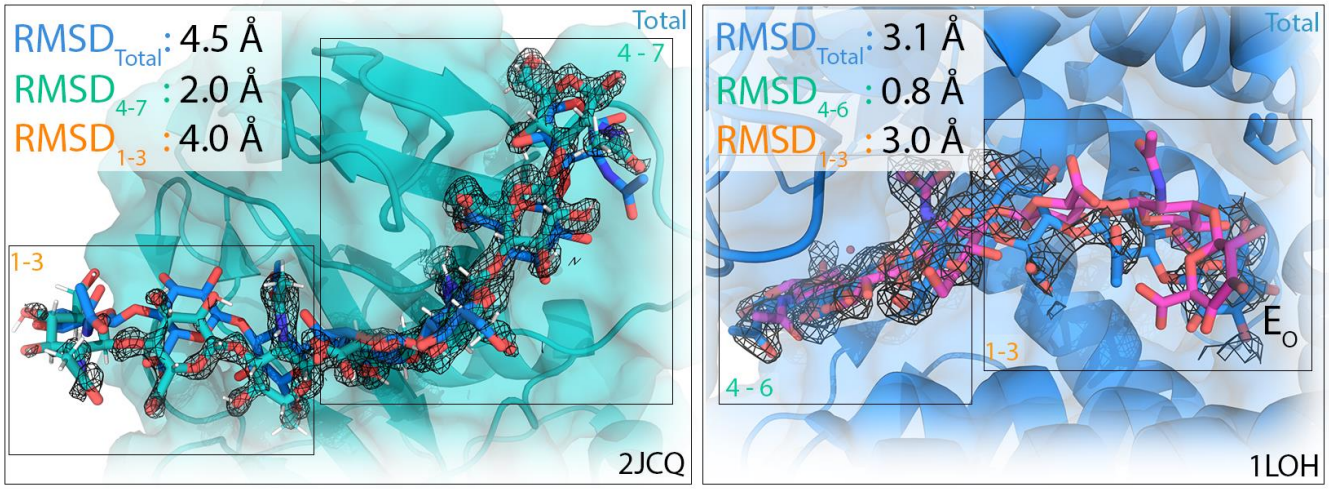

Figure 10. The local quality of the model, inferred from relative B factors or local electron density, correlates with the average RMSD of poses produced by GlycoTorch Vina. An example top-scoring pose is superimposed over the native pose for models of the cell surface receptor, Cd44 (left), and a hyaluronate lyase (right). Portions of the $2F_o$-F_c electron density maps, shown contoured at 1 σ, corresponding to the native ligands is also shown.

None of the programs tested were able to reproduce the crystallographic pose of two crystal structures (PDB codes: 2JCQ and 1LOH). These structures model the cell surface receptor, Cd44,[60] and a hyaluronate lyase,[61] with a hyaluronic acid hexa- and heptasaccharide, respectively. The former (2JCQ) was resolved with excellent resolution (1.25 Å). Despite this, portions of the model ligand (sugars one through three, counting from the NRE) display poor fit with the deposited $2F_o$-F_c electron density map (Figure 10, left). The other residues sit within a groove on the surface of the receptor, while the residues with partial electron density extend upwards and away from the primary binding site. Interestingly, when redocking this structure with GTV, the average RMSD (over ten trials) associated with the portion of the model with lower quality (4.0 Å) was significantly higher when compared to the portion of the model that better fit the experimental electron density ($RMSD_{average}$ = 2.0 Å). Similar observations can be made regarding the hyaluronate lyase structure (1LOH, Figure 10, right). The active site of this enzyme is comprised of a catalytic tunnel. The substrate (hyaluronic acid) is threaded through this passageway,

where it is subsequently cleaved. In this model, sugars four through six (counting from the NRE) are held in place inside the catalytic tunnel. These highly stabilized residues show good agreement with the reported electron density map. Sugars one through three lie in the larger catalytic cleft, positioned between the α- and β-domains of the enzyme. The quality of the model in this region suggests a substantial amount of structural disorder. Again, the average RMSD for this region (3.0 Å) was higher in comparison with the portion of the model with higher occupancy (0.8 Å). Another factor likely to influence the RMSD of the pose produced by GTV was the inclusion of a terminal sugar in the E_O envelope conformation in the structure, as disaccharide fragments containing sugars in this conformation are not supported by GTV or VC. Although the correlation between local model quality and average accuracy of pose prediction presented here is hypothetical, we hope it provides the reader with more realistic expectations when attempting to draw conclusions from molecular docking.

Conclusions

The pose scoring functions derived in this work can act as a valuable tool in understanding and predicting GAG-protein complexes. Torsional energy scoring functions developed for the ϕ and ψ linkages of sugars in the ^2S_O conformation identified low energy conformations that were consistent with experimentally determined torsions. GTV, a docking program which implements these pose scoring functions, demonstrated robust redocking accuracy, based on results evaluated from ten high-quality GAG crystal structures. Poses produced by ADV and Glide displayed glycosidic torsions that frequently deviated from consensus low energy conformations. This was attributed to the programs prioritizing the strength of intermolecular interactions over forces of internal strain. As both programs were developed with the goal of high-throughput screening of drug-like molecules, it is unsurprising that these programs failed to reproduce systems with many conformational degrees of freedom such as GAGs.

A statistical and theoretical investigation into common ionic interactions frequently observed in GAG-protein interactions suggested that water-bridges (i.e. solvent mediated salt-bridges) play an important role in influencing the structure of the complex. There was mixed agreement between statistical

observations and theoretical predictions. Multiple confounding factors influencing these complexes were identified, most of which were not captured by the simple models employed. Despite this, this work identified two common binding modes observed in arginine-carboxylate interactions (close-range coplanar and long-range water-bridge interactions) which showed good agreement with the corresponding theoretical models. These data may be useful at informing the future development of scoring functions that could be designed to better reproduce the orientation of these interactions. To this end, we have increased the scope and functionality of GTV to include user-defined pose scoring functions.

Acknowledgements

The authors acknowledge the work of Anita K. Nivedha and the Woods group, whose open-source program, VC, laid the foundations for GTV. We gratefully recognize support from the Australian Research Council (DP170104431 to V.F.), and the University of Queensland. N.S.G. acknowledges funding support from the Advance Queensland Industry Research Fellowship. Schrodinger software was purchased through QUT Science and Engineering Faculty Small Equipment grant. E.D.B. and J.M.B. also acknowledge assistance and resources from the National Computational Infrastructure (NCI Australia), a National Collaborative Research Infrastructure Strategy (NCRIS) enabled capability supported by the Australian Government and Queensland Cyber Infrastructure Foundation (QCIF) (http://www.qcif.edu.au), as well as computational resources from the University of Queensland (UQ). The high-performance computing resources provided by Queensland University of Technology (QUT) are also gratefully acknowledged. We thank Alan Mark and Elizabeth Krenske (UQ) for their insightful discussions relevant to this work.

Supporting Information

The Supporting Information is available free of charge at URL. Links to the source code and executables for Linux and Windows operating systems available at www.glycotorch.com.

Figure S1 – S7: The energy penalty functions developed in this work and their comparison to experimental glycosidic torsions; Figure S8: analysis of protein/ligand B-factors used in refining the redocking test set; Table S1: RMSD (averaged over ten trials), for each model, for each program; Table S2: Description of the models used in the docking tests in this study; links to and description of (secondary) empirical data used in this study. (PDF)

REFERENCES

1. Gandhi, N. S.; Mancera, R. L., The Structure of Glycosaminoglycans and their Interactions with Proteins. *Chem. Biol. Drug Des.* **2008,** *72* (6), 455-482.
2. Capila, I.; Linhardt, R. J., Heparin-protein interactions. *Angew. Chem. Int. Eng.* **2002,** *41*, 390-412.
3. Soares da Costa, D.; Reis, R. L.; Pashkuleva, I., Sulfation of Glycosaminoglycans and Its Implications in Human Health and Disorders. *Annu. Rev. Biomed. Eng.* **2017,** *19* (1), 1-26.
4. Swarup, V. P.; Mencio, C. P.; Hlady, V.; Kuberan, B., Sugar glues for broken neurons. *Biomol. Concepts* **2013,** *4* (3), 233-257.
5. Vlodavsky, I.; Friedmann, Y., Heparan sulfate proteoglycans Molecular properties and involvement of heparanase in cancer metastasis and angiogenesis. *J. Clin. Invest.* **2001,** *108* (3), 341-347.
6. Sasisekharan, R.; Shriver, Z.; Venkataraman, G.; Narayanasami, U., Roles of heparan-sulphate glycosaminoglycans in cancer. *Nat. Rev. Cancer* **2002,** *2* (7), 521-528.
7. Scholefield, Z.; Yates, E. A.; Wayne, G.; Amour, A.; McDowell, W.; Turnbull, J. E., Heparan sulfate regulates amyloid precursor protein processing by BACE1, the Alzheimer's beta-secretase. *J. Cell Biol.* **2003,** *163* (1), 97-107.
8. LeBrasseur, N., Heparin sustains the brain. *J. Cell Biol.* **2003,** *163* (1), 1-10.
9. Gama, C. I.; Tully, S. E.; Sotogaku, N.; Clark, P. M.; Rawat, M.; Vaidehi, N.; Goddard, W. A.; Nishi, A.; Hsieh-Wilson, L. C., Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. *Nat. Chem. Biol.* **2006,** *2* (9), 467-473.
10. Habuchi, H.; Habuchi, O.; Kimata, K., Sulfation pattern in glycosaminoglycan: Does it have a code? *Glycoconjugate J.* **2004,** *21* (1-2), 47-52.
11. Kreuger, J.; Spillmann, D.; Li, J. P.; Lindahl, U., Interactions between heparan sulfate and proteins: The concept of specificity. *J. Cell Biol.* **2006,** *174* (3), 323-327.
12. Alibay, I.; Bryce, R. A., Ring Puckering Landscapes of Glycosaminoglycan-Related Monosaccharides from Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2019,** *59* (11), 4729-4741.
13. Woods, R. J., Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* **2018,** *118* (8005-8024).
14. Gandhi, N. S.; Mancera, R. L., Can current force fields reproduce ring puckering in 2-O-sulfo-α-l-iduronic acid? A molecular dynamics simulation study. *Carbohydr. Res.* **2010,** *345* (5), 689-695.
15. van Boeckel, C. A. A.; van Aelst, S. F.; Wagenaars, G. N.; Mellema, J. R.; Paulsen, H.; Peters, T.; Pollex, A.; Sinnwell, V., Conformational analysis of synthetic heparin-like oligosaccharides containing α-L-idopyranosyluronic acid. *Recl. Trav. Chim. Pays-Bas* **1987,** *106* (1), 19-29.
16. Ferro, D. R.; Provasoli, A.; Ragazzi, M.; Torri, G.; Casu, B.; Gatti, G.; Jacquinet, J. C.; Sinaÿ, P.; Petitou, M.; Choay, J., Evidence for Conformational Equilibrium of the Sulfated L-Iduronate Residue in Heparin and in Synthetic Heparin Mono- and Oligosaccharides: NMR and Force-Field Studies. *J. Am. Chem. Soc.* **1986,** *108* (21), 6773-6778.
17. Johnson, D. J. D.; Huntington, J. A., Crystal Structure of Antithrombin in a Heparin-Bound Intermediate State. *Biochemistry* **2003,** *42* (29), 8712-8719.
18. Das, S. K.; Mallet, J.-M.; Esnault, J.; Driguez, P.-A.; Duchaussoy, P.; Sizun, P.; Herault, J.-P.; Herbert, J.-M.; Petitou, M.; Sinaÿ, P., Synthesis of Conformationally Lockedl-Iduronic Acid Derivatives:

Direct Evidence for a Critical Role of the Skew-Boat2S0 Conformer in the Activation of Antithrombin by Heparin. *Chem. Eur. J.* **2001,** *7* (22), 4821-4834.
19. Canales, A.; Angulo, J.; Ojeda, R.; Bruix, M.; Fayos, R.; Lozano, R.; Giménez-Gallego, G.; Martín-Lomas, M.; Nieto, P. M.; Jiménez-Barbero, J., Conformational Flexibility of a Synthetic Glycosylaminoglycan Bound to a Fibroblast Growth Factor. FGF-1 Recognizes Both the 1C4 and 2SO Conformations of a Bioactive Heparin-like Hexasaccharide. *J. Am. Chem. Soc.* **2005,** *127* (16), 5778-5779.
20. Opal, S. M.; Kessler, C. M.; Roemisch, J.; Knaub, S., Antithrombin, heparin, and heparan sulfate. *Crit. Care Med.* **2002,** *30* (5), S325-S331.
21. Smythe, M. A.; Priziola, J.; Dobesh, P. P.; Wirth, D.; Cuker, A.; Wittkowsky, A. K., Guidance for the practical management of the heparin anticoagulants in the treatment of venous thromboembolism. *J. Thrombosis Thrombolysis* **2016,** *41* (1), 165-186.
22. Nahain, A. A.; Ignjatovic, V.; Monagle, P.; Tsanaktsidis, J.; Ferro, V., Heparin mimetics with anticoagulant activity. *Med. Res. Rev.* **2018,** *5* (December 2017), 1582-1613.
23. Petitou, M.; van Boeckel, C. A. A., A Synthetic Antithrombin III Binding Pentasaccharide Is Now a Drug! What Comes Next? *Angew. Chem. Int. Ed.* **2004,** *43* (24), 3118-3133.
24. de Meirelles, J. L.; Nepomuceno, F. C.; Peña-García, J.; Schmidt, R. R.; Pérez-Sánchez, H.; Verli, H., Current Status of Carbohydrates Information in the Protein Data Bank. *J. Chem. Inf. Model* **2020,** *60* (2), 684-699.
25. Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R., Principles of docking: An overview of search algorithms and a guide to scoring functions. *Prot. Struct. Funct. Gen.* **2002,** *47*, 409-443.
26. Comeau, S. R.; Kozakov, D.; Brenke, R.; Shen, Y.; Beglov, D.; Vajda, S., ClusPro: Performance in CAPRI rounds 6-11 and the new server. *Prot. Struct. Funct. Bioinf.* **2007,** *69* (4), 781-785.
27. Griffith, A. R.; Rogers, C. J.; Miller, G. M.; Abrol, R.; Hsieh-Wilson, L. C.; Goddard, W. A., Predicting glycosaminoglycan surface protein interactions and implications for studying axonal growth. *Proc. Nat. Acad. Sci. USA* **2017,** *114* (52), 13697-13702.
28. Samsonov, S. A.; Zacharias, M.; Chauvot de Beauchene, I., Modeling large protein–glycosaminoglycan complexes using a fragment-based approach. *J. Comput. Chem.* **2019,** *40* (14), 1429-1439.
29. Nivedha, A. K.; Makeneni, S.; Foley, B. L.; Tessier, M. B.; Woods, R. J., Carbohydrate Docking: Sorting the Wheat from the Chaff. *J. Comput. Chem.* **2015,** *35* (7), 526-539.
30. Csonka, G. I.; Kaminsky, J., Accurate conformational energy differences of carbohydrates: A complete basis set extrapolation. *J. Chem. Theory Comput.* **2011,** *7*, 988-997.
31. Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A., Evaluation of Density Functionals and Basis Sets for Carbohydrates. *J. Chem. Theory Comput.* **2009,** *5*, 679-692.
32. Trott, O.; Olson, A. J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010,** *31*, 455-461.
33. Nivedha, A. K.; Thieker, D. F.; Makeneni, S.; Hu, H.; Woods, R. J., Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *J. Chem. Theory Comput.* **2016,** *12* (2), 892-901.
34. Sarkar, A.; Yu, W.; Desai, U. R.; MacKerell, A. D.; Mosier, P. D., Estimating glycosaminoglycan–protein interaction affinity: water dominates the specific antithrombin–heparin interaction. *Glycobiology* **2016,** *26* (10), 1041-1047.
35. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004,** *47* (7), 1739-1749.
36. Cremer, D.; Pople, J. A., General definition of ring puckering coordinates. *J. Am. Chem. Soc.* **1975,** *97* (6), 1354-1358.
37. Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A., Binding MOAD (Mother Of All Databases). *Proteins: Struct., Funct., Bioinf.* **2005,** *60* (3), 333-340.

38. Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M., PLIP: fully automated protein--ligand interaction profiler. *Nucleic Acids Res.* **2015,** *43* (W1), 443-7.
39. Lee, T.; Sanzogni, A.; Zhangzhou, N.; Burn, P. L.; Mark, A. E., Morphology of a Bulk Heterojunction Photovoltaic Cell with Low Donor Concentration. *ACS Appl. Mater. Interfaces* **2018,** *10* (38), 32413-32419.
40. Freedman, D.; Diaconis, P., On the Histogram as a Density Estimator: L2 Theory. *Probab. Theory Relat. Fields* **1981,** *57*, 453-476.
41. Zhao, Y.; Truhlar, D. G.; Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals and inorganometallic chemistry and for noncovalent interactions. *Theor. Chem. Acc.* **2008,** *120*, 215-241.
42. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G., Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009,** *113* (18), 6378-6396.
43. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J., Gaussian 16 Revision B.01. 2016.
44. Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F., NBO.
45. Reed, A. E.; Weinstock, R. B.; Weinhold, F., Natural population analysis. *J. Chem. Phys.* **1985,** *83* (2), 735-746.
46. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminf.* **2012,** *4* (1), 17-17.
47. Makeneni, S.; Foley, B. L.; Woods, R. J., BFMP: A method for discretizing and visualizing pyranose conformations. *J. Chem. Theory Comput.* **2014,** *54*, 2744–2750-2744–2750.
48. Samsonov, S. A.; Pisabarro, M. T., Computational analysis of interactions in structurally available protein–glycosaminoglycan complexes. *Glycobiology* **2016,** *26* (8), 850-861.
49. Clerc, O.; Mariethoz, J.; Rivet, A.; Lisacek, F.; Pérez, S.; Ricard-Blum, S., A pipeline to translate glycosaminoglycan sequences into 3D models. Application to the exploration of glycosaminoglycan conformational space. *Glycobiology* **2018,** *29*, 36-44.
50. Li, W.; Huntington, J. A., The Heparin Binding Site of Protein C Inhibitor Is Protease-dependent. *J. Biol. Chem.* **2008,** *283* (51), 36039-36045.
51. Retraction for Ganesh et al., Structure of vaccinia complement protein in complex with heparin and potential implications for complement regulation. *Proc. Nat. Acad. Sci. USA* **2018,** *115* (29), E6965.
52. Borrell, B., Fraud rocks protein community. *Nature* **2009,** *462* (7276), 970-970.
53. Kumar, K.; Woo, S. M.; Thomas, S.; Cortopassi, W. A.; Duarte, F.; Paton, R. S., Cation-π interactions in protein-ligand binding: theory and data-mining reveal different roles for lysine and arginine. *Chem. Sci.* **2018,** *9*, 2655-2665.
54. Kurczab, R.; Paweł; Liwa, S.; Rataj, K.; Kafel, R.; Bojarski, A. J., Salt Bridge in Ligand−Protein Complexes--Systematic Theoretical and Statistical Investigations. *J. Chem. Inf. Model* **2018,** *58*, 2224-2238.

55. Li, S.; Hong, M., Protonation, tautomerization, and rotameric structure of histidine: a comprehensive study by magic-angle-spinning solid-state NMR. *J. Am. Chem. Soc.* **2011,** *133* (5), 1534-1544.
56. Kerzmann, A.; Fuhrmann, J.; Kohlbacher, O.; Neumann, D., BALLDock/SLICK: A new method for protein-carbohydrate docking. *J. Chem. Theory Comput.* **2008,** *8*, 1616-1625.
57. Liebeschuetz, J.; Hennemann, J.; Olsson, T.; Groom, C. R., The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *J Comput Aided Mol Des* **2012,** *26* (2), 169-183.
58. Bordogna, A.; Pandini, A.; Bonati, L., Predicting the accuracy of protein-ligand docking on homology models. *J. Comput. Chem.* **2011,** *32* (1), 81-98.
59. Deller, M. C.; Rupp, B., Models of protein-ligand crystal structures: trust, but verify. *J Comput Aided Mol Des* **2015,** *29* (9), 817-836.
60. Banerji, S.; Wright, A. J.; Noble, M.; Mahoney, D. J.; Campbell, I. D.; Day, A. J.; Jackson, D. G., Structures of the Cd44–hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. *Nat. Struct. Mol. Biol.* **2007,** *14* (3), 234-239.
61. Li, S.; Jedrzejas, M. J., Hyaluronan binding and degradation by Streptococcus agalactiae hyaluronate lyase. *J. Biol. Chem.* **2001,** *276* (44), 41407-16.