Chapter 7: Computational Modeling of Glycosaminoglycans

Parts of this chapter are from Clark, P.M., Rogers, C.J., Tully, S.E., Garcia, K.C., Abrol, R., Goddard, W.A. & Hsieh-Wilson, L.C. Microarray and Computational Approaches to Understanding Glycosaminoglycan-Growth Factor Interactions. *In preparation* (2010).

Glycosaminoglycans (GAGs) are sulfated linear polysaccharides that are important in neuronal development, viral invasion, and cancer. Recent work in our lab has shown that chondroitin sulfate (CS) tetrasaccharides, a type of GAG, are able to promote neuronal outgrowth in a manner that depends on the pattern of their sulfation. Here, we use computational approaches to better understand how the CS sulfation patterns affect their activity. We modeled the solution structure of CS-A, CS-C, CS-E, and CS-R and found that each CS tetrasaccharide favors a distinct set of torsion angles and presents a unique electrostatic surface. We further employed computational docking algorithms to determine the CS-E binding sites on a variety of proteins, including BDNF, NGF, and TNF. We found that CS-E binds to a general CS-E binding site characterized by two closely placed basic amino acids and a more distant third basic amino acid. Based on the modeled CS-E binding sites, we predict that CS-E stabilizes the interaction between certain NGF family of neurotrophins and their Trk receptors, results that were supported using carbohydrate microarrays.

Introduction

Glycosaminoglycans (GAGs) are a set of diverse sulfated carbohydrates that are important in numerous biological processes including neuronal development[1,2], angiogenesis[3], and viral invasion[4]. They have also been implicated in a number of diseases including cancer[5], spinal cord injury[6,7], and Alzheimer's disease[8]. GAGs are linear polysaccharides that are composed of repeating disaccharide units. These linear polysaccharides are usually found on the cell surface or in the extracellular matrix and are

attached to a protein core as a post-translational modification[9]. Different GAGs have been shown to bind to a wide variety of proteins, and it is through these interactions that the GAGs have their effects[10].

There are multiple classes of GAGs, with the two most heavily studied being heparin sulfate (HS) or heparin and chondroitin sulfate (CS). HS and heparin consists of repeating units of D-glucosamine and D-glucuronic acid (GlcA) or L-iduronic acid whereas CS consists of repeating units of N-acetylglucosamine (GlcN) and GlcA. Within the linear GAG polysaccharide, any free hydroxyl group can be sulfated[11], leading to a diverse combination of sulfate motifs, and the pattern of sulfation uniquely identifies each disaccharide unit[9]. For example, Chondroitin Sulfate A (CS-A) consists of a CS disaccharide with a sulfate group on C-4 position of the GlcN residue whereas Chondroitin Sulfate C (CS-C) consists of a CS disaccharide with a sulfate group on the C-6 position of the GlcN residue. Recent work in our lab has begun to show that the molecular level activity of these GAGs depends intimately on their sulfation pattern. Tetrasaccharides of CS-E have been shown to produce neurite outgrowth whereas tetrasaccharides of CS-C or CS-A do not[12]. Similarly CS-E tetrasaccharides have been shown to inhibit the interaction between Tumor Necrosis Factor α (TNF- α) and Tumor Necrosis Factor Receptor 1 (TNF-R1)[13].

Because GAGs control their activity through intricately positioned molecular interactions, they have interested structural biologists who want to study how these sulfate groups affect CS or HS structure and their interactions with different protein targets[14-17]. Crystallographers performed some of the early studies on the molecular level interactions between GAGs and their binding partners. Doug Rees and his group

discovered the crystal structures of heparin bound to fibroblast growth factor 1 and 2 (FGF-1 and FGF-2)[18,19]. More recent studies have focused on the ternary complex between FGF-1, heparin, and its receptor[20], and FGF-2, heparin, and its receptor[21] as well as heparin bound to other proteins including antithrombin III[22], annexin V[23], and annexin II[24].

These crystallography studies point to some common themes in GAG–protein interactions[25]. Confirming experimental evidence, the structures indicate that the positions of the GAG sulfate groups are critical for specific interactions with key lysine and arginine residues on the protein surface. Yet in those cases where the GAGs bind to a protein monomer, all of the sulfate groups on the GAG are not positioned to interact with the protein, but rather many of the GAG sulfate groups are positioned out into solution where they interact with salts and water. Furthermore, unlike small molecule–protein interactions, which usually occur in a deep pocket of the protein, GAG–protein interactions occur much closer to the surface of the protein and in shallow pockets.

Computational chemists have also studied GAGs and their interactions with proteins. Computational approaches are particularly useful in this field as they are able to elucidate molecular details that would be otherwise difficult to obtain experimentally. Some of the first computational studies looked at the structure of GAGs in solution. Perez et al.[26] used molecular dynamics to look at the lowest energy torsion angles between the GlcA monomer and the GlcN monomer for CS-A and CS-C. Similarly Mulloy et al. employed molecular modeling, in combination with limited NMR data, to investigate the conformation of heparin in solution[27]. These molecular models indicated

that CS and heparin form repeating helical chains in solution, and that the degree of rotation of these chains depends on the sulfation pattern.

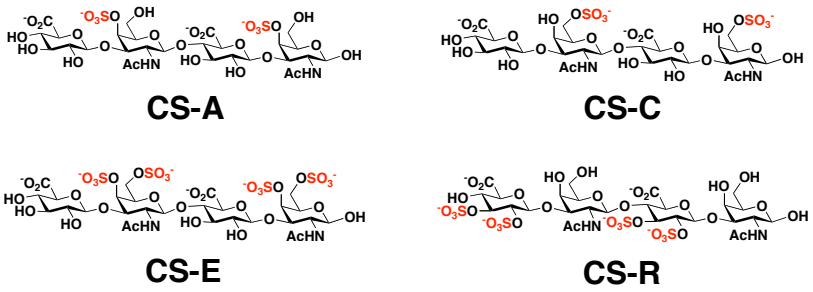

Figure 1: CS-A, CS-C, CS-E, and CS-R tetrasaccharides

Recently a number of methods have been reported for determining the binding site of heparin to different proteins[28-31]. Most of these methods work by identifying potential binding sites around the protein and sampling each of these sites with a heparin saccharide to assay which site affords the lowest complex energy or highest surface complementarities. However, it is not known whether similar methods could be used to model the relationship of less highly charged CS with proteins. Additionally how these methods could be used to investigate larger GAG protein ternary complexes remains unknown. New validated methods for determining the interactions between CS and individual proteins or larger protein complexes are needed.

Here, I describe work I have done to elucidate the solution structures of CS-A, CS-C, CS-E, and an unnatural CS motif, CS-R tetrasaccharides (**Fig. 1**). These tetrasaccharides have been previously synthesized in our laboratory, and they have been shown to have different biological activity from one another. Next, I describe a computational method I developed for determining chondroitin sulfate binding sites on proteins and show that it correctly predicts the binding sites of heparin on FGF-1 and FGF-2, as well as the likely binding site of CS-E on midkine. Finally, I use this method

to predict the CS-E binding sites on a number of proteins. These binding sites demonstrate a common CS-E binding motif and predict a role for CS-E in stabilizing complexes between neurotrophins and their receptors.

Results and Discussion

Solution Structures of CS-A, CS-C, CS-E, and CS-R. To understand how the different chondroitin sulfate molecules exert their unique effects, we chose to first model the solution structure of the CS-A, CS-C, CS-E, and CS-R tetrasaccharides. By examining the CS solution structures, we reasoned that we could see how the sulfate groups could affect the structure. We build each of the tetrasaccharides (CS-A, CS-C, CS-E, and CS-R) into the Cerius2 program (Accelrys Inc.), charged them, and minimized them in vacuum within the confines on the Dreiding force field, which had to be first modified to accommodate the sulfate groups. Although each tetrasaccharide carries a formal charge of -4 or -6, we chose to afford each atom of the tetrasaccharide with partial charges but to keep the tetrasaccharide overall neutral. This was done to account for the fact that actual tetrasaccharides are in polarizable water with counter-ions that would dampen the charges of the sulfate and carboxyl groups[32].

Once the tetrasaccharides had been built and minimized, each was subject to a Boltzmann jump algorithm to create 1000 different conformations. A Boltzmann jump works by taking the original conformation of a molecule and rotating a specific set of torsions to create a new conformation. If the new conformation is lower in energy than the original conformation, the Boltzmann jump algorithm accepts the new conformation. If the new conformation has a higher energy than the original conformation, then the

Boltzmann jump algorithm accepts the new conformation at a probability of exp(-ΔE/RT) where ΔE is the difference in energy between the old and the new conformation. The Boltzmann jump algorithm continues to follow this method until it reaches a predefined number of conformations. In our case that limit was 1000 conformations.

Once the 1000 conformations of a given tetrasaccharide had been created, these conformations were sorted into five groups based upon the RMSD to each other. The five groups were chosen in such a way as to maximize the average RMSD between each group. This was done to ensure that over the next steps, a diverse set of structures would be explored to enhance the probability that the global minimum energy structure was indeed reached. We scored the energy of all of the structures in each of the five groups and chose the top two lowest-energy structures from each of the groups to bring on to the next step. We then took each of these ten structures, immersed them in a water box, and ran 300 ps of molecular dynamics on each of the structures to allow them to find their minimum energy structure. The structures from the last 100 ps of a given run were averaged and the energy of this average structure was calculated. We used the lowest-energy structure among these ten to represent the predicted solution structure for each tetrasaccharide (**Fig. 2**).

We found that each CS tetrasaccharide favors a distinct set of torsion angles and presents a unique electrostatic and van der Waals surface for interaction with proteins. Whereas the negatively charged sulfate and carboxylate groups on CS-C point toward either the top or bottom face of the molecule, as oriented in Figure 2a, the same charges on CS-A point in several different directions. Similarly, although CS-E and CS-R have the same number of sulfate groups, the relative orientation of these groups along the

carbohydrate backbone leads to distinctly different predicted solution structures. Whereas the CS-R tetrasaccharide has the sulfate groups distributed along several faces of the molecule, the CS-E tetrasaccharide presents all four sulfate groups along a single face, which may position the groups to interact with basic residues characteristic of glycosaminoglycan binding sites on proteins[25].

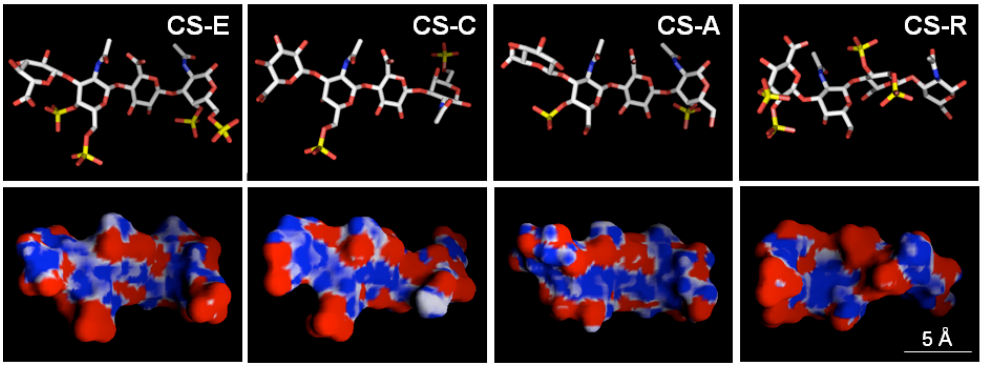

Figure 2: *Top:* The lowest energy structures of CS-A, CS-C, CS-E, and CS-R tetrasaccharides. *Bottom:* Electrostatic representations of these structures

Predicting CS Binding Sites on Proteins. CS interacts with a variety of different proteins, including VAR2CSA[33], TNF-α[13], BDNF[12], and NGF[34]. Thus having determined the solution structure of CS molecules, we next wanted to investigate the CS binding sites on these and other proteins.

Method Development and Confirmation. The program ScanBindSite has previously been employed to correctly predict the binding sites of small molecules into proteins[35,36]. Thus we chose to use this program to investigate CS binding sites on proteins. The input for ScanBindSite is a protein and a ligand file. ScanBindSite calculates the molecular surface of the protein and then determines the surface innervations from the negative image of the molecular surface. These innervations are represented by spheres and are

grouped into potential binding sites. ScanBindSite then maps the ligand atoms onto the spheres in each of these binding sites and calculates the energy of that docked conformation. At the end these energies are tabulated and can be used to determine which of the potential regions is the likely site for the ligand to bind.

To test whether we could use ScanBindSite to predict the binding sites of GAG tetrasaccharides, we first predicted the GAG binding sites on structures for which the GAG binding sites are known. We improved and then validated the method using two heparin–protein co-crystal structures, heparin binding to the FGF-2 monomer and heparin binding to the FGF-1 dimer[37,38], and two domains of the protein VAR2CSA, DBL3X and DBL6ε, for which the CS-A binding site has been proposed by mutagenesis[33]. We reasoned that these would be good structures to test our approach as they represent the interaction between charged GAG polysaccharides and proteins, which is similar to the system we were interested in exploring.

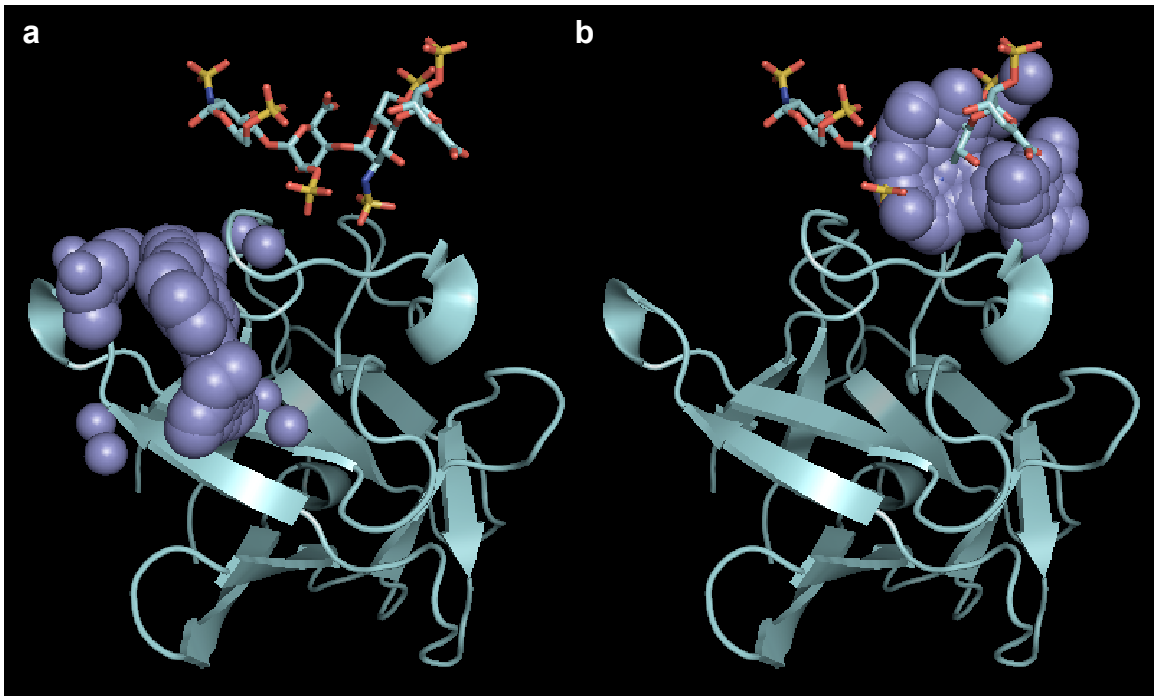

Figure 3: a, Predicted heparin binding site of FGF-2 from 1BFB using default ScanBindSite parameters. **b,** Predicted heparin binding site of FGF-2 from 1BFB using modified ScanBindSite parameters

ScanBindSite requires a number of input parameters that can be optimized for a given application. To determine the correct parameters to use for predicting CS binding sites on proteins, we tested which set of parameters best predicted the binding site of the heparin tetrasaccharide onto the surface of FGF-2. Initially, we found that the default ScanBindSite parameters failed to predict the heparin binding site on FGF-2 (**Fig. 3a**). Further analysis indicated that ScanBindSite failed to even identify the heparin binding site as a potential binding site, instead favoring more innervated regions of the protein. We also found that the potential binding sites determined by the program were much smaller than the size of a tetrasaccharide. To expand the potential binding sites determined by the ScanBindSite program, we changed the *radmax* parameter from 4.0 to

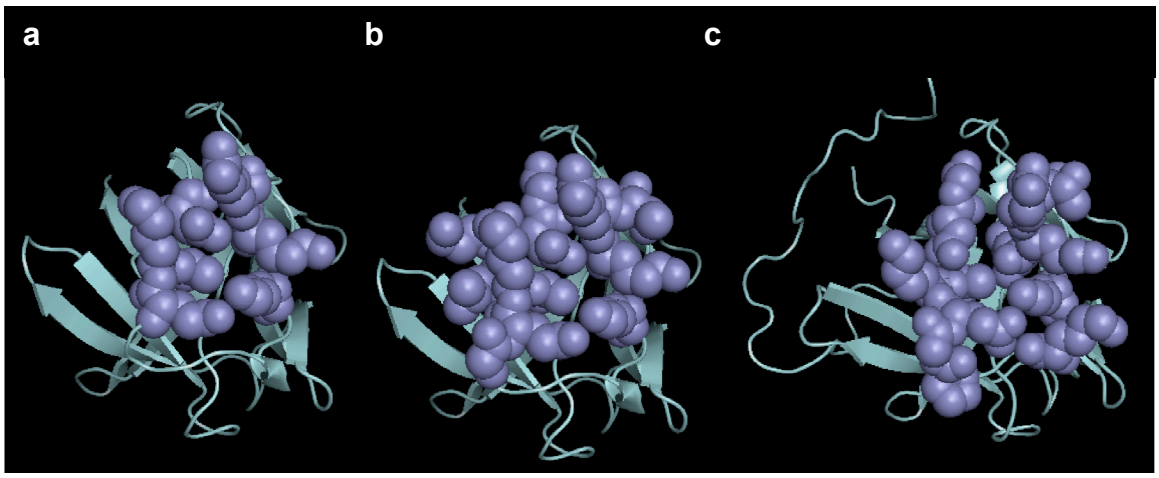

Figure 4: a, Heparin binding site on FGF-2 (from 1BFB). **b,** Predicted heparin binding site (from 1BFB crystal structure). **c,** Predicted heparin binding site (from 1BLA crystal structure).

5.0, which favors the formation of larger binding regions and the *dotlim* parameter from 0.0 to -0.5, which favors the formation of potential binding sites within the flatter surface regions of the protein that characterize GAG binding sites. Using these new parameters on the same heparin–FGF-2 systems, we correctly determined the heparin binding site of FGF-2 (**Fig. 3b**). Further optimization of these parameters by changing the *radmax* to 6.0 and the *dotlim* to -0.75 and -1.0 gave worse results.

Residues within four angstroms of heparin in the 1BFB crystal structure	**Residues within four angstroms of heparin in the computationally determined binding site from 1BFB**	**Residues within four angstroms of heparin in the computationally determined site from 1BLA***
	Lys 2 7	
Asn 28	Asn 28	Asn 28
	Gly 29	Gly 29
		Leu 1 1 9
Lys 1 2 0	Lys 1 2 0	Lys 1 2 0
Arg 1 2 1	Arg 1 2 1	Arg 1 2 1
	Thr 1 2 2	Thr 1 2 2
Lys 1 2 6	Lys 1 2 6	Lys 1 2 6
		Lys 1 3 0
		Pro 133
	Gly 134	Gly 134
Glcn 135	Gln 1 3 5	Gln 1 3 5
	Lys 1 3 6	Lys 1 3 6
Ala 1 3 7	Ala 1 3 7	Ala 1 3 7
	Leu 1 3 9	

Table 1: Residues that interact with heparin from the 1BFB crystal structure and in the predicted heparin binding site from FGF-2 in the 1BFB and 1BLA crystal structure. * Eight was subtracted from each residue number to make the 1BLA residue numbers align with the 1BFB residue numbers.

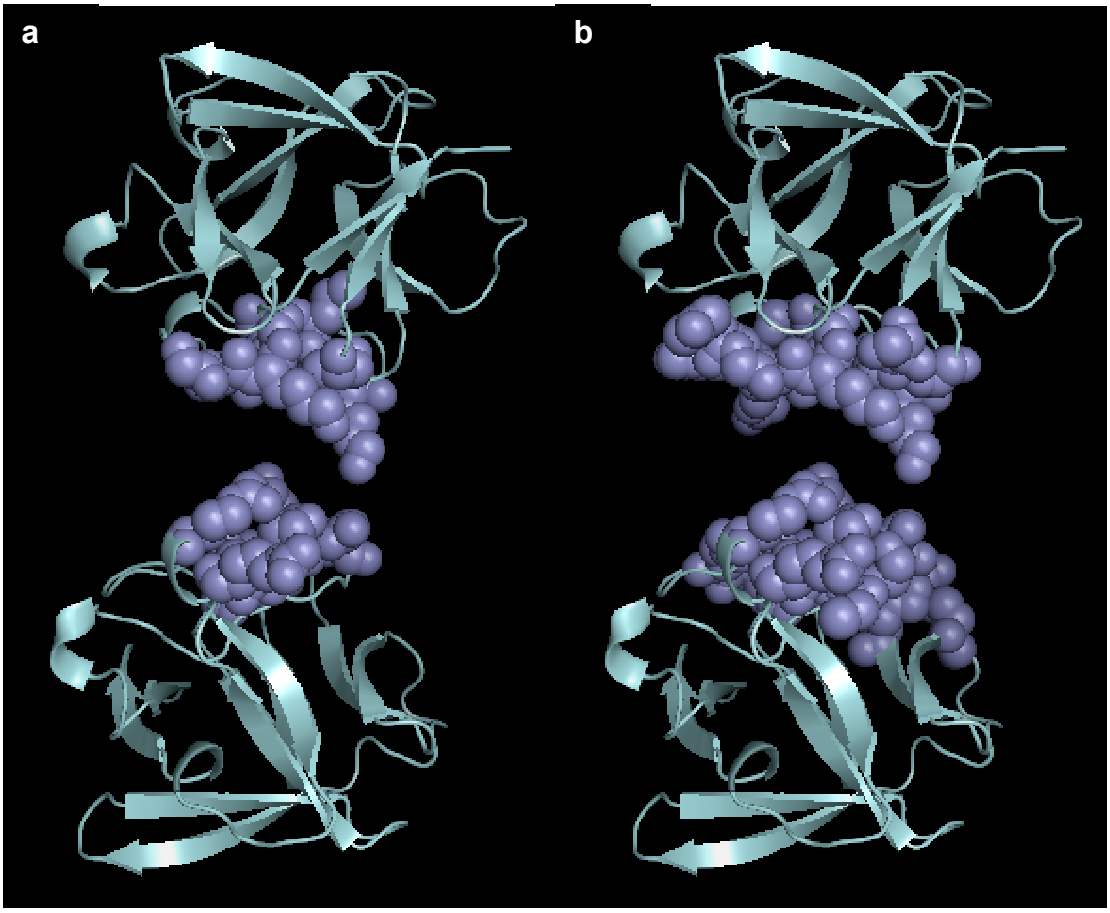

Figure 5: a, Heparin binding site on FGF-1 (from 2AXM). **b,** Predicted heparin binding site

We next wanted to investigate whether the protein residues with which the GAG interacts could be extracted from the calculated binding sites. To do this, we examined the five lowest-energy heparin–FGF-2 structures within the calculated binding site and determined which residues interact with the docked heparin molecules in these structures. We found that our models predicted all of the residues that heparin interacts with in the crystal structure while predicting a limited number of extra residues (**Fig. 4a, b, Table 1**). Furthermore to determine whether the GAG binding site could be correctly predicted from the apoprotein as well as the co-crystal structure, the binding site of heparin was predicted from a crystal structure of the FGF-2 apoprotein, which differs from the heparin

–FGF-2 co-crystal structure by an RMSD of 1.172 angstroms. We found that the calculated binding site of heparin on the FGF-2 apoprotein structure was similar to the binding site calculated from the co-crystal structure (**Fig. 4c**). Furthermore this binding site correctly predicted all of the residues that interact with heparin in the co-crystal structure while predicting few extra residues with which heparin does not interact (**Table 1**).

To test whether these modified parameters worked more generally for predicting glycosaminoglycan binding sites, we predicted the heparin binding site on FGF-1. No crystal structure of a heparin tetrasaccharide bound to FGF-1 exists, so we extracted the central heparin tetrasaccharide from the FGF-1–heparin hexasaccharide crystal structure. We then used this tetrasaccharide and the FGF-1 structure to predict the heparin binding site on FGF-1 (**Fig. 5a, b**). Again we successfully identified the heparin binding site on FGF-1 and predicted all but one of the amino acids with which the heparin tetrasaccharide interacts in the crystal structure (**Table 2**). This suggests that the modified ScanBindSite program could successfully predict GAG binding site on proteins.

Finally, we wanted to determine whether our approach could successfully predict the binding sites of CS on proteins. CS-A binds to two domains on the protein VAR2CSA, DBL3X and DBL6ε[33]. Crystal structures of these domains are available and previous work has proposed the CS-A binding sites by mutagenesis[33]. We predicted the CS-A binding site on DBL3X and DBL6ε using ScanBindSite (**Fig. 6, Table 3, 4**). Excitingly, the predicted CS-A binding site on DBL3X contains seven of the eight basic

Residues within four angstroms of heparin in the 2AXM crystal structure	Residues within four angstroms of heparin in the predicted binding site from 2AXM
	Ser 17, Chain A
Asn 18, Chain A	Asn 18, Chain A
	Gly 19, Chain A
Leu 111, Chain A	
Lys 112, Chain A	Lys 112, Chain A
Lys 113, Chain A	Lys 113, Chain A
	Asn 114, Chain A
Lys 118, Chain A	Lys 118, Chain A
	Arg 119, Chain A
	Arg 122, Chain A
	His 124, Chain A
	Tyr 125, Chain A
Gly 126, Chain A	Gly 126, Chain A
Gln 127, Chain A	Gln 127, Chain A
Lys 128, Chain A	Lys 128, Chain A
Ala 129, Chain A	Ala 129, Chain A
Asn 18, Chain B	Asn 18, Chain B
	Gly 19, Chain B
	His 21, Chain B
	Arg 35, Chain B
Lys 112, Chain B	Lys 112, Chain B
Lys 113, Chain B	Lys 113, Chain B
Asn 114, Chain B	Asn 114, Chain B
	Ser 116, Chain B
	Cys 117, Chain B
Lys 118, Chain B	Lys 118, Chain B
Arg 119, Chain B	Arg 119, Chain B
Arg 122, Chain B	Arg 122, Chain B
	Gln 127, Chain B
	Lys 128, Chain B
Ala 129, Chain B	Ala 129, Chain B

Table 2: Residues that interact with heparin from the 2AXM crystal structure and in the predicted heparin binding site from FGF-1 in the 2AXM crystal structure

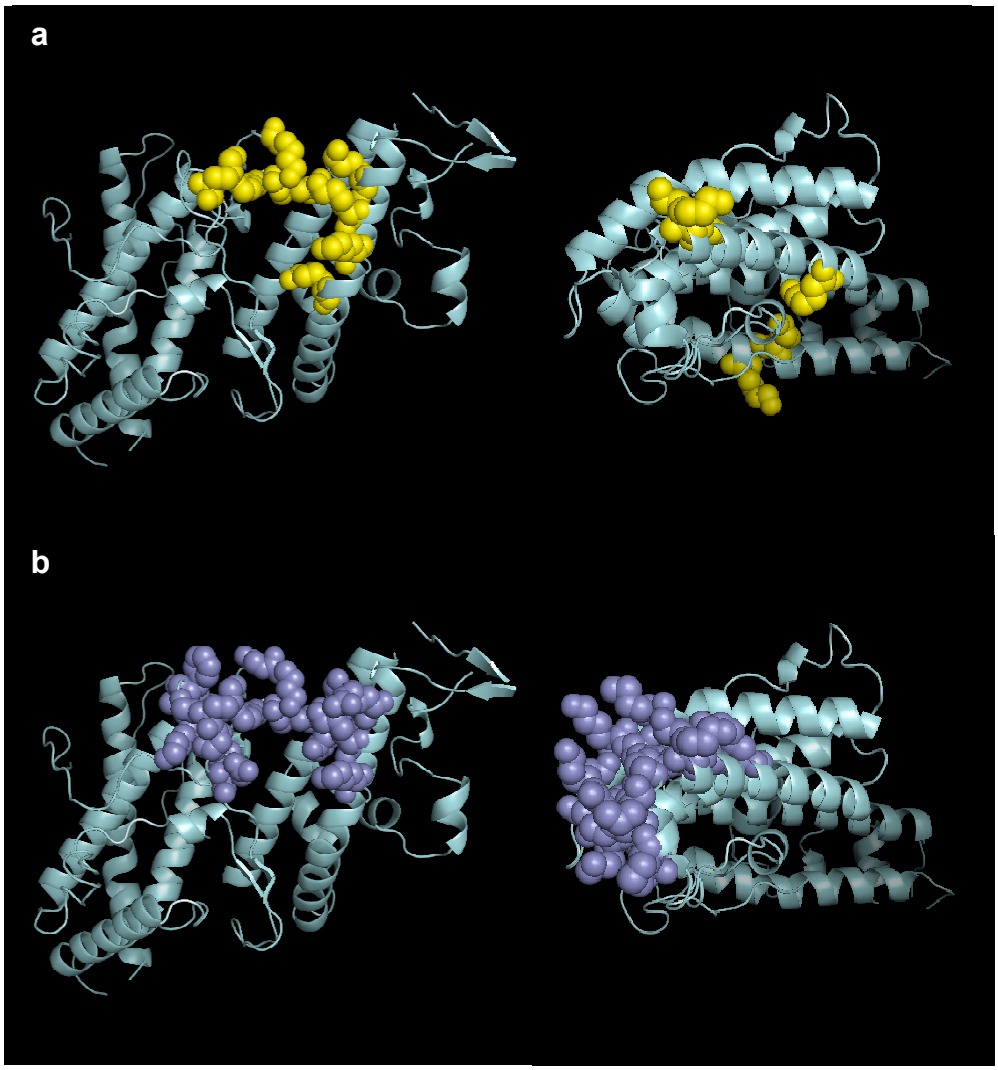

Figure 6: a, Residues important for CS-A binding to DLB3X (left) and DLB6 (right), as previously determined by mutagenesis experiments (Khunrae et al. 2009). **b,** Predicted CS-A binding site on DLB3X (left) and DLB6 (right)

residues shown previously to be important for CS-A binding. Furthermore, the predicted CS-A binding site on DBL6ε contains the two basic amino acids shown to be most important for CS-A binding, K2392 and K2395. Although the predicted CS-A binding site did not span all of the residues shown to be important for CS-A binding in the DBL6ε mutagenesis experiments, the extra residues not predicted to be part of the binding site by our program were found by mutagenesis studies using endogenous CS,

Predicted CS-A Binding Site	Mutagenesis[a]	Kd
	WT	33 mM
Asp 1236		
Gly 1237		
Lys 1238		
Phe 1240		
Gly 1242		
Lys 1243	Lys1243Ala	367 mM
Gly 1244		
Glu 1246		
Thr 1317		
Gly 1318		
Thr 1319		
Lys 1324	K1324A	122 mM
Lys 1328	Lys1328Ala	89 mM
Gly 1329		
Arg 1467	Arg1467Ala	122 mM
Tyr 1468		
Arg 1503		
Lys 1504	Lys1504Ala	172 mM
Lys 1507	Lys1510Ala	193 mM
Lys 1510		
	Lys1515Ala	488 mM

[a] Mutagenesis values from Khunrae *et al*, 2009

Table 3: Predicted CS-A binding site on DBL3X and residues experimentally determined important for CS-A binding

Predicted CS-A Binding Site	Mutagenesis[a]	Kd
	WT	80 mM
	Lys2346Ala[b]	190 mM
Ile 2384		
Cys 2385		
Lys 2388		
Arg 2389		
Pro 2391	Lys2392Ala	ND[c]
Lys 2392	Lys2395Ala	ND[c]
Lys 2395		
Tyr 2399		
	Arg2408Ala	151 mM
Lys 2451		
Ile 2452		
Leu 2453		
Gly 2454		
Lys 2462		
Lys 2465		
Trp 2466		
Met 2469		
	Lys2565Ala	215 mM
	Lys2567Ala	440 mM

[a] Mutagenesis values from Khunrae *et al*, 2009

[b] Residue is not resolved in the crystal structure

[c] CS-A bound these mutants too weakly to accurately determine a Kd value

Table 4: Predicted CS-A binding site on DBL6 and residues experimentally determine important for CS-A binding

which is likely to be longer than a tetrasaccharide, and indeed the residues not found by our predictions are 19.4 angstroms and 26.9 angstroms from the major CS-E binding site and thus are unlikely to interact with a CS-A tetrasaccharide. Furthermore, based on other experimental evidence, Khunrae and coworkers propose that K2392 and K2395 represent the true CS-A binding site whereas the other amino acids determined by mutagenesis are likely an artifact of using only the DBL6ε domain for binding studies[33].

Finally CS-A has been co-crystallized with cathepsin K[39]. Again using ScanBindSite, we successfully predicted the CS-A binding site on cathepsin K (**Fig. 7**) as well as many of the CS-A interacting residues (**Table 5**). This suggests that our method for predicting GAG binding sites is successful at predicting known binding sites and can be used to predict GAG binding sites on protein where the site is unknown.

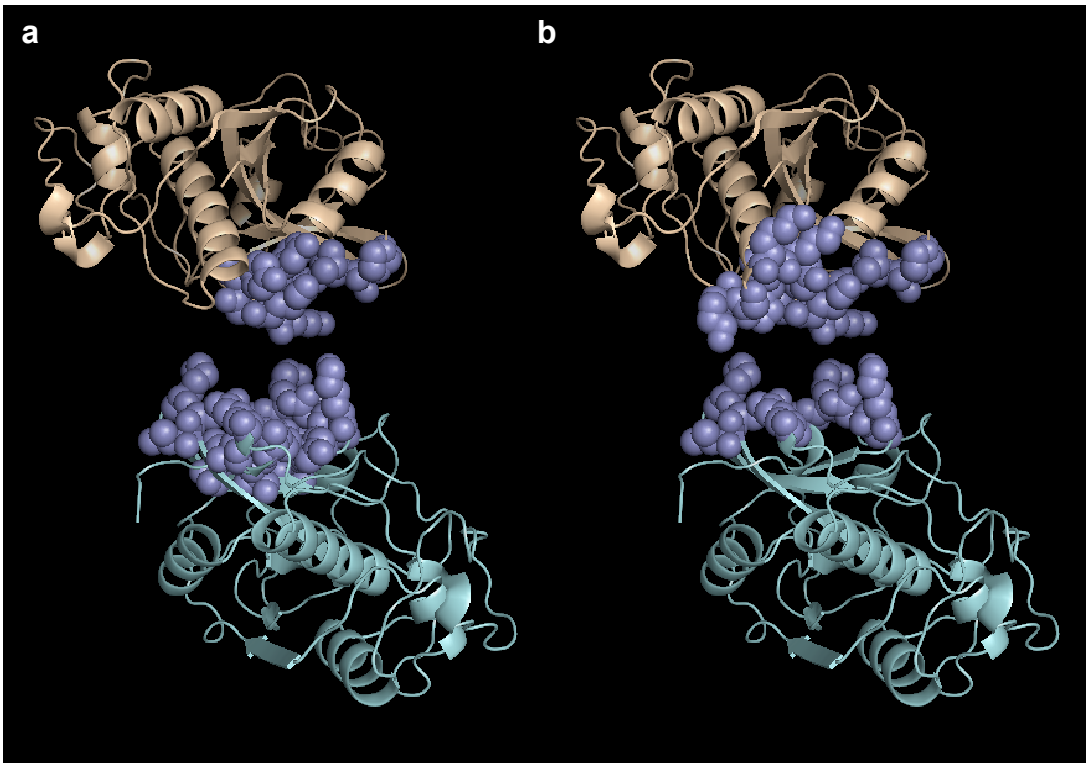

Figure 7: a, Crystal structure CS-A binding site on Cathepsin K (blue). **b,** Computationally-predicted CS-A binding site on Cathepsin K (blue)

Residues within four angstroms of CS-A in the 3C9E crystal structure	Residues within four angstroms of CS-A in the predicted binding site from 3C9E
Pro 2, Chain A	Pro 2, Chain A
Ser 4, Chain A	Ser 4, Chain A
Val 5, Chain A	Val 5, Chain A
Asp 6, Chain A	Asp 6, Chain A
Tyr 7, Chain A	
Lys 9, Chain A	Lys 9, Chain A
Lys 10, Chain A	Lys 10, Chain A
Gly 11, Chain A	Gly 11, Chain A
Tyr 12, Chain A	Tyr 12, Chain A
	Lys 39, Chain A
	Lys 40, Chain A
	Lys 41, Chain A
	Gly 43, Chain A
	Lys 44, Chain A
Asp 6, Chain B	
Arg 8, Chain B	
Lys 9, Chain B	Lys 9, Chain B
Tyr 145, Chain B	
Ser 146, Chain B	
Lys 147, Chain B	Lys 147, Chain B
Gly 148, Chain B	Gly 148, Chain B
Ile 171, Chain B	Ile 171, Chain B
Gln 172, Chain B	Gln 172, Chain B
Lys 173, Chain B	Lys 173, Chain B
His 177, Chain B	
Ile 179, Chain B	
Gly 189, Chain B	Gly 189, Chain B
Asn 190, Chain B	Asn 190, Chain B
Lys 191, Chain B	Lys 191, Chain B
Tyr 193, Chain B	
Ile 194, Chain B	
Leu 195, Chain B	

Table 5: Residues that interact with CS-A from the 3C9E crystal structure and in the predicted CS-A binding site from Cathepsin K in the 3C9E crystal structure

Given our success at predicting the binding sites of heparin of FGF-1 and FGF-2 and CS-A on DBL3 and DBL6ε, we decided to employ our approach to predict the binding site of CS-E on a number of proteins with which it is known to interact, including TNF-α, BDNF, NGF, NT-3, NT-4/5, TrkA, TrkB, TrkC, midkine, GDNF receptor alpha 1, Nogo-66, and Nogo receptor (S. Tully, C. Rogers, unpublished data).

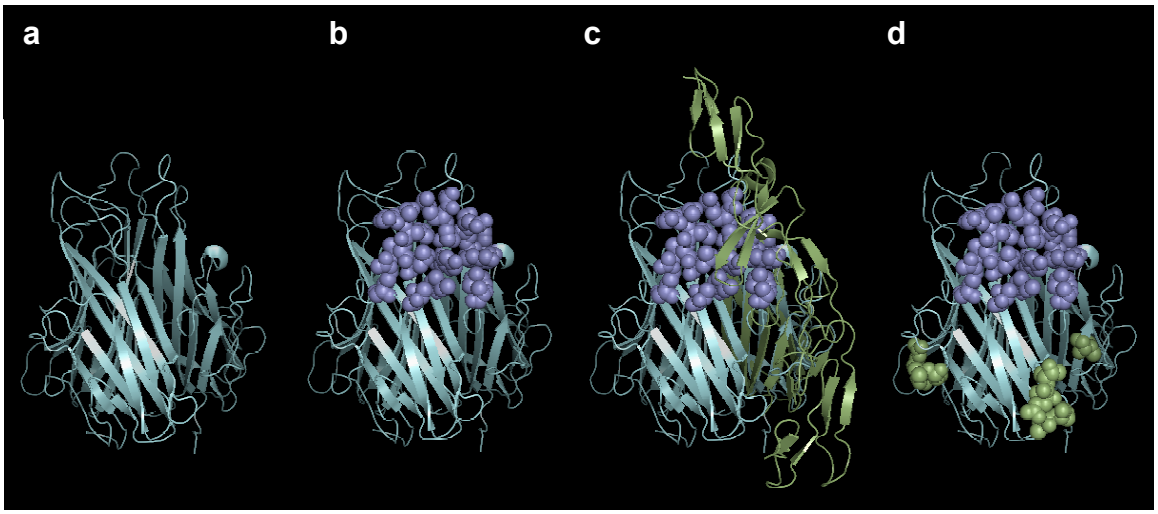

Figure 8: a, Crystal structure of TNF trimer (from 1TNF). **b,** Predicted CS-E binding site on TNF trimer. **c,** Overlay between predicted CS-E binding site (slate) and predicted TNF-R1 (green, from 1TNR) complex. **d,** Overlay between predicted CS-E binding site (slate) and residues important for TNF-R2 (green) binding

TNF-α. TNF-α is a molecule important in the pathogenesis of such diseases as rheumatoid arthritis, Chrohn's diease, and psoriasis[40]. TNF-α interacts with two receptors TNF-R1 (p55) or TNF-R2 (p75) that modulate its biological functions. We predicted the CS-E binding site from the TNF-α structure in 1TNF, which consists of amino acids 6 through 157 of human TNF-α (**Fig. 8a, b**). The CS-E binding site on the TNF trimer spans two of the three monomers and includes basic amino acids on the first monomer — Arg103 and Arg138 — as well as basic amino acids on the second monomer — Lys65 and Lys112.

We next wanted to know how the interaction between CS-E and TNF-α might affect the interaction between TNF-α and its receptors. To determine how TNF-α interacts with TNF-R1, we constructed a homology model of this complex based on the crystal structure of TNF-β and TNF-R1[41]. This model, along with other mutagenesis

studies[42], indicates that TNF-R1 interacts with TNF-α at the same interface as CS-E interacts with TNF-α (**Fig. 8c**). Thus one would predict from the CS-E binding site that the CS-E might block the interaction between TNF-α and TNF-R1. Alternatively, mutagenesis studies indicate that TNF-R2 interacts with TNF-α on a different part of the protein from the predicted CS-E binding site[42] suggesting that CS-E should not block the interaction between TNF-α and TNF-R2 (**Fig. 8d**). Excitingly, previous work by Tully and coworkers[13,34] has demonstrated that CS-E blocks the interaction between TNF-α and TNF-R1 but not TNF-R2, confirming the computational predictions.

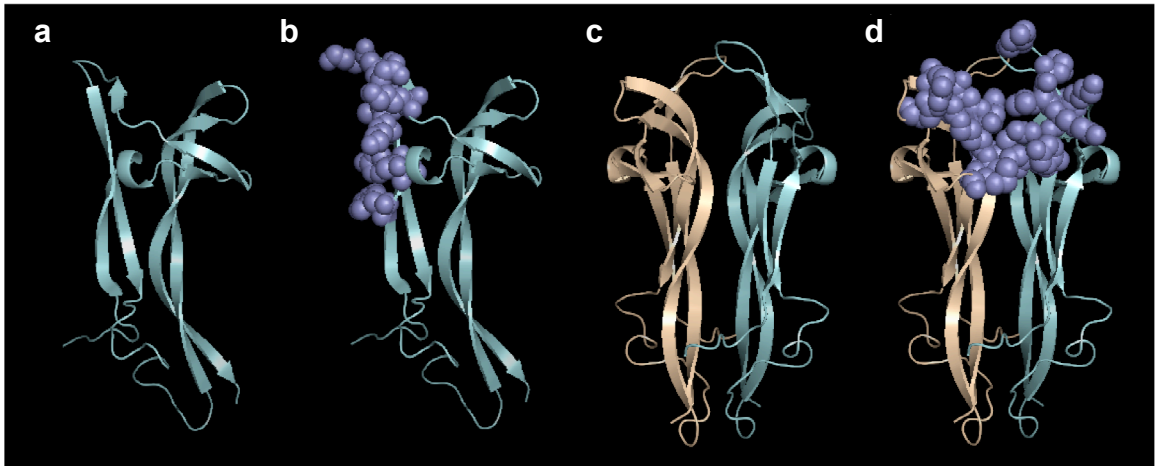

Figure 9: a, BDNF monomer crystal structure (from 1BND). **b,** CS-E binding site (slate). **c,** Homology model of BDNF dimer crystal structure (wheat and cyan). **d,** CS-E binding site (slate)

Neurotrophins and Trk Receptors. The NGF family of neurotrophins contains BDNF, NGF, NT-3, and NT-4/5 and shows a high degree of structural homology between these structures (average RMSD for Cα atoms = 0.926 angstroms). Neurotrophins function generally to regulate growth, survival, and differentiation of neurons[43] and have been shown to signal predominantly through the common p75 neurotrophin receptor and through distinct Trk receptors, TrkA, TrkB, and TrkC. CS-E has been shown to bind to

all four members of this family[44].

BDNF. BDNF is an important molecule for synaptic plasticity and learning and memory and functions, in part, through its interaction with the high-affinity TrkB receptor[45]. We predicted the CS-E binding site from the BDNF structure in 1BND, which contains amino acids 8 through 116 of human BDNF. The CS-E binding site on the BDNF monomer is predominantly across a beta-sheet and within a charged loop region that contains three basic residues — Lys41, Lys46, and Lys50 (**Fig. 9a, b**). The CS-E binding site was also predicted from a homology model of the BDNF dimer (**Fig. 9c, d, Table 6**). The CS-E binding site on the dimer structure is within a similar region to the BDNF monomer but also contains amino acids from the second dimer molecule. This includes basic residues Arg88, Arg97, and Arg101 on the second BDNF molecule.

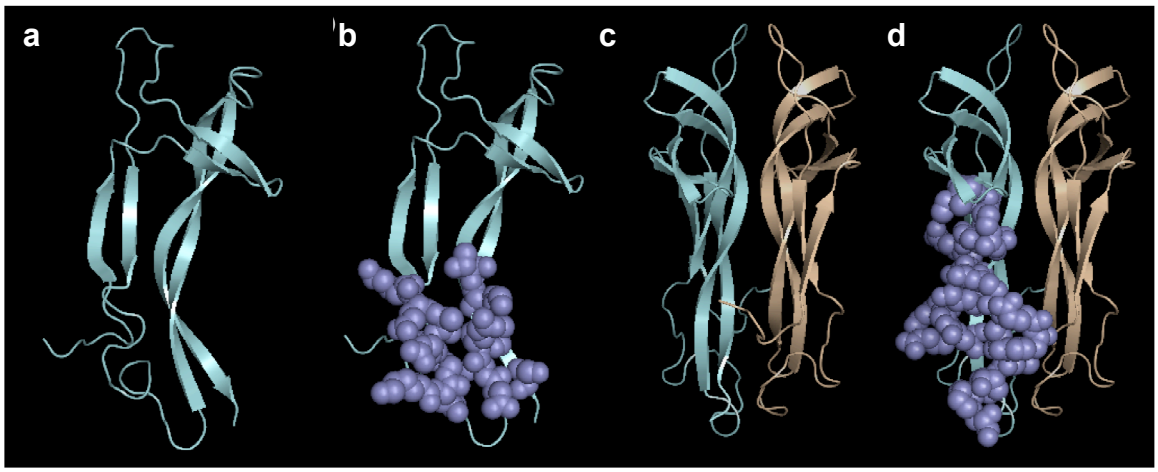

Figure 10: a, NT-3 monomer crystal structure (from 1BND). **b,** CS-E binding site (slate). **c,** NT-3 dimer crystal structure (from 1NT3). **d,** CS-E binding site (slate)

NT-3. NT-3 contributes to neuronal survival, neurotransmission, and synaptic plasticity[46]. NT-3 interacts preferentially with TrkC although it has also been shown to signal through TrkA and TrkB in certain cellular contexts[47]. We predicted the CS-E

BDNF	NGF	NT-4/5	NT-3
Chain A, Lys41	Chain A, Asn46	Chain A, Asp32	Chain A, Ile28
Chain A, Lys46	Chain A, Ser47	Chain A, Leu33	Chain A, Arg56
Chain A, Gln48	Chain A, Val48	Chain A, Arg34	Chain A, Cys57
Chain A, Leu49	Chain A, Phe49	Chain A, Arg36	Chain A, Glu59
Chain A, Lys50	Chain A, Lys50	Chain A, Arg98	Chain A, Ala60
Chain A, Tyr52	Chain A, Tyr52	Chain A, Asp103	Chain A, Arg61
Chain B, Met31	Chain B, Lys32	Chain A, Gln105	Chain A, Asn76
Chain B, Arg88	Chain B, Lys34	Chain A, Arg107	Chain A, Gln78
Chain B, Asp93	Chain B, Lys88	Chain A, Val108	Chain A, Lys80
Chain B, Arg97	Chain B, Asp93	Chain A, Gly109	Chain A, Thr81
Chain B, Ile98	Chain B, Gly94	Chain A, Trp110	Chain A, Gln83
Chain B, Gly99	Chain B, Lys95	Chain A, Arg111	Chain A, Arg103
Chain B, Trp100	Chain B, Gln96	Chain A, Trp112	Chain A, Asp105
Chain B, Arg101	Chain B, Ala98	Chain B, Trp23	Chain A, Ala111
Chain B, Phe102	Chain B, Trp99	Chain B, Ala47	Chain A, Leu112
	Chain B, Arg100	Chain B, Leu52	Chain A, Ser113
	Chain B, Phe101	Chain B, Arg53	Chain A, Lys115
		Chain B, Tyr55	Chain B, Arg8
			Chain B, Glu10
			Chain B, Tyr11

Table 6: Predicted CS-E binding sites on BDNF, NGF, NT-4/5, and NT-3

binding site on NT-3 from NT-3 in 1BND, which contains amino acids 8 through 116 of human NT-3. The predicted CS-E binding site on the NT-3 monomer is predominantly within a loop region between the fourth and fifth beta sheet of the structure and contains four basic amino acids — Lys58, Arg61, Lys64, and Lys80 (**Fig. 10a, b**). Interestingly, although NT-3 has a high degree of structural homology to BDNF (RMSD 0.967 angstroms), the CS-E binding site on NT-3 is different from the CS-E binding site on BDNF. The preference for one binding site over the other is likely due to changes in basic residues in the loop regions of BDNF and NT-3 (**Fig. 11**). In particular, the loop that contains the CS-E binding site in BDNF is very different from the homologous loop within NT-3, with Lys41 and Lys46 in BDNF being homologous to Glu40 and Asn45 in NT-3. Similarly the CS-E binding site in NT-3 contains Lys62, which is homologous to

Gly63 in BDNF. The CS-E binding site on the NT-3 dimer is similar to the CS-E binding site on the NT-3 monomer but fails to contain Lys58 and Arg64, although it does contain other basic residues such as Arg103 and Lys115 on the first NT-3 monomer and Arg8 on the second one (**Fig. 10c, d, Table 6**).

```
BDNF            ----HSDPARRGQLSVCDSISEWVTAADKKTAVDMSGGTVTVLEKVPVSKGQ-LKQYFYE  55
NGF             --SSSHPIFHRGEFSVCDSVSVWVG--DKTTATDIKGKEVMVLGEVNINNSV-FKQYFFE  55
NT-4/5          GVSETAPASRRGELAVCDAVSGWVT--DRRTAVDLRGREVEVLGEVPAAGGSPLRQYFFE  58
NT-3            ---YAEHKSHRGEVSVCDSESLWVT--DKSSAIDIRGHQVTVLGEIKTQNSP-VKQYFYE  54

BDNF            TKCNPMGYTKEG--------CRGIDKRHWNSQCRTTQSYVRALTMDSKKRIGWRFIRIDTS 108
NGF             TKCRDPNPVDSG--------CRGIDSKHWNSYCTTTHTFVKALTMDG-KQAAWRFIRIDTA 107
NT-4/5          TRCKADNAEEGGPGAGGGGCRGVDRRHWVSECKAKQSYVRALTADAQGRVGWRWIRIDTA 118
NT-3            TRCKEARPVKNG--------CRGIDDKHWNSQCKTSQTYVRALTSENNKLVGWRWIRIDTS 107

BDNF            CVCTLTIKRGR--  119
NGF             CVCVLSRKAVRRA  120
NT-4/5          CVCTLLSRTGRA-  130
NT-3            CVCALSRKIGRT-  119

Beta Sheets
```

Figure 11: Cationic amino acids (yellow) in the respective CS-E binding sites.

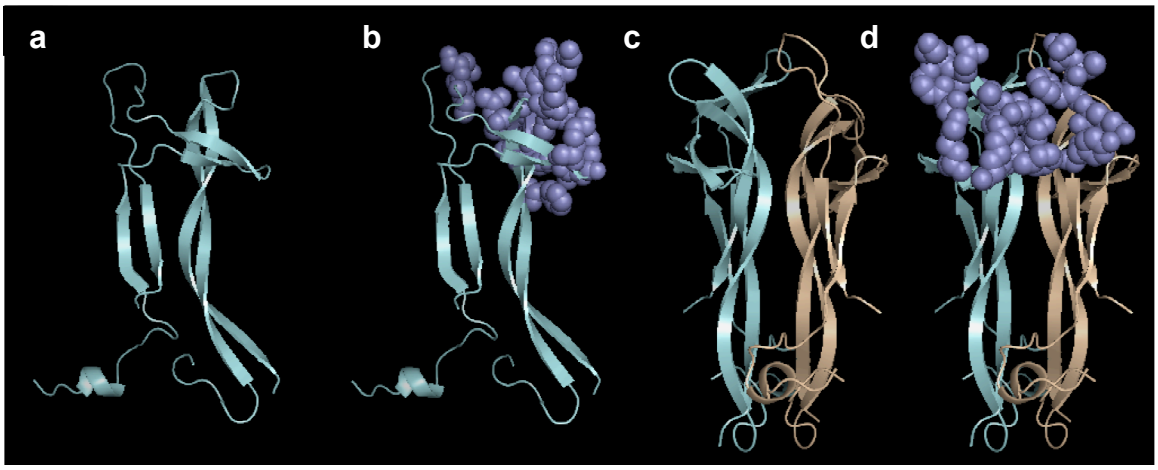

Figure 12: a, NGF monomer crystal structure (from 2IFG). **b,** CS-E binding site (slate). **c,** NGF dimer crystal (from 2IFG). **d,** CS-E binding site (slate)

NGF. NGF is also a neurotrophin that is involved in maintenance and survival of peripheral and sensory neurons[48] and has been shown to signal through TrkA. We predicted the CS-E binding site on the NGF monomer from NGF in chain E of 2IFG,

which contains amino acids 2 through 115 of human NGF. CS-E binds to two hairpin loops and the adjacent beta sheets and contains five basic amino acids — Lys32, Lys34, Lys88, Lys95, and Arg100 (**Fig. 12a, b**). The CS-E binding site on NGF is distinct from the CS-E binding site both on BDNF as well as NT-3. Again this can be attributed to differences in the amino acid sequences in the loops with which CS-E interacts (**Fig. 11**). For example, the homologous amino acids to Lys41 and Lys46 in the CS-E binding site of BDNF are Glu41 and Asp46 on NGF, and the homologous amino acids to Arg61 and Lys64 in the CS-E binding site of NT-3 are Asn62 and Asp64 in NGF. Correspondingly, Lys32 and Lys34 that make up the CS-E binding site on NGF are homologous to Arg31 and His33 in NT-3 and homologous to Ser32 and Gly34 in BDNF, respectively. Although CS-E interacts with different loops of the NGF and BDNF monomers, CS-E interacts with the same loops of the NGF and BDNF dimers (**Fig. 12c, d, Table 6**). This

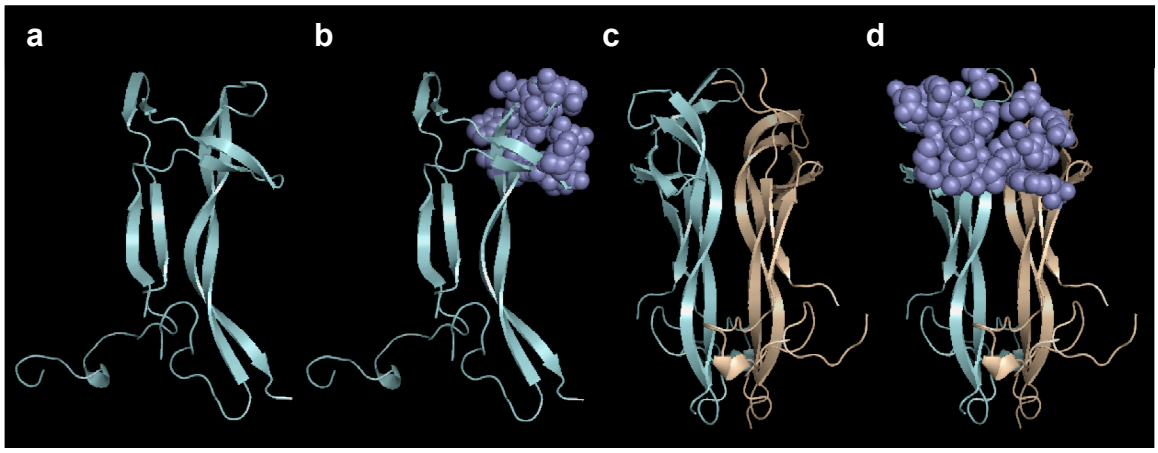

Figure 13: a, NT-4/5 monomer crystal structure (from 1HCF). **b,** CS-E binding site (slate). **c,** NT-4/5 dimer crystal (from 1HCF). **d,** CS-E binding site (slate)

is due, in part, to the fact that when the NGF and BDNF dimers form, loops that are distant from each other in the monomer structures and now close to each other in the dimer structure.

NT-4/5. NT-4/5 is a neurotrophin that has been shown to promote peripheral sensory and symphathetic neuronal survival, and, like BDNF, signals through TrkB[49]. We predicted the CS-E binding site on NT-4/5 from the structure of the NT-4/5 monomer in 1HCF, which contains amino acids 1 through 127 of human NT-4/5. CS-E binds predominantly in two loop regions between the second and third beta-sheet and the seventh and eighth beta-sheet of NT-4/5 (**Fig. 13a, b**). The CS-E binding site contains five basic amino acids — Arg34, Arg36, Arg98, Arg107, and Arg111 — and is very similar to the CS-E binding site on NGF. Indeed, four of the five basic residues are homologous between the CS-E binding sites on NGF and NT-4/5 (**Fig. 11**). The CS-E binding site on the NT-4/5 dimer is also similar to the CS-E binding site on the NGF dimer and BDNF dimer and includes further interaction with the second NT-4/5 molecule including with Arg53 (**Fig. 13c, d, Table 6**).

The predicted neurotrophin CS-E binding sites share are number of common features. Although different in its exact binding site, CS-E predominantly binds in the loops that connect that beta sheets. In the case of NT-3, this corresponds to loop 3, while in the remaining neurotrophins, this corresponds to loops 1, 2, and 4. Furthermore, none of the binding sites fall within the dimerization interface or the actual or predicted Trk or p75 receptor interface, suggesting that CS-E would not block these interactions. In the dimer structures, the CS-E binding site is always across the face of the two neurotrophins with potential electrostatic interactions between CS-E and both monomers in the complex. For example, in the NGF monomer, CS-E is predicted to interact with Lys32, 34, 88, and 95, and Arg100 while in the NGF dimer, CS-E is predicted to interact with these amino

acids on the first protein in the monomer as well as Lys 50 on the second protein. Similarly in the NT-3 monomer, CS-E is predicted to interact with Arg 61 and 103 and Lys 80 and 115 while in the NT-3 dimer, CS-E is predicted to also interact with Arg 8 on the second NT-3 molecule. Since neurotrophins are suggested to exist predominantly as dimers in nature[50], this data suggests that CS-E may primarily interact with the neurotrophin dimers rather than contribute to dimer formation.

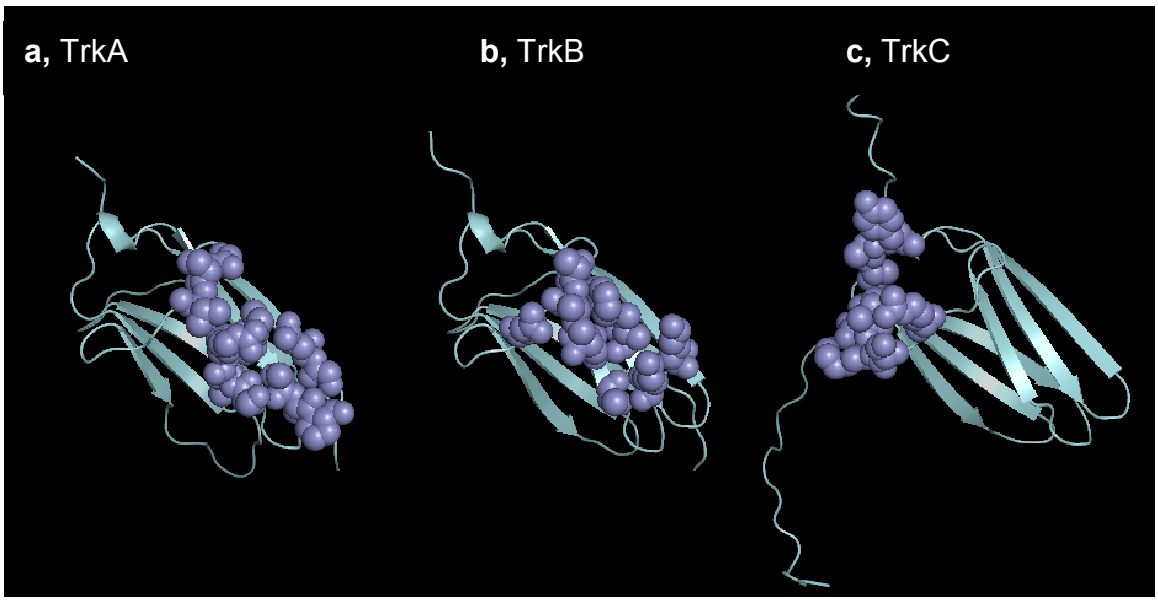

Figure 14: CS-E binding sites on the Trk family of receptors. **a,** TrkA **b,** TrkB **c,** TrkC

Trk Receptors. The Trk family of proteins, which includes TrkA, TrkB, and TrkC, are one set of receptors for the NGF family of neurotrophins and interact weakly with CS-E. We modeled the CS-E binding sites on TrkA, TrkB, and TrkC proteins from their crystal structures in 1WWW, which contain amino acids on 282 through 382 of human TrkA, 1HCF, which contain amino acids 286 through 383 of human TrkB, and 1WWC, which contain amino acids 300 through 404 of human TrkC (**Fig. 14, Table 7**). TrkA and TrkB

TrkA	TrkB	TrkC
Ser312	Ala314	Arg343
Leu313	Gln316	Ser345
Arg314	Phe318	Lys346
Gly319	Ala322	Ile347
Ser320	Ile323	Asn366
Val321	Asn325	Lys367
Leu362	Ile362	Pro368
Ala364	Lys364	Thr369
Asn365	Lys369	Tyr371
Pro366		
Gly368		
Gln369		

Table 7: Predicted CS-E binding sites on TrkA, TrkB, and TrkC

have similar CS-E binding sites, which reside mostly across the face of three beta sheets. The TrkA binding site contains only one basic residue — Arg314 — whereas the TrkB binding site contains two basic residues — Lys364 and 369. The CS-E binding site on TrkC, however, is in a different region of the protein from the homologous sites on TrkA and TrkB. The CS-E binding site on TrkC is near the top of the β-barrel structure and contains three basic residues – Arg343, Lys346, and Lys367.

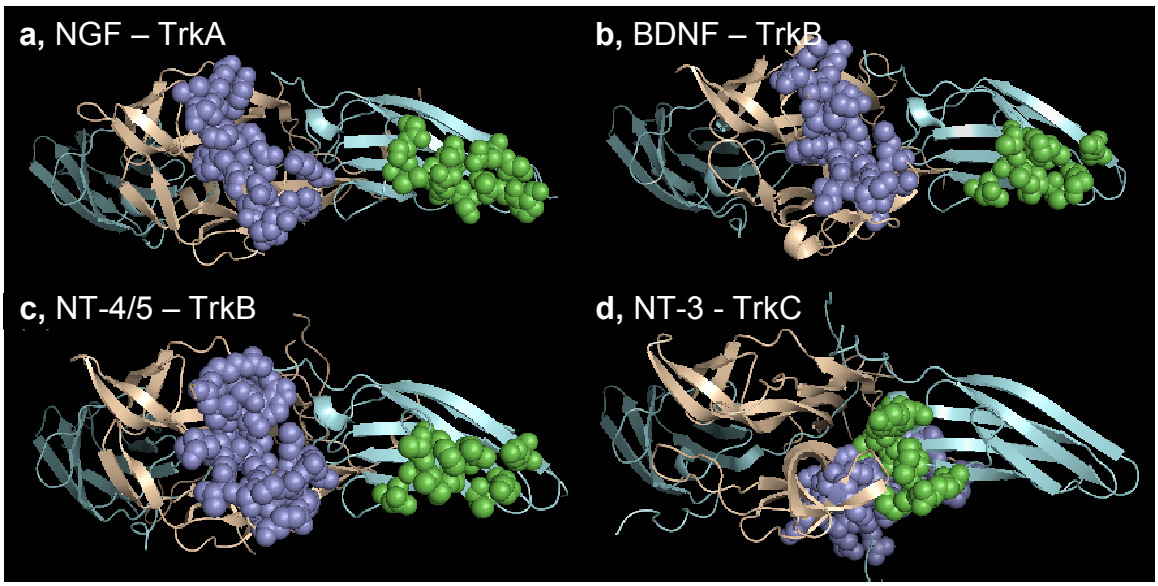

Figure 15: CS-E binding sites for the neurotrophins (slate) and Trk receptors (green), projected onto the neurotrophin (wheat) – receptor (cyan) complexes. **a,** NGF – TrkA, **b,** BDNF – TrkB, **c,** NT-4/5 – TrkB, **d,** NT-3 – TrkC

Complex Formation. Heparin polysaccharides interact with multiple proteins in a protein complex[20,21] and facilitate in signaling[51]. CS-E binds to neurotrophins such as NT-4/5 and NGF as well as weakly to their receptors, including TrkA and TrkB[34,44]. To investigate whether CS-E might facilitate or stabilize these neurotrophin - receptor complexes, we plotted the CS-E binding sites for the neurotrophin dimers and the Trk proteins onto predicted or actual crystal structures of the neurotrophin - Trk complexes, including NT-3 / TrkC, NT-4/5 / TrkB, BDNF / TrkB, and NGF / TrkA (**Fig. 15**). In the case of the predicted NT-3 / TrkC complex, the CS-E binding site for TrkC falls within the NT-3 / TrkC protein-protein interaction interface, thus making it unlikely that CS-E would facilitate the complex formation (although at the same time, the interaction between CS-E and TrkC would appear too weak to block the formation of the NT-3 / TrkC complex). In every case besides the NT-3 / TrkC complex, however, the CS-E binding sites on the neurotrophin dimer and the Trk protein occur on the same face of the protein complex. Thus, one long CS molecule, with CS-E motifs spaced at the correct

distance, could potentially span both the CS-E binding site on the Trk protein and the CS-E binding site on the neurotrophin dimer. Furthermore, the distance between the basic amino acids on the neurotrophin dimer binding sites and those of the Trk binding sites is such that these amino acids would be correctly positioned to interact with the sulfate groups on repeating CS units. For example, the average distance between the exposed basic amino acids within the CS-E binding site on NGF and TrkA is 25.9 angstroms, which is approximately twice the distance between the sulfate groups on the four position of CS-E or approximately the distance between the furthermost four sulfate groups on a CS-E hexasaccharide. This suggests that CS-E might facilitate the formation or stabilization of the neurotrophin / Trk protein complex. Indeed, previous studies have reported that mutations of residues within the predicted CS-E binding site on NGF — in particular, Lys32, Lys34, and Glu35 to alanine or Lys32, and Arg95 to alanine — decreased binding of NGF to a fibroblast cell line that expresses only TrkA by 45 and 60% even though none of these residues make direct side-chain contacts with TrkA[52]. Nevertheless, the CS-E binding sites on the Trk molecules do not contain a high density of basic amino acids that is characteristic of traditional CS-E binding sites and thus are more likely secondary binding sites of CS-E molecules, suggesting that CS-E may not necessarily assist in bringing the neurotrophin and the Trk receptor together but rather may stabilize the preformed neurotrophin-Trk complex.

Other Proteins

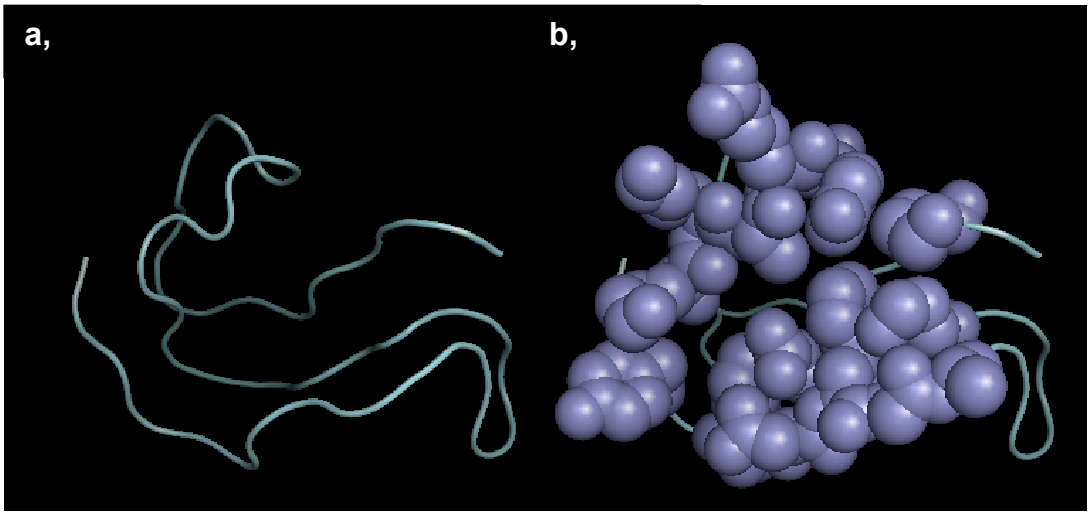

Figure 16: a, Midkine (from 1MKN). **b,** Predicted CS-E binding site (slate)

Residues whose NMR chemical shift changes with addition of Heparin 12-mer	Predicted CS-E Binding Site
	Tyr 64
	Phe 66
Glu 67	
Asn 68	Asn 68
Trp 69	Trp 69
	Gly 70
	Ala 71
Lys 79	Lys 79
Val 80	
Arg 81	Arg 81
Leu 85	
Lys 86	Lys 86
Lys 87	Lys 87
Ala 88	Ala 88
Arg 89	Arg 89
Tyr 90	Tyr 90
Asn 91	
Cys 94	
	Lys 102

Table 8: Residues that interact with heparin as previously determined by NMR (Iwasaki 1997) and the predicted CS-E binding site

Midkine. Midkine is a 13 kDa protein whose expression is regulated by retinoic acid and which has been shown to enhanced neurite outgrowth and survival[53]. We predicted the CS-E binding site using the structure of midkine from 1MKN, which contains amino acids 23 through 81 of human midkine. The predicted CS-E binding site is within the N-terminal region of the protein and spans most of one face of the protein (**Fig. 16**). The binding site contains five basic amino acids and contains nine of the fourteen amino acids whose NMR chemical shifts[53] were affected upon the addition of a heparin 12mer (**Table 8**).

GDNF. GDNF is a neuronal survival factor that is structurally distinct from the NGF family of neurotrophins. GDNF has been implicated in neuronal differentiation, survival, and protection[54]. We predicted the binding site on CS-E based on the structure of GDNF from 2V5E, which contains amino acids 34 through 134 of human GDNF and chain D of 3FUB, which contains amino acids 32 through 134 of human GDNF. The CS-E binding site on GDNF is near the N-terminus of the protein and within a long stretch of basic amino acids, RRGQRGKNR, that is devoid of traditional secondary structure (**Fig. 17, Table 9**). The two GDNF structures are from different crystallization of GDNF, one as a monomer and the other as a dimer, and although they differ significantly (RMSD = 1.480 angstroms), they have similar CS-E binding sites, with the largest difference in the binding site being a consequence of differences in the placement of the alpha helix in the protein structure. This suggests that our method is robust to differences in protein structure such as might occur during different crystallization processes. Furthermore since the GDNF structure appears very mobile around the CS-E binding site, this may indicate a role for CS-E in stabilizing specific structural conformations.

GDNF Receptor. The GDNF-family receptor α 1 binds to GDNF and signals through the receptor tyrosine kinase RET[55]. We predicted the CS-E binding site on the GDNF receptor from the structure of the GDNF receptor from Chain C in 3FUB, which contains amino acids 150 through 384 of the rat GDNF receptor. The predicted CS-E binding site on the isolated GDNF receptor is on the surface on two alpha helixes and is close to but separate from the binding interface between GDNF and the GDNF receptor (**Fig. 18a, b,**

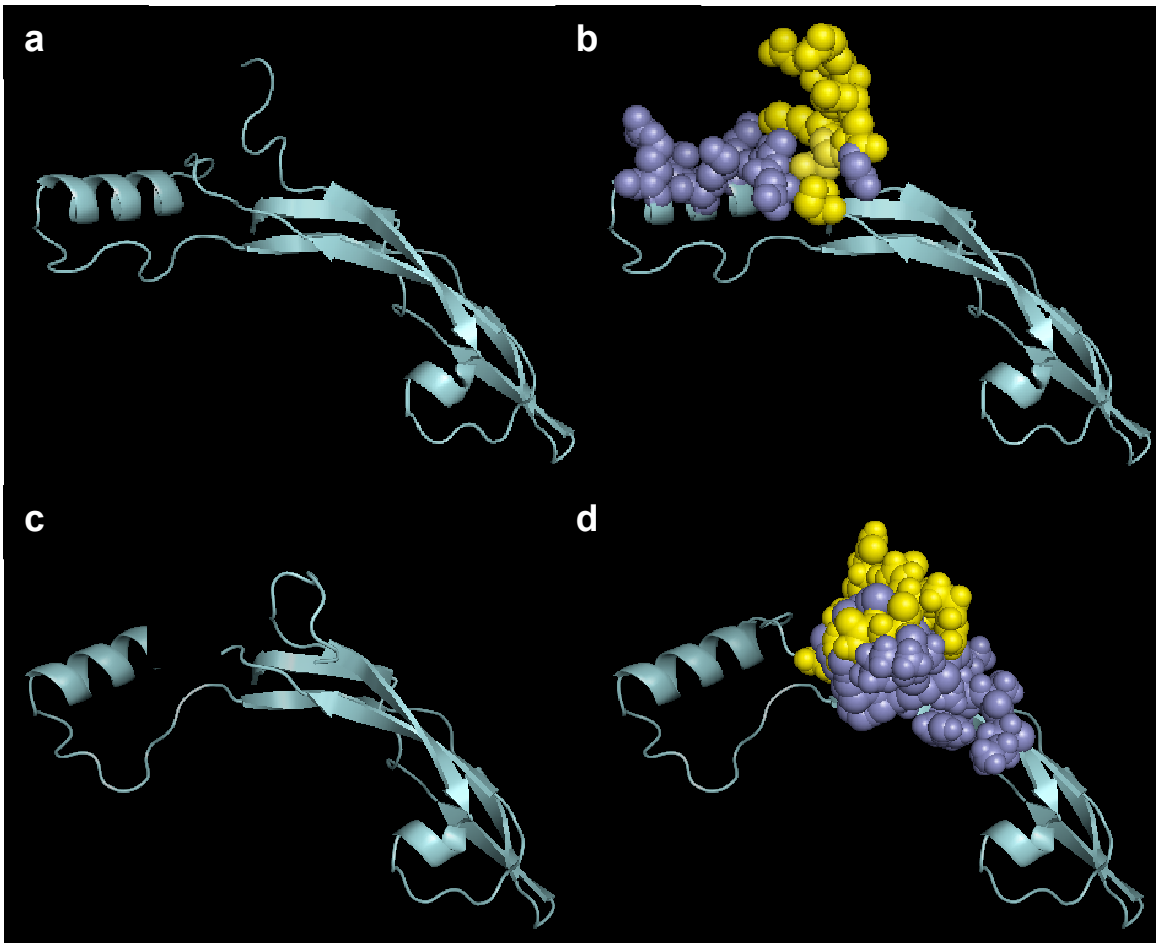

Figure 17: a, GDNF crystal structure (from 2V5E). **b,** Predicted CS-E binding site on GDNF from (**a**) (slate and yellow). **c,** GDNF crystal structure (from Chain D 3FUB). **d,** Predicted CS-E binding site on GDNF (slate and yellow) from (**c**). Residues predicted to be in the CS-E binding site for both GDNF crystal structures are yellow; those predicted to be in the CS-E binding site for only one GDNF crystal structure are colored slate.

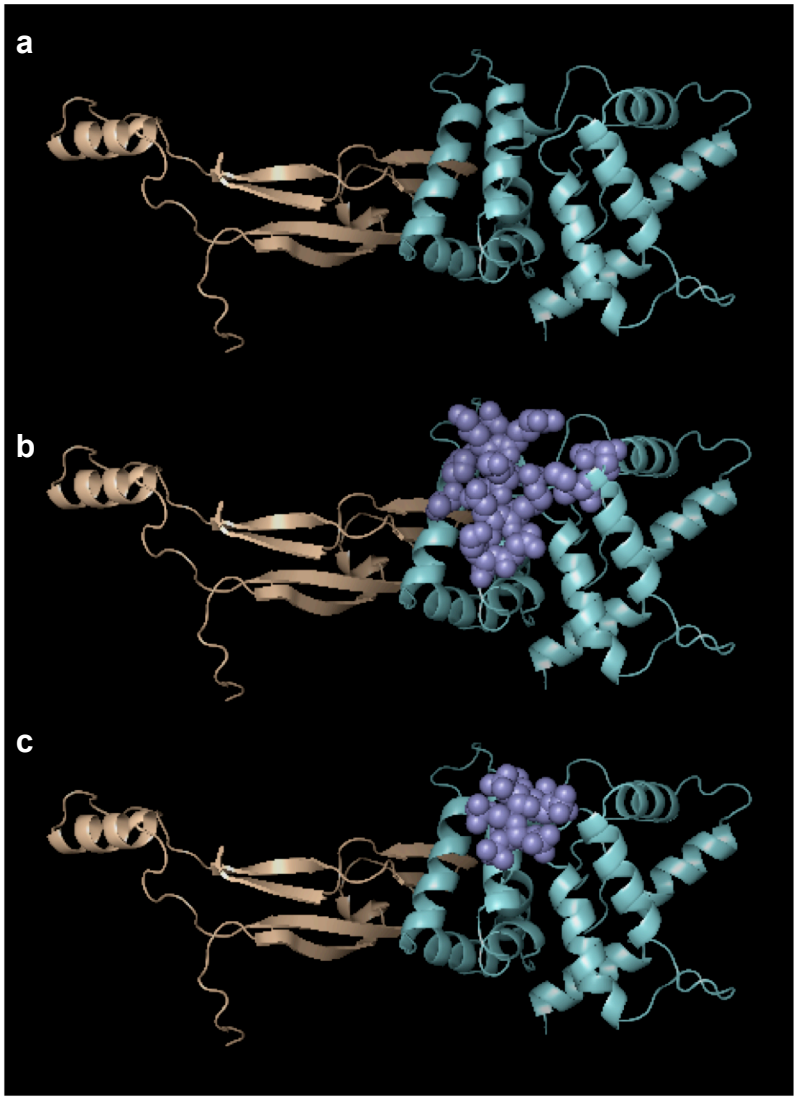

Figure 18: a, GDNF protein–receptor complex (wheat and cyan, from 3FUB). **b,** Predicted CS-E binding site (slate). **c,** Binding site of heparin mimic sucrose octasulfate (slate) as determined from the co-crystal structure (2V5E)

Table 9). The CS-E binding site consists of four charged residues — Lys169, Lys191, Lys194, and Lys202. The CS-E binding site on the GDNF protein - receptor complex is shifted slightly toward GDNF compared with the binding site on isolated GDNF receptor and overlaps with the binding site of the heparin mimic sucrose octasulfate found in the GDNF - GDNF receptor crystal structure (**Fig. 18c**).

Unlike the neurotrophins and their receptors, the CS-E binding site on GDNF and

GDNF Monomer (3FUB)	GDNF Monomer (2V5E)	GDNF Receptor	Nogo
Arg32	Gln34	Lys169	Thr243
Gly33	Arg35	Tyr170	Leu246
Gln34	Gly36	Ala173	Ala247
Arg35	Lys37	Thr176	Pro248
Gly36	Asn38	Pro177	Leu249
Lys37	Arg39	Asn188	Arg250
Asn38	Gly40	Arg190	Ala251
Arg39	Ser71	Lys191	Gln253
Val42	Asp73	Lys194	Arg269
Leu43	Ala74	Ala195	Pro270
Thr44	Ala75	Gln198	Ala273
Ala45	Lys81	Lys202	Trp274
Ile46	Lys84		Lys277
Tyr67	Asn85		Phe278
Ser69	Arg88		Arg279
Gly70			
Ser71			

Table 8: CS-E binding site on GDNF monomer (3FUB, 2V5E crystal structure), GDNF receptor, and Nogo

its receptor are not close to each other but rather on opposite sides of the protein, and even the binding site on the adjacent receptor appears to be in the wrong orientation to allow one CS-E molecule to span both binding sites (**Fig. 19**). Thus it does not look likely that CS-E would facilitate or necessarily stabilize a complex between GDNF and its receptor. Nevertheless since the CS-E binding site on GDNF is within an unstructured region of the protein that is structurally different in different crystallizations, it is possible that the CS-E binding site on GDNF could orient in such a way as to interact with the CS-E binding site on the GDNF receptor.

Nogo Receptor. The Nogo receptor interacts with Nogo and is important for axonal regeneration in the adult vertebrae central nervous system[56]. The Nogo receptor consists of a signal peptide followed by eight leucine-rich repeats (LRR), a leucine-rich region C-terminal domain (LRRCT), and a predicted transmembrane / glycosylphosphatidylinositol linkage[57]. We predicted the CS-E binding site from the Nogo receptor structure in 1P8T, which contains amino acids 27–311 of the human Nogo receptor. The CS-E binding site is at the end of the final LRR and within the LRRCT and contains four basic residues — Arg250, Arg269, Lys277, and Arg279 (**Fig. 20, Table 9**). This binding site is also separate from the predicted ligand-binding regions on the protein[56], suggesting that CS-E is not likely to block the interaction between the Nogo receptor and its ligands.

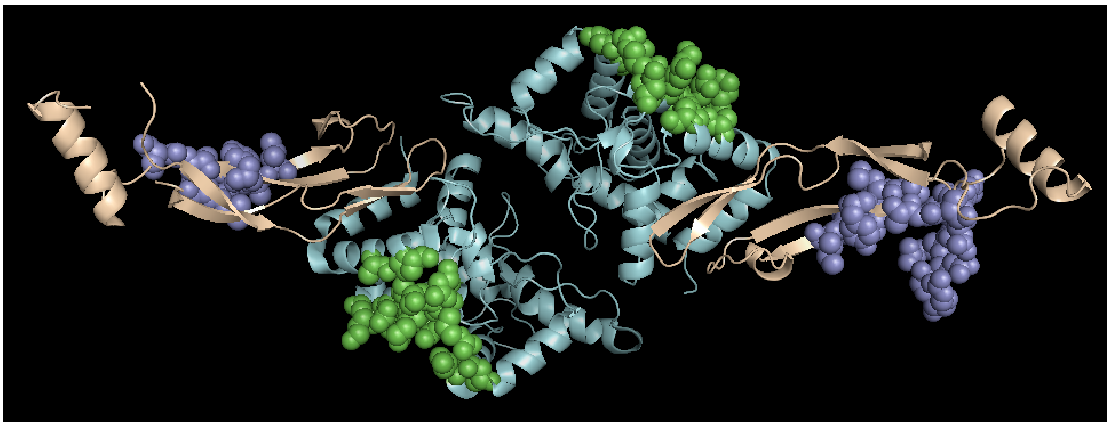

Figure 19: CS-E binding sites (slate) on GDNF (wheat) and CS-E binding sites (yellow) on the GDNF receptor (cyan) as mapped onto the GDNF – GDNF receptor complex (3FUB)

CS-E Binding Site Characteristics. Although CS-E binds to a variety of different proteins with distinct structural motifs, we found that certain general features characterize many of the CS-E binding sites. CS-E binding sites are enriched in basic residues, as might be expected given the six acidic groups on the CS-E tetrasaccharide. Of the proteins modeled that strongly interact with CS-E, the median number of basic residues

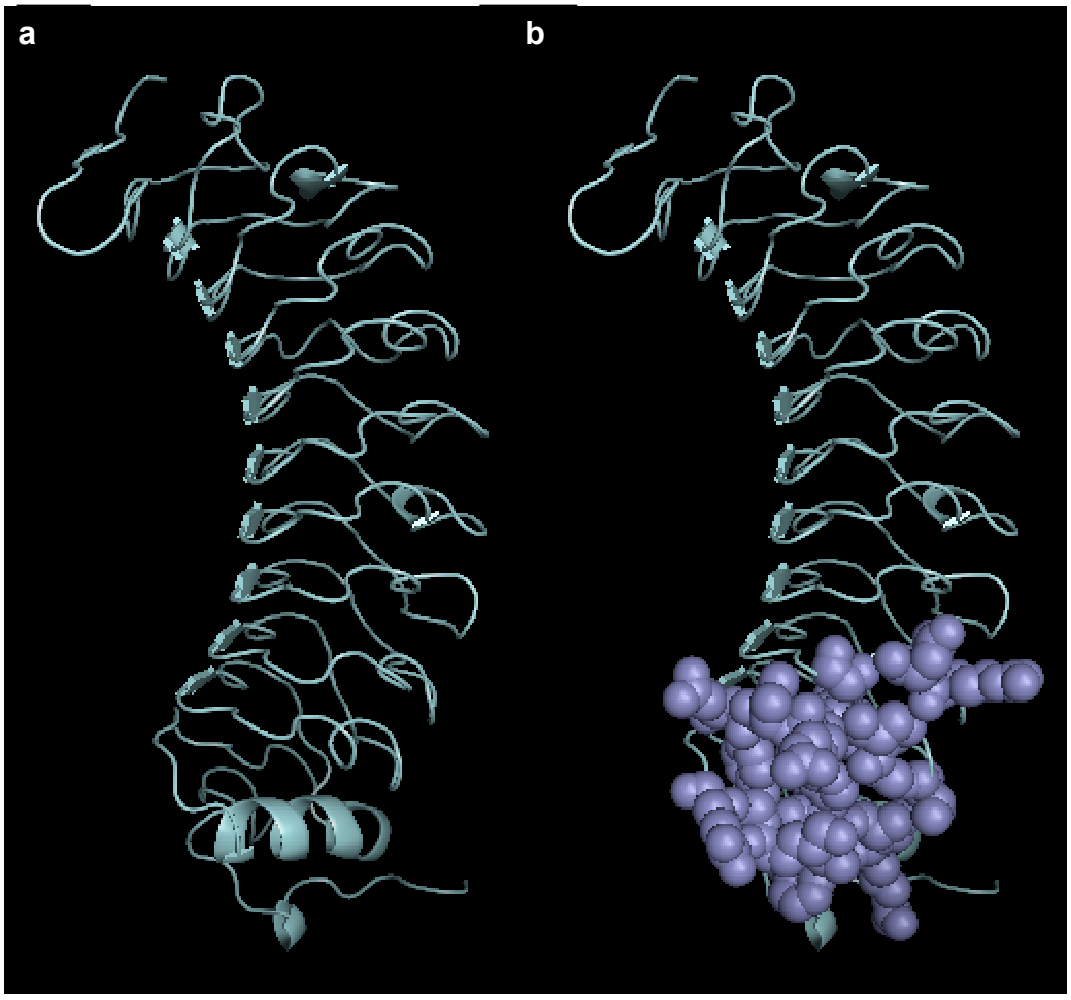

Figure 20: a, Nogo receptor crystal structure (from 1P8T). **b,** CS-E binding site (slate)

in the CS-E binding sites is four. Furthermore, some of the CS-E binding sites are characterized in primary sequence by two basic amino acids that are usually within a few

residues from one another and a third basic amino acid that is more distant from the first two. One example of this type of binding site is the NT-4/5 dimer. The CS-E binding site on the NT-4/5 dimer consists of two arginine residues at amino acid positions 34 and 36 on the first protein and then a third arginine at amino acid position 53 on the second. Nevertheless, this binding site also contains additional basic residues that do not fit within this simplified tetrasaccharide binding site motif and may be important in further stabilizing the CS-E tetrasaccharide or for making extra contacts with longer CS-E chains.

Further analysis of the secondary structure of the CS-E binding sites reveals more similarities between proteins. The secondary structure of the CS-E binding sites are characterized by two basic amino acids that are approximately 5 Å from each other and a third basic amino acid approximately 15 Å away from the first two but that can be connected by a line to the first two without bisecting the protein. Thus, in the NT-4/5 dimer, the average distance between the terminal nitrogen of the guanidinium groups of Arg34 and Arg36 is 4.1 Å while the terminal nitrogen of the guanidinium group of Arg53 are an average of 16.9 Å from the terminal nitrogen of the guanidinium group of Lys34. Similarly in the BDNF dimer, the terminal nitrogen of the guanidinium groups of Arg97 and Arg101 are an average of 7.0 Å from each other and the epsilon nitrogen of Lys46 is an average 14.7 Å from the terminal nitrogen of the guanidinium group of Arg97. Interestingly, these distances correspond quite closely to the distances between the sulfate groups on the CS-E tetrasaccharide. The distance between the sulfur atoms on the four and the six position of the same sugar is on average 5.5 Å while the furthest distance between the sulfate atoms on the four position of the first sugar and the six position of the

third sugar is 15.0 Å and the average distance between the sulfate atoms on the two sugars is 12.9 Å. Indeed, these characteristic distances between the positively charged side chains more accurately characterize the CS-E binding site than does the primary sequence characterization. For example, the CS-E binding site on the TNF trimer structure contains four positively charged residues, Arg103 and Arg138 on one monomer of the trimer and Lys65 and Lys112 on a second monomer of the trimer, which does not fit well into the proposed primary sequence identification of a CS-E binding site. Yet, the distance between the epsilon nitrogen of Lys65 and Lys112 is 4.9 Å and the average distance between epsilon nitrogen on Lys112 in the first monomer and the terminal guanidinium nitrogen on Arg138 in the second monomer is 15.5 Å, again corresponding well to the distances between sulfate groups on CS-E.

Yet, surprisingly the third basic amino acid that lies approximately 15 Å away from the first two is almost exclusively found at a minimum of 10 Å away from any other basic amino acids in the CS-E binding site. This suggests that CS-E may need a strong positively charged region, consisting of two or more basic residues, to anchor it and then other, less positively charged regions, to orient it. Indeed the mutations that distinguish the CS-E binding sites on BDNF, NGF, NT-3, and NT-4/5 occur within one or both of the two closely positioned basic amino acids, whereas the third more distantly placed basic amino acids is more conserved between the proteins. Such a situation could explain the preference of these proteins for CS-E molecules over other chondroitin sulfate molecules with less concentrated charge. In particular, it would suggest that CS-E is able to most strongly bind to these proteins due to the ability of the negatively-charged sulfates on the four and six positions of the GlcN sugar to make strong salt bridges with

two correctly positioned positively-charged amino acids in the CS-E binding site for these proteins. This strong salt bridge may be required, in part, because unlike other small molecules, GAGs do not bind into a deep binding pocket and thus fail to make the full set of molecular interactions that they would otherwise make in a deep pocket. This would imply that a full CS-E tetrasaccharide may not be necessary for binding to these molecules but rather a tetrasaccharide consisting of one CS-E disaccharide unit and another singly charged CS-motif may be sufficient.

Non-traditional binding sites. The predicted CS-E binding sites for the Nogo receptor and the GDNF receptor do not immediately fit within this CS-E binding site rubric. The distance between any two of the positively-charged amino acids in these binding sites is greater than 10 Å, making them too far away to make meaningful contacts with the sulfate groups on the four and six position of a CS-E sugar. Nevertheless these binding sites each have amino acids that could rotate upon CS-E binding to afford a more characteristic CS-E binding site that would provide strong interactions between the CS-E sulfate groups and basic residues.

Limitations. Since the predicted CS-E binding sites are based on calculations, they suffer from a number of limitations. Beyond the inherent limitations in trying to model a complex system with a limited number of equations and variables, other limitations exist in this system. In particular, for most of the proteins, a greater region of the protein was used for the CS-E binding experiments than was resolved in the crystal structure. That is, CS-E could be binding to parts of the protein not found in the crystal structures and thus the predicted CS-E binding site based on the crystal structure would be incorrect. An

example of where this almost becomes a problem is in the case of GDNF. The CS-E binding studies were performed of amino acids 1 – 134 of GDNF, whereas the crystal structures of GDNF begin at amino acids 40, 32, and 34 and end at amino acids 134, 134, and 134 for chain B of 3FUB, chain D of 3FUB, and 2V5E, respectively. The predicted CS-E binding site based on the structures that start at amino acids 32 and 34 predict that CS-E interacts heavily with the amino acids 32 through 40 thus making this region a key part of the predicted CS-E binding site. Indeed, calculating the CS-E binding site based on the GDNF structure that starts at amino acid 40 predicts that the CS-E binding site is on the opposite side of the protein to that predicted from the structures that start at amino acids 32 and 34. Thus, if only the structure for GDNF that starts at amino acid 40 were available, then the predicted binding site would likely be incorrect.

Methods

Forcefields: The Dreiding force field[58], adapted to include sulfate groups, was used throughout the calculations. The force field was modified by optimizing the bond lengths and angles of a model CH_3OSO_3Na system through quantum mechanics (Jaguar)[59] and adjusting the Dreiding force field parameters based on this optimum geometry. All charges for the ligands were calculated using the charge equilibrium (QEq)[60] method. CHARM22[61] charges were used for the protein.

Molecular dynamics simulations: For each tetrasaccharide, charge equilibrium (QEq)[60] charges were assigned and 1,000 conformations were generated using a Boltzmann jump method with rotation around the glycosidic bonds followed by structural minimization. The resulting conformations were sorted into five groups by root mean square deviation

in coordinates and ranked by their potential energies. 300 ps of explicit water molecular dynamics was run at 300 K on the two lowest-energy conformations from each of the five groups. For each of the ten molecular dynamics runs, the tetrasaccharide conformations were averaged from the last 100 ps and their potential energies were calculated with explicit solvation. The lowest-energy structure among these ten was used to represent the predicted solution structure. All Boltzmann jumps and the molecular dynamics calculations were performed using Cerius2 (Accelrys Inc.)[62].

Preparation of the Proteins: The pdb file for each protein was downloaded from the RCSB Protein Data Bank (www.pdb.org). Water and other non-protein molecules were removed, missing residues were added using Swiss PDB Viewer, and hydrogen were added using the WhatIf program. CHARM22 charges were added, NaCl atoms were added to neutralize the protein, and the protein was minimized using the MPSim program in SGB implicit solvation.

Binding Site Calculations: We determined the binding sites as per previously described[63] with the following changes: We used a buried surface criteria of 10%, the parameter 'Grow' rather than 'Pass', the *radmax* = 5.0, and the *dotlim* = -0.5. Once the potential binding sites were identified and the top docked conformations and corresponding binding sites were ranked by energy, the predicted GAG binding site was identified by the following. The top twenty-five docked conformations and corresponding binding sites were tabulated and the sum of the inverse energy ranks for each binding site was determined. Any binding site in which this value was greater than zero was considered a GAG binding site. To determine which residues contributed to the predicted binding site,

the five lowest-energy dock conformations for GAG binding site were determined and those residues within 4 Å of any of those conformations were taken to be part of the potential GAG binding site. Similarly to determine the heparin binding site from the heparin containing crystal structures, residues within 4 angstroms of heparin were determined and were considered to contribute to the heparin binding site.

References

1. Carulli, D., Laabs, T., Geller, H.M. & Fawcett, J.W. Chondroitin sulfate proteoglycans in neural development and regeneration. *Curr Opin Neurobiol* **15**, 116-20 (2005).
2. Bulow, H.E. & Hobert, O. Differential sulfations and epimerization define heparan sulfate specificity in nervous system development. *Neuron* **41**, 723-36 (2004).
3. Goretzki, L., Lombardo, C.R. & Stallcup, W.B. Binding of the NG2 proteoglycan to kringle domains modulates the functional properties of angiostatin and plasmin(ogen). *J Biol Chem* **275**, 28625-33 (2000).
4. Trybala, E. et al. Structural and functional features of the polycationic peptide required for inhibition of herpes simplex virus invasion of cells. *Antiviral Res* **62**, 125-34 (2004).
5. Sasisekharan, R., Shriver, Z., Venkataraman, G. & Narayanasami, U. Roles of heparan-sulphate glycosaminoglycans in cancer. *Nat Rev Cancer* **2**, 521-8 (2002).
6. Huang, W.C. et al. Chondroitinase ABC promotes axonal re-growth and behavior recovery in spinal cord injury. *Biochem Biophys Res Commun* **349**, 963-968 (2006).
7. Bradbury, E.J. et al. Chondroitinase ABC promotes functional recovery after spinal cord injury. *Nature* **416**, 636-40 (2002).
8. Pangalos, M.N., Shioi, J., Efthimiopoulos, S., Wu, A. & Robakis, N.K. Characterization of appican, the chondroitin sulfate proteoglycan form of the Alzheimer amyloid precursor protein. *Neurodegeneration* **5**, 445-51 (1996).
9. Gama, C.I. & Hsieh-Wilson, L.C. Chemical approaches to deciphering the glycosaminoglycan code. *Curr Opin Chem Biol* **9**, 609-19 (2005).
10. Capila, I. & Linhardt, R.J. Heparin-protein interactions. *Angew Chem Int Ed Engl* **41**, 391-412 (2002).
11. Kinoshita-Toyoda, A. et al. Structural determination of five novel tetrasaccharides containing 3-O-sulfated D-glucuronic acid and two rare oligosaccharides containing a beta-D-glucose branch isolated from squid cartilage chondroitin sulfate E. *Biochemistry* **43**, 11063-74 (2004).
12. Gama, C.I. et al. Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. *Nat Chem Biol* **2**, 467-73 (2006).

13. Tully, S.E., Rawat, M. & Hsieh-Wilson, L.C. Discovery of a TNF-alpha antagonist using chondroitin sulfate microarrays. *J Am Chem Soc* **128**, 7740-1 (2006).
14. Yates, E.A. et al. Protein-GAG interactions: new surface-based techniques, spectroscopies and nanotechnology probes. *Biochem Soc Trans* **34**, 427-30 (2006).
15. Shaya, D. et al. Crystal structure of heparinase II from Pedobacter heparinus and its complex with a disaccharide product. *J Biol Chem* **281**, 15525-35 (2006).
16. Tan, K. et al. The structures of the thrombospondin-1 N-terminal domain and its complex with a synthetic pentameric heparin. *Structure* **14**, 33-42 (2006).
17. Mohammadi, M., Olsen, S.K. & Goetz, R. A protein canyon in the FGF-FGF receptor dimer selects from an a la carte menu of heparan sulfate motifs. *Curr Opin Struct Biol* **15**, 506-16 (2005).
18. Zhu, X. et al. Three-dimensional structures of acidic and basic fibroblast growth factors. *Science* **251**, 90-3 (1991).
19. Zhu, X., Hsu, B.T. & Rees, D.C. Structural studies of the binding of the anti-ulcer drug sucrose octasulfate to acidic fibroblast growth factor. *Structure* **1**, 27-34 (1993).
20. Pellegrini, L., Burke, D.F., von Delft, F., Mulloy, B. & Blundell, T.L. Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature* **407**, 1029-34 (2000).
21. Schlessinger, J. et al. Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization. *Mol Cell* **6**, 743-50 (2000).
22. Mourey, L. et al. Crystal structure of cleaved bovine antithrombin III at 3.2 A resolution. *J Mol Biol* **232**, 223-41 (1993).
23. Bewley, M.C., Boustead, C.M., Walker, J.H., Waller, D.A. & Huber, R. Structure of chicken annexin V at 2.25-A resolution. *Biochem* **32**, 3923-9 (1993).
24. Burger, A. et al. The crystal structure and ion channel activity of human annexin II, a peripheral membrane protein. *J Mol Biol* **257**, 839-47 (1996).
25. Faham, S., Linhardt, R.J. & Rees, D.C. Diversity does make a difference: fibroblast growth factor-heparin interactions. *Curr Opin Struct Biol* **8**, 578-86 (1998).
26. Rodriguez-Carvajal, M.A., Imberty, A. & Perez, S. Conformational behavior of chondroitin and chondroitin sulfate in relation to their physical properties as inferred by molecular modeling. *Biopolymers* **69**, 15-28 (2003).
27. Mulloy, B., Forster, M.J., Jones, C. & Davies, D.B. N.m.r. and molecular-modelling studies of the solution conformation of heparin. *Biochem J* **293 (Pt 3)**, 849-58 (1993).
28. Bitomsky, W. & Wade, R.C. Docking of Glycosaminoglycans to Heparin-Binding Proteins: Validation for aFGF, bFGF, and Antithrombin and Application to IL-8. *J. Am. Chem. Soc.* **121**, 3004-3013 (1999).
29. Forster, M. & Mulloy, B. Computational approaches to the identification of heparin-binding sites on the surfaces of proteins. *Biochem Soc Trans* **34**, 431-434 (2006).

30. Sachchidanand et al. Mapping the heparin-binding site on the 13-14F3 fragment of fibronectin. *J Biol Chem* **277**, 50629-35 (2002).
31. Gandhi, N.S., Coombe, D.R. & Mancera, R.L. Platelet endothelial cell adhesion molecule 1 (PECAM-1) and its interactions with glycosaminoglycans: 1. Molecular modeling studies. *Biochemistry* **47**, 4851-62 (2008).
32. Millane, R.P., Mitra, A.K. & Arnott, S. Chondroitin 4-sulfate: comparison of the structures of the potassium and sodium salts. *J Mol Biol* **169**, 903-20 (1983).
33. Khunrae, P., Philip, J.M.D., Bull, D.R. & Higgins, M.K. Structural comparison of two CSPG-binding DBL domains from the VAR2CSA protein important in malaria during pregnancy. *J Mol Biol* **393**, 202-13 (2009).
34. Tully, S.E. & Hsieh-Wilson, L.C. Unpublished data. (2010).
35. Floriano, W.B. et al. Modeling the human PTC bitter-taste receptor interactions with bitter tastants. *J Mol Model* **12**, 931-41 (2006).
36. Floriano, W.B., Vaidehi, N. & Goddard, W.A. Making sense of olfaction through predictions of the 3-D structure and function of olfactory receptors. *Chem Senses* **29**, 269-90 (2004).
37. Faham, S., Hileman, R.E., Fromm, J.R., Linhardt, R.J. & Rees, D.C. Heparin structure and interactions with basic fibroblast growth factor. *Science* **271**, 1116-20 (1996).
38. DiGabriele, A.D. et al. Structure of a heparin-linked biologically active dimer of fibroblast growth factor. *Nature* **393**, 812-7 (1998).
39. Li, Z., Kienetz, M., Cherney, M.M., James, M.N.G. & Brömme, D. The crystal and molecular structures of a cathepsin K:chondroitin sulfate complex. *J Mol Biol* **383**, 78-91 (2008).
40. Tracey, D., Klareskog, L., Sasso, E.H., Salfeld, J.G. & Tak, P.P. Tumor necrosis factor antagonist mechanisms of action: a comprehensive review. *Pharmacol Ther* **117**, 244-79 (2008).
41. Banner, D.W. et al. Crystal structure of the soluble human 55 kd TNF receptor-human TNF beta complex: implications for TNF receptor activation. *Cell* **73**, 431-45 (1993).
42. Mukai, Y. et al. Structure-function relationship of tumor necrosis factor (TNF) and its receptor interaction based on 3D structural analysis of a fully active TNFR1-selective TNF mutant. *J Mol Biol* **385**, 1221-9 (2009).
43. Kaplan, D.R. & Miller, F.D. Neurotrophin signal transduction in the nervous system. *Curr Opin Neurobiol* **10**, 381-91 (2000).
44. Rogers, C. & Hsieh-Wilson, L.C. Unpublished data. (2010).
45. Numakawa, T. et al. BDNF function and intracellular signaling in neurons. *Histol Histopathol* **25**, 237-58 (2010).
46. Pae, C.U., Marks, D.M., Han, C., Patkar, A.A. & Steffens, D. Does neurotropin-3 have a therapeutic implication in major depression? *Int J Neurosci* **118**, 1515-22 (2008).
47. Patapoutian, A. & Reichardt, L.F. Trk receptors: mediators of neurotrophin action. *Curr Opin Neurobiol* **11**, 272-80 (2001).
48. Sofroniew, M.V., Howe, C.L. & Mobley, W.C. Nerve growth factor signaling, neuroprotection, and neural repair. *Annu Rev Neurosci* **24**, 1217-81 (2001).

49. Berkemeier, L.R. et al. Neurotrophin-5: a novel neurotrophic factor that activates trk and trkB. *Neuron* **7**, 857-66 (1991).
50. Kolbeck, R., Jungbluth, S. & Barde, Y.A. Characterisation of neurotrophin dimers and monomers. *Eur J Biochem* **225**, 995-1003 (1994).
51. Irie, F., Okuno, M., Matsumoto, K., Pasquale, E.B. & Yamaguchi, Y. Heparan sulfate regulates ephrin-A3/EphA receptor signaling. *Proc Natl Acad Sci USA* **105**, 12307-12312 (2008).
52. Ibáñez, C.F. et al. Disruption of the low affinity receptor-binding site in NGF allows neuronal survival and differentiation by binding to the trk gene product. *Cell* **69**, 329-41 (1992).
53. Iwasaki, W. et al. Solution structure of midkine, a new heparin-binding growth factor. *The EMBO Journal* **16**, 6936-46 (1997).
54. Parkash, V. et al. The structure of the glial cell line-derived neurotrophic factor-coreceptor complex: insights into RET signaling and heparin binding. *J Biol Chem* **283**, 35164-72 (2008).
55. Baloh, R.H., Enomoto, H., Johnson, E.M. & Milbrandt, J. The GDNF family ligands and receptors - implications for neural development. *Curr Opin Neurobiol* **10**, 103-10 (2000).
56. He, X.L. et al. Structure of the Nogo receptor ectodomain: a recognition module implicated in myelin inhibition. *Neuron* **38**, 177-85 (2003).
57. Fournier, A.E., GrandPre, T. & Strittmatter, S.M. Identification of a receptor mediating Nogo-66 inhibition of axonal regeneration. *Nature* **409**, 341-6 (2001).
58. Mayo, S.L., Olafson, B.D. & Goddard, W.A. Dreiding - a Generic Force-Field for Molecular Simulations. *Journal of Physical Chemistry* **94**, 8897-8909 (1990).
59. Greeley, B.H. et al. New Pseudospectral Algorithms for Electronic-Structure Calculations - Length Scale Separation and Analytical 2-Electron Integral Corrections. *Journal of Chemical Physics* **101**, 4028-4041 (1994).
60. Rappe, A.K. & Goddard, W.A. Charge Equilibration for Molecular-Dynamics Simulations. *Journal of Physical Chemistry* **95**, 3358-3363 (1991).
61. MacKerell, A.D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* **102**, 3586-3616 (1998).
62. Lim, K.T. et al. Molecular dynamics for very large systems on massively parallel computers: The MPSim program. *Journal of Computational Chemistry* **18**, 501-521 (1997).
63. Kalani, M.Y. et al. The predicted 3D structure of the human D2 dopamine receptor and the binding site and binding affinities for agonists and antagonists. *Proc Natl Acad Sci U S A* **101**, 3815-20 (2004).