DARWINDOCK & GAG-DOCK: METHODS AND APPLICATIONS FOR SMALL MOLECULE DOCKING

Thesis by

Adam Reid Griffith

In Partial Fulfillment of the Requirements for
the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2017
(Defended September 26, 2016)

© 2016

Adam Reid Griffith

All Rights Reserved

ACKNOWLEDGEMENTS

I would first and foremost like to thank my advisor, William A. Goddard III, for his guidance, support, and patience during my time at Caltech. I continue to be in awe of his knowledge, wisdom, and passion for knowledge.

I would also like to thank my thesis committee: Dr. James Heath, Dr. Douglas Rees, and Dr. Linda Hsieh-Wilson. They have been extremely patient with me, much more than they should have been.

My interactions with Dr. Ravi Abrol as friend and mentor have been invaluable. He has provided guidance and direction on most of my projects, as well as providing encouragement and a friendly ear.

The countless members of the MSC and especially the Biogroup have been my friends and colleagues. Chief among them have been Dr. Soo-Kyung Kim, Dr. Ismet Caglar Tanrikulu, Dr. Heather Wiencko, Dr. Julius Su, Dr. John Keith, Dr. Jenelle Bray, Dr. Andrea Kirkpatrick, Dr. Caitlin Scott, Dr. Vaclav Cvicek, Dr. Fan Liu, and… many others. (I sincerely apologize to those I haven't mentioned.)

I have been blessed with many friends in Pasadena. Michael Torrice, Graef Allen, Michael D. Adams, Michelle Robbins, Heidi Privett, Dillon Ross, Ariele Hanek, Gavin and Jill Doughtie, Heather Wiencko, Ger Norton, Joel Louwsma, Greg ver Steeg, Jeff Krimmel, and Eve Stenson to name a few.

Last, certainly not least, I have had endless support and encouragement from my family, both immediate and extended. My cousins Margaret Griffith and Jamison Carter and Suzanna Griffith had never met me before I moved to California, yet they took me in during my first Thanksgiving out here and have been friends ever since. I've thoroughly enjoyed spending time discussing art and science with you. Thank you for providing an escape when I needed one. Nobody has been more supportive of me than my parents, Boyce and Diana Griffith, and my brother, Boyce E. Griffith. I wouldn't be here or doing what I'm doing without you. I can't even put into words how much you've done for me, or how much you mean to me.

ABSTRACT

Computational modeling is an effective tool in studying complex biological systems. Docking of small molecule ligands in particular is useful both in understanding the functioning of proteins as well as in the development of pharmaceuticals. Together with experiment, modeling can often provide a thorough picture of a given system. Computation can often provide details that are difficult or impossible to determine experimentally, while experiments provide guidance on what calculations are useful or interesting. Our goal is to extend computational modeling, specifically ligand docking, to systems not previously possible, such as the challenging glycosaminoglycan (GAG) systems. In order to do this it was first necessary to develop an automatic way of performing docking without extensive user input and experimental knowledge to narrow the list of candidate poses. DarwinDock represents our efforts in this respect. It is a method for small-molecule docking that separates pose generation and scoring into separate stages, which allows for complete binding site sampling followed by efficient, hierarchical sampling. Our convergence criteria for complete sampling allows for diverse systems to be studied without prior knowledge of how large a set of poses needs to be to span a given binding site, making the procedure more automatic. We also replace bulky, nonpolar residues with alanine, which we refer to as "alanization". This allows the ligand to interact more closely with polar sidechains, which help to orient the ligand. Additionally, alanization reduces the impact of incorrect sidechain placement on ligand placement, a concern that sometimes requires user intervention. With DarwinDock working for standard small molecules, it was then necessary to modify the procedure to work on challenging GAG ligands, which are large and have strong negative charges. A modification to DarwinDock – GAG-Dock – allows the method to be applied to GAGs and protein surface interactions. GAGs are large, linear polysaccharides with strong negative charge. They typically interact with the surfaces of proteins, rather than the cavities favored by most small-molecule drugs. GAG-Dock systematically samples the protein surface for unknown binding sites and modifies the pose generation to allow for large, surface-interacting ligands. GAG-Dock allowed us to study several systems important for neuronal development and answer interesting questions posed by experiment. Finally, we needed a

way to validate our predictions for GAG binding sites. We used a systematic approach to identify sets of beneficial mutations to the GAG binding sites by building up from individual *in silico* mutations. Standard mutation experiments typically employ large mutations, such as arginine to alanine, which decrease or destroy binding. However, such information is not always definitive, as large mutations can have wide-ranging effects beyond direct protein-ligand interactions. Mutations that *increase* binding, however, are less ambiguous because they must form new interactions with the ligand in order to affect binding energies or affinity. Therefore, we have identified and proposed sets of mutations for our GAG predictions for PTPs, NgR1, NgR3, and EphB3. We encourage our experimentalist colleagues to try these mutations and validate our predictions.

CONTRIBUTIONS

Chapter 2: Adam R. Griffith, Ismet Caglar Tanrikulu, Ravinder Abrol, William A. Goddard III, THE DARWINDOCK ALGORITHM FOR SMALL MOLECULE DOCKING

A.R.G. was involved in method development, wrote software to handle input/output of DOCK 6 calls, implemented the CNS clustering algorithm for the calculations reported, made modifications to the main DarwinDock program, prepared the DUD systems for calculations, performed extensive testing during method development, and performed all calculations reported

I.C.T. was involved in method development, wrote the main DarwinDock program, prepared DUD systems for testing during development, and performed testing on the CSN method

R.A. was involved in method development and wrote a development implementation of the CNS algorithm for testing

W.A.G. provided guidance for the project and valuable feedback

Chapter 3: Adam R. Griffith, Claude J. Rogers, Ravinder Abrol, Greg Miller, William A. Goddard III, Linda C. Hsieh-Wilson, PREDICTING GLYCOSAMINOGLYCAN-SURFACE PROTEIN INTERACTIONS: IMPLICATIONS FOR STUDYING AXONAL GROWTH

A.R.G. was involved in GAG-Dock method development, performed extensive testing, wrote software to automate docking calculations, developed the method for predicting mutations, and performed the mutation calculations

C.J.R. was involved in method development, prepared and evaluated the validation systems, prepared GAG ligands for docking, and performed calculations for the predicted systems

R.A. provided assistance with homology modeling and guidance

G.M. assisted with preparation of GAG ligands

W.A.G. provided guidance for the project and valuable feedback

L.C.H.-W. initiated the project and provided guidance on identifying relevant systems for calculation

Chapter 4: Adam R. Griffith, Claude J. Rogers, Ravinder Abrol, Greg Miller, William A. Goddard III, Linda C. Hsieh-Wilson, EXPLORING NOVEL INTERACTIONS BETWEEN CHONDROITIN SULFATE AND THE EPHB3 RECEPTOR

A.R.G. implemented APBS mapping of charges to *sphgen* spheres, performed all docking calculations, and performed mutation analysis

C.J.R. prepared GAG ligands for docking and provided extensive discussion and analysis relevant to the project

R.A. generated EphB2 and EphB3 homology models and provided guidance

G.M. assisted with preparation of GAG ligands

W.A.G. provided guidance for the project and valuable feedback

L.C.H.-W. initiated the project

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

List of Figures & Tables: Chapter 2

List of Figures & Tables: Chapter 3

List of Figures & Tables: Chapter 4

Chapter 1

SMALL MOLECULE DOCKING

Introduction

Computational modeling has become an important part of studying complex biological systems. It often provides details and information that either cannot or cannot easily be obtained via experiments. Rather, computation and experiment often work hand-in-hand, with insights from one driving questions for the other. A recent paper by Brian Shoichet[1] exemplifies this. G-protein coupled receptors (GPCRs) produce signaling via two pathways: a G protein or β-arrestin. The same GPCR can use both pathways to produce different effects. Experiments suggest that while the primary analgesic effect of the μ-opioid receptor (mOR) is carried out via G protein signaling, side effects such as the buildup of tolerance may be due to the β-arrestin pathway[2-4]. This suggests that a drug capable of activating the G protein pathway and *not* the β-arrestin pathway would be an improvement on current opioid drugs. While it may be possible to identify such a drug using an experimental approach, such an approach would be lengthy and expensive. Computation, on the other hand, is ideally suited to this sort of task. Shoichet, *et. al.*, used a virtual ligand screening (VLS) method to screen 3 million drug candidates against mOR to identify drug candidates that interact with the protein in a manner different than current opioid drugs. After refinement of the candidates, one drug was identified for testing. The result of the project was a drug that preferentially activates mOR G protein signaling over β-arrestin signaling. This drug will now provide a starting point for development of better mOR analgesics as well as providing a tool for the study of the differences of G protein and

β-arrestin signaling. Such a project likely would not have been successful without the use of computational methods for small-molecule docking.

Purpose of the Study

Our goal in this work is to tackle challenging computational problems, such as those related to glycosaminoglycan (GAG) docking, to answer questions raised by experiment, and to help guide new experiments. GAGs are large, linear polysaccharides that are heavily negatively charged, which makes them difficult to treat computationally. Furthermore, they typically interact with the surfaces of proteins, instead of the deeper cavities typically used by standard small-molecule ligands. These surface-protein interactions are less well defined, which requires much more extensive sampling to identify a binding site than a protein with a known binding cavity. We therefore need an automated docking method that is capable of identifying interesting poses and binding sites with little intervention by the user and little or no experimental information. We further need to be able to account for the size and charge of GAG ligands, as well as the ability to predict ligand binding to protein surfaces. And finally, we need a way to validate our predictions.

Our strategy for automated docking is DarwinDock. DarwinDock divides the problem of docking into two stages: geometry and scoring. The geometry or "completeness" stage thoroughly samples a putative binding site without performing any energy calls or using any scoring methods. Our method for generating a complete set of ligand poses differs from other docking methods in that the number of poses generated depends on the system, and not a default or assumed number of poses. Instead, poses are generated until a convergence threshold is met. This is important because each system – protein, ligand, or

binding site – is different. There is no default number of poses that will work in all situations. The scoring stage of DarwinDock efficiently and accurately evaluates the poses that are generated. Efficiency is derived from our hierarchical scoring method. Poses are clustered into families, and only the family head (centroid) pose is scored initially. This family head score is used to eliminate 90% of the families from consideration, significantly reducing computational cost. The children of the remaining 10% of families are then evaluated, and a final set of 120 best poses is output. The goal of the method is for the final set of 120 poses to contain correct poses for further, more detailed evaluations. We will describe the DarwinDock method and validate it against a set of known ligand-protein co-crystals.

We address the complexity of GAG ligand binding with GAG-Dock. GAG-Dock is a variation of DarwinDock that is designed to work for an extremely challenging and interesting problem: the study of glycosaminoglycan (GAG) binding to proteins. GAGs are linear polysaccharides that carry a strong negative charge. Their size and charge makes them particularly difficult candidates for docking. Furthermore, they typically interact with the surfaces of proteins. Typical small-molecule ligands tend to bury into cavities in a protein, which provides a clear, contained region to sample during docking. Interactions on the surface of a protein are less well defined and require much greater computational sampling. However, the interactions between GAGs and proteins are also very interesting. For instance, GAGs have been shown to be involved in directing neuronal development via interactions with the PTPs, NgR1, and NgR3 receptors. GAG-Dock is designed to study just these sorts of systems, for which structural data is often not known. We will show that

GAG-Dock is effective for these challenging systems by validating it against crystal structures with bound GAG polysaccharides. We will then apply it to the systems PTPs, NgR1, NgR3, and EphB3 in order to predict the binding sites of several GAG ligands and answer interesting questions posed by experimental evidence.

Finally, we use a systematic approach to identify sets of beneficial mutations to validate our GAG binding sites. Mutations that decrease or eliminate binding are often easy, but ambiguous. Mutating an arginine to an alanine could simply be removing a contact with the ligand, or it could be fundamentally altering the protein structure. Beneficial mutations that *increase* ligand binding energy/affinity, however, require that the mutation stabilize the binding site or provide new interactions with the ligand. They provide a much clearer signal. Therefore we identify sets of mutations for each of our predicted GAG cases where no crystal structure is known. This is particularly important for our CS-E/EphB3 predictions where even the general binding region of the protein was not known.

Overall, we show that we have developed methods capable of handling challenging, relevant systems that could not be approached before. Furthermore, we show that computational modeling and experiment can work together to guide and complement each other.

Outline

Chapter 2: Description of the DarwinDock method and validation against the DUD set, a set of protein-ligand co-crystals intended for testing small-molecule ligand docking methods.

Chapter 3: Description of GAG-Dock, validation against known GAG-protein co-crystals, and application to PTPs, NgR1, and NgR3 for the GAG ligands CS-A, CS-D, CS-E, and heparin.

Chapter 4: Application of GAG-Dock to the novel system of EphB3. We predict the binding site of CS-E to EphB3. We also explain why CS-E and not CS-A binds to EphB3, and why neither binds to EphB2.

References

1. Manglik, A., Lin, H., Aryal, D. K., McCorvy, J. D., Dengler, D., Corder, G., et al. (2016). Structure-based discovery of opioid analgesics with reduced side effects. *Nature*, *537*(7619), 185–190. http://doi.org/10.1038/nature19112
2. Bohn, L. M., Gainetdinov, R. R., Lin, F.-T., Lefkowitz, R. J., & Caron, M. G. (2000). μ-Opioid receptor desensitization by β-arrestin-2 determines morphine tolerance but not dependence. *Nature*, *408*(6813), 720–723. http://doi.org/10.1038/35047086
3. Bohn, L. M., Lefkowitz, R. J., Gainetdinov, R. R., Peppel, K., Caron, M. G., & Lin, F.-T. (1999). Enhanced Morphine Analgesia in Mice Lacking β-Arrestin 2. *Science*, *286*(5449), 2495–2498. http://doi.org/10.1126/science.286.5449.2495
4. Raehal, K. M., Walker, J. K. L., & Bohn, L. M. (2005). Morphine Side Effects in β-Arrestin 2 Knockout Mice. *Journal of Pharmacology and Experimental Therapeutics*, *314*(3), 1195–1201. http://doi.org/10.1124/jpet.105.087254
5. Shen, Y., Tenney, A. P., Busch, S. A., Horn, K. P., Cuascut, F. X., Liu, K., et al. (2009). PTPσ Is a Receptor for Chondroitin Sulfate Proteoglycan, an Inhibitor of Neural Regeneration. *Science*, *326*(5952), 592–596. http://doi.org/10.1126/science.1178310
6. Dickendesher, T. L., Baldwin, K. T., Mironova, Y. A., Koriyama, Y., Raiker, S. J., Askew, K. L., et al. (2012). NgR1 and NgR3 are receptors for chondroitin sulfate proteoglycans. *Nature Neuroscience*, *15*(5), 703–712. http://doi.org/10.1038/nn.3070

Chapter 2

THE DARWINDOCK ALGORITHM FOR SMALL MOLECULE DOCKING

Adam R. Griffith, Ismet Caglar Tanrikulu, Ravinder Abrol, William A. Goddard III

Introduction

The first step in understanding how binding of a ligand to a protein affects its function is to determine its binding site, conformation, and binding energy. The ultimate arbiter for the structure is the x-ray structure, and the arbiter for the binding energy uses radioisotopes to measure an equilibrium constant or competitive binding to obtain an IC50. However, such experimental procedures are far too slow for broad searching for new ligand scaffolds (e.g. virtual ligand screening) or for optimizing hits from such a screening. Here we use theory and computation to identify the most likely binding sites and configurations and to rank ligands in terms of binding (or, ideally, in terms of function). This computational process is referred to as "docking". Various methods for docking have been developed over the last 50 years and are widely practiced in industry and academia. A typical application involves a coupling of theory and experiment where a number of putative poses from energy minimization and molecular dynamics computations might be tested by doing mutation experiments to identify which poses are most likely correct, followed by computational modifications of the ligand to improve binding by competitive binding experiments to select the best ligand.

My goal has been to develop a purely computational algorithm that can predict the best ligands and binding sites, without the intervention of experiments. This requires a very

complete sampling of ligand poses throughout the possible binding region and requires an energy scoring that is accurate enough to discriminate the binding strength of various ligands to various binding sites. But carrying out accurate calculations of binding strength generally involves extensive calculations for every possible pose and every possible ligand conformation, which is not practical. DarwinDock is an algorithm that I in collaboration with others in the Goddard group have developed to solve this conundrum.

The DarwinDock method is a new strategy for docking that we refer to as Complete Sampling-Hierarchical Scoring (CS-HS). The idea is:

1. **Alanization** – First, properly prepare the protein system to minimize the chances of bad contacts due to improper sidechain placement and maximize the interaction of the ligand with polar groups in the protein. We achieve this via a process referred to as "alanization", which is the replacing of bulky, nonpolar sidechains with alanine.
2. **Completeness** – Then, a complete set of ligand poses is generated that completely samples the putative binding site, but is done so quickly by eschewing any energy calls. Poses are generated in iterations of 5000 and clustered by RMSD into families. When the number of new families reaches a convergence threshold we consider this to be a complete sampling of the binding site. This typically yields ~50,000 poses.
3. **Scoring** – Finally, the poses are scored in a hierarchical manner in order to minimize the number of energy calls necessary. We score each of the family heads

(as determined by RMSD) using a single energy calculation. Based on these energies we eliminate 90% of the families. The children of the remaining families are then scored and all remaining poses are ranked. From this set we choose the best 120 poses based on the scoring energy and output them for detailed examination

This overall procedure is referred to as DarwinDock. DarwinDock is aimed at being automatic, relying on our scoring algorithms to select interesting ligand candidates without human intervention. This makes DarwinDock useful for virtual ligand screening (VLS) applications where the DarwinDock method might be used to rank the output of a pharmacophore-driven VLS process.

Indeed, as I have been developing and optimizing DarwinDock over the last few years, it has been used in the Goddard group for numerous successful applications.[1-9]

The goal of this chapter is to document and explain the full DarwinDock procedure, in particular how the optimum settings for the procedure are determined. That is, to identify the settings that provide the highest probability of success with minimal computational time.

Evaluating the performance of such an algorithm is difficult since biological systems are complex; therefore, we use sets of pre-determined systems for validation. Specifically, we use the Directory of Useful Decoys[10] (DUD set), which is a selection of 40 diverse systems

based on high quality x-ray structures of ligand/protein co-crystals and is intended for validation of docking programs and methods.

DarwinDock

The first stage of DarwinDock – **alanization** – addresses a critical problem in docking a ligand to a protein. The optimum binding site and conformation of the ligand depend on the conformations of the protein sidechains; simultaneously, the conformations of the protein sidechains depend on the binding site and conformation of the ligand. Thus we need to dock the ligand simultaneously with optimized protein sidechain conformations. We have solved this "chicken-egg" problem by replacing the bulky, nonpolar sidechains with alanine prior to docking, which we refer to as "alanization". We consider valine, leucine, isoleucine, methionine, phenylalanine, tyrosine, and tryptophan to be primarily nonpolar in character and thus less likely to be essential to orienting a ligand in a binding site. Furthermore, these residues are bulky enough that they might block significant portions of the binding site if placed incorrectly, eliminating what might be the ideal binding pose. Alanizing these residues additionally opens up the binding site, allowing the ligand to sample a larger space in the binding site without being bumped so that it has the best chance of interacting more directly with polar sidechains. The tradeoffs are that we miss a significant part of van der Waals (VDW) interactions with the ligand, and we must do greater sampling to span the more open binding site. While not a required part of preparing a system for DarwinDock, it is recommended and we consider this as the default approach.

The second stage of DarwinDock addresses **complete sampling** of the binding. Ligand poses are generated in iterations of 5000 using DOCK 6[11] and are clustered by RMSD into families with a diversity of 2Å. When the number of new families falls below a threshold of 2% we consider the set of poses to be a complete sampling of the putative binding site. This procedure typically results in ~50,000 poses and ~5,000 families. The only evaluation performed on poses during the "completeness" stage of DarwinDock is a bump test to ensure that the ligand does not clash with the protein. *No energy calls* are made during this stage.

The number of poses generated in the completeness stage is quite large, making evaluating accurate energies for each pose impractical. Instead we use a hierarchical approach for scoring the poses in the final, **scoring** stage of DarwinDock, beginning with scoring of the family heads. The head of each 2Å family from the completeness stage is selected as the centroid of the family based on the heavy-atom RMSD and then its energy is evaluated. We use the DREIDING[12] forcefield in MPSim[13] to evaluate the non-bond energy between the ligand and the rest of the protein. Based on the energy of each family head, we eliminate 90% of the families and focus on the children of the remaining 10%. Assuming that the family head is broadly representative of the children, this allows us to dramatically reduce the number of energy evaluations necessary to finally select the best pose. At the end of the scoring stage we obtain 120 best poses for further, more accurate evaluations.

We the use closest-neighbor seeded (CNS) algorithm to cluster poses based on heavy-atom RMSD. The pairwise RMSDs for all ligand poses are calculated and the pairs are ordered

by increasing RMSD. Each pose is initially placed in its own family. The list of RMSDs is then traversed for each *i, j* pair of poses. If the RMSD of each member of family *i* to each member of family *j* is less than the cutoff value (2Å), then the families are merged. This builds up clusters of poses starting from the most closely related poses. The centroid of the family is then labeled as the family head.

It should be noted that our choices of DOCK for pose generation, the CNS algorithm for clustering, and DREIDING force field for energy evaluation are made for convenience in our implementation. Any other methods of pose generation, clustering, and scoring could be used.

Many of the parameters in DarwinDock can be adjusted in order to better suit a particular project. For instance, the default completeness threshold of 2% can be increased to produce a faster, less complete calculation. The clustering diversity can be increased from 2Å so that the family head is more representative of the children. The percent of families scored can be increased above 10% for a slower, but more thorough consideration of possible poses. The polar and nonpolar components of the scoring energy can be adjusted based on the composition of the ligand. Changing forcefield parameters such as the dielectric constant can also alter scoring. The DUD set provides a straightforward test for identifying the default settings for DarwinDock. Specifically, we will derive the defaults for:

- Completeness threshold (2%)
- Clustering diversity (2Å)

- Percent of families fully scored (10%)
- Polar (100%) and nonpolar (10%) scaling for scoring
- Dielectric constant (2.5, distance dependent)

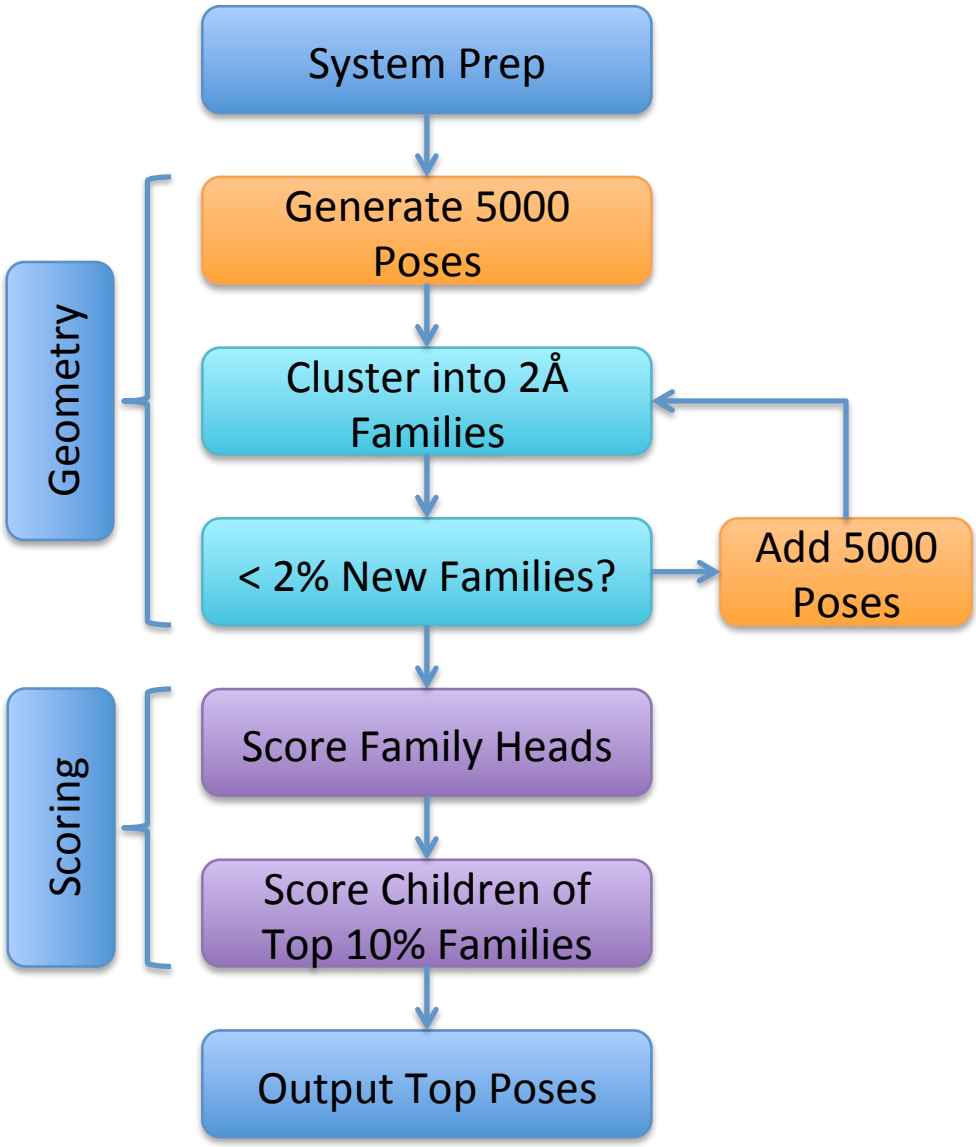

Figure 2-1 - Diagram of the DarwinDock algorithm. Pose generation and scoring are partitioned into two completely separate stages. The geometry or "completeness" stage generates 5000 poses, clusters them into 2Å families, and adds 5000 additional poses until the completeness threshold has been reached. The scoring stage initially only evaluates the 2Å family heads. 90% of the families are eliminated based on the family head energy, and the children of the remaining 10% are scored fully. Typically, 120 interesting candidate poses are output from this final list.

The DUD Set

The DUD set contains 40 systems for which accurate x-ray structures are available for a co-crystal of a small molecule ligand bound to a protein. It was intended to provide a reasonable test for docking programs and methods so that their accuracy can be assessed. Of these 40 systems we rejected three of the systems as inappropriate for small-molecule docking validation.

1. The version of the system 'thrombin' (pdb: 1ba8) included in the DUD set contains a covalent ligand, which represents a wholly different class of ligands than standard small-molecules.
2. The system 'pdgfrb' in the DUD set is in fact derived from a computational model and not an experimental structure; therefore, it cannot be used to provide accurate validation of another docking model.
3. The ligand in 'cdk2' is not completely resolved, with several missing heavy atoms.
4. Additionally, while we do not reject the 'comt' system, it should be noted that it contains two copies of the target ligand within the binding site. The presence of the second ligand is only obvious when one includes neighboring unit cells from the crystal in the structure, which may explain why this system was included in the DUD set. The positioning of the two ligands is shown in Figure 2.

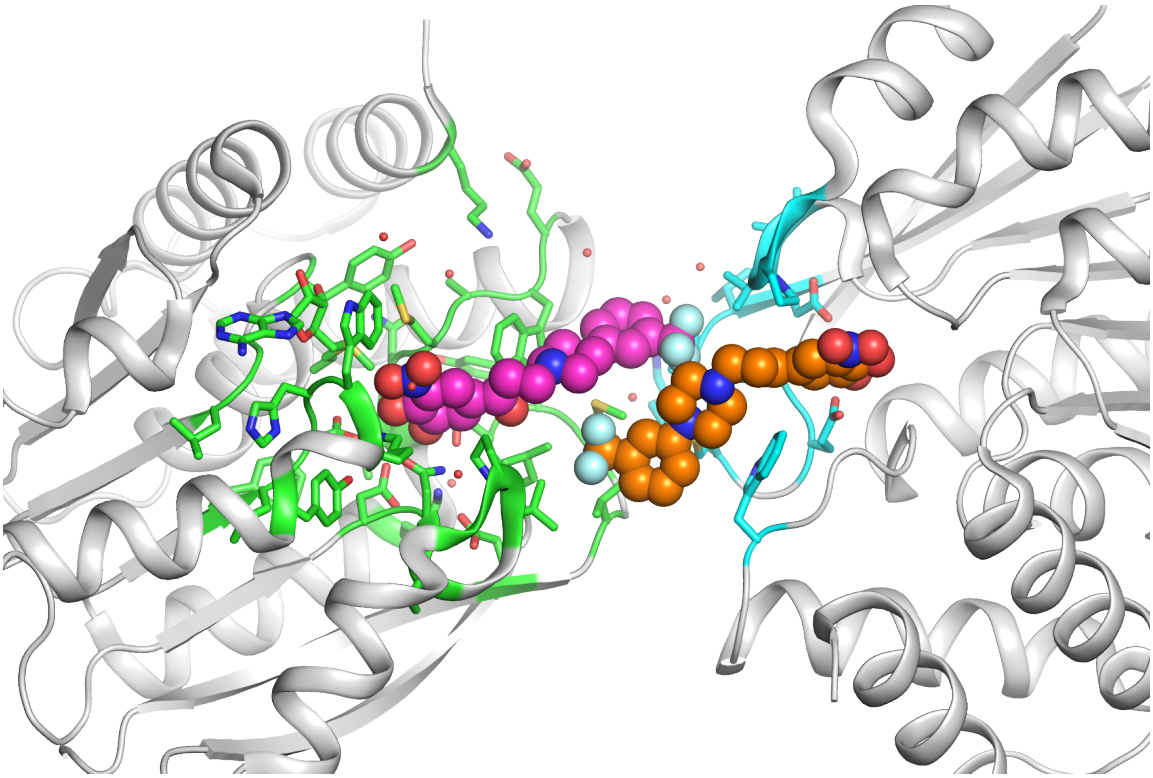

Figure 2-2 - The system 'comt' contains two copies of the target ligand within the binding site, shown as spheres in magenta and orange. This oddity is only obvious when one includes neighboring unit cells from the crystal in the structure.

Table 2-1 - DUD Systems used in validation with corresponding PDB ID number, the residue ID information of the ligand, and general system information. Three of the forty systems have been rejected as inappropriate for inclusion in benchmarking, with reasons listed.

Name	PDB ID	Lig ID	Lig Num	Lig Chn	System Name	System Class
ace	1o86	LPR	702	A	Angiotensin-converting enzyme	Metalloenzyme
ache	1eve	E20	2001	A	Acetylcholine esterase	Other enzyme
ada	1ndw	FR2	1001	A	Adenosine deaminase	Metalloenzyme
alr2	1ah3	TOL	320	A	Aldose reductase	Other enzyme
ampc	1xgj	HTC	777	A	AmpC beta lactamase	Other enzyme
ar	2ao6	R18	1001	A	Androgen receptor	Nuclear hormone receptor
comt	1h1d	BIA	335	A	Catechol O-methyltransferase	Metalloenzyme
cox1	1q4g	BFL	701	A	Cyclooxygenase 1	Other enzyme
cox2	1cx2	S58	701	A	Cyclooxygenase 2	Other enzyme
dhfr	3dfr	MTX	164	A	Dihydrofolate reductase	Folate enzyme
egfr	1m17	AQ4	999	A	Epidernam growth factor receptor kinase	Kinase
er_ag	1l2i	ETC	600	A	Estrogen receptor agonist	Nuclear hormone receptor
er_ant	3ert	OHT	600	A	Estrogen receptor antagonist	Nuclear hormone receptor
fgfr1	1agw	SU2	1001	A	Fibroblast growth factor receptor kinase	Kinase
fxa	1f0r	815	401	A	Factor Xa	Serine protease
gart	1c2t	NHS	222	A	Glycinamide ribonucleotide transformylase	Folate enzyme
gpb	1a8i	GLS	998	A	Glycogen phosphorylase beta	Other enzyme
gr	1m2z	DEX	301	A	Glucocorticoid receptor	Nuclear hormone receptor
hivpr	1hpx	KNI	900	B	HIV protease	Other enzyme
hivrt	1rt1	MKC	999	A	HIV reverse transcriptase	Other enzyme
hmga	1hw8	114	3	D	Hydroxymethylglutaryl-CoA reductase	Other enzyme
hsp90	1uy6	PU3	1224	A	Human heat shock protein 90 kinase	Kinase
inha	1p44	GEQ	350	A	Enoyl ACP reductase	Other enzyme
mr	2aa2	AS4	201	A	Mineralcorticoid receptor	Nuclear hormone receptor
na	1a4g	ZMR	466	A	Neuraminidase	Other enzyme
p38	1kv2	B96	391	A	P38 mitogen activated protein kinase	Kinase
parp	1efy	BZC	201	A	Poly(ADP-ribose) polymerase	Other enzyme
pde5	1xp0	VDN	201	A	Phosphodiesterase V	Metalloenzyme
pnp	1b8o	IMH	600	A	Purine nucleoside phosphorylase	Other enzyme
ppar	1fm9	570	200	D	Peroxisome proliferator activated receptor gamma	Nuclear hormone receptor
pr	1sr7	MOF	302	B	Progesterone receptor	Nuclear hormone receptor
rxr	1mvc	BM6	200	A	Retinoic X receptor alpha	Nuclear hormone receptor
sahh	1a7a	ADC	435	A	S-adenosyl-homocysteine hydrolase	Other enzyme
src	1y57	MPZ	600	A	Tyroside kinase SRC	Kinase
tk	1kim	THM	2	B	Thymidine kinase	Kinase
trypsin	1bju	GP6	910	A	Trypsin	Serine protease
vegfr2	1fgi	SU1	1001	A	Vascular endothelial growth factor receptor kinase	Kinase

Rejected					
thrombin	1ba8	0IT	1	B	covalent ligand
pdgfrb					computational model
cdk2	1ckp	PVB	299	A	incompletely resolved ligand

System Preparation

Despite being part of a curated set, the DUD systems required careful preparation before being used in validating our method. While DUD provides pre-prepared files for each system, we found it useful to return to the original PDB source. In particular this allowed

us to use the full protein structures and symmetry information in order to generate the final systems used for docking. The general procedure is as follows:

1. Generate neighboring unit cells using the symmetry information
2. Remove parts of the system that are distant from the target ligand (8Å cutoff)
3. Add hydrogens
4. Optimize ligand protonation states
5. Optimize asparagine, glutamine, and histidine flips, as well as histidine protonation
6. Minimize ligand separately and assign partial charges
7. Assign forcefield types
8. Perform conjugate-gradient energy minimization on the system
9. Alanize bulky residues (V, L, I, M, F, Y, W)
10. Generate sphere points for use with DOCK

Steps 1 and 2 were performed using PyMol[14]. Steps 3-5 were performed using the Maestro Protein Preparation Wizard[15-17]. CHARMM[18] charges were used for protein atoms. Ligands were minimized using the Maestro OPLS forcefield minimization[19] before generating Mulliken charges. Single-point energy calculations were performed using Jaguar[20] and the B3LYP level of DFT with the 6-311G** basis set except for the ligand containing bromine, where the ERMLER**++ basis set was used. Conjugate gradient energy minimization of the final system was performed using MPSim. Sphere generation was performed using the standard DOCK *sphgen* parameters and methods, with the exception that the maximum sphere radius is set to be 12Å instead of 4Å. This allows the

spheres to span larger binding sites without gaps or voids in the binding site. All spheres within 5Å of the ligand were selected, and were then clustered using the "cns" algorithm to reduce the number of spheres to below 400 spheres. This is necessary for the DOCK calculation to fit within available memory.

Systems prepared thus represent the final pre-docking crystal structures; however, there are other considerations that must be made. In a real-world use of DarwinDock there would be no pre-existing crystal structure containing both protein and ligand. At best there would be an apo crystal structure or a structure containing a different ligand. Thus we will not know the conformations of protein sidechains in a real-world application of any docking method. Therefore, in addition to testing DarwinDock against structures with crystal sidechains, our most important tests are for systems in which the sidechain conformations are predicted. Here we use the SCREAM[21] method to predict the sidechain conformations of the apo-protein. This allows us to test how well DarwinDock would do in a real ligand discovery project. Some sidechains were kept fixed during the predictions due to obvious strong interactions with ions or non-target ligands in the protein. These are listed in Table 2. Alanization of the bulky, nonpolar residues was applied after sidechain placement with SCREAM. It should be noted that while many of the structures showed waters present in the binding sites, we removed all waters prior to any calculations. In a real-world test the placement of waters in the binding site would not be known before docking. Coordinated ions and other ligands, however, might be known or inferred from related structures; therefore, these were left in place.

We tested several combinations of parameters in SCREAM in order to identify the best way to predict sidechains without the presence of the ligand. Specifically, we considered flat dielectric constants of 2.5, 3.33, and 5.0, as well as distance-dependent dielectric constants of 1.0, 2.5, 3.33, and 5.0.

The presence of histidines within the binding sites of some systems required special consideration. No system had more than two histidines within 4Å of the ligand, excluding histidines that were fixed due to interactions with ions. Therefore each possible combination of neutral and positively charged histidine (denoted as "B" instead of "H" in our terminology) was attempted. We also tested an additional combination where all flexible histidines in the binding site were replaced with alanine. The histidines treated in this way are listed in Table 2. Using this approach up to five different sidechain predictions were made for each dielectric constant.

Table 2-2 - Residues in the binding site that are fixed due to interactions with ions or non-target ligands, and histidines in the binding site that are tested as both neutral and charged.

System	Fixed Residues	Histidines
ace	H383_A H387_A E411_A	H353_A H513_A
ache		H440_A
ada	H15_A H17_A H214_A H238_A D295_A	H157_A
alr2		H110_A
comt	D141_A D169_A	H142_A
cox2		H90_A
dhfr		H28_A
er_ag		H524_A
er_ant		H524_A
fxa		H57_A
gart		H108_A H137_A
gpb		H377_A
hivrt		H235_A
p38		H148_A
parp		H862_A
pde5	H617_A H653_A D654_A D764_A	H613_A
pnp	S33_A H64_A R84_A H86_A S220_A	H257_A
ppar		H323_D H449_D
rxr		H435_A
sahh		H55_A H353_A
src	R388_A	
tk		H58_B
trypsin	H57_A	

Predicted Sidechain Sets

We assessed predicted sidechains for each system using the dielectric constants and histidine considerations mentioned above, both with and without alanization of the nonpolar residues. The calculations were performed in the absence of the ligand, but the

ligand was replaced in order to evaluate the number of heavy atoms with close contacts to the ligand. Based on these calculations we identified three sets of sidechains to dock to.

The first set is referred to as the "best case". We identified the best sidechain predictions for each of the 37 systems using the number of close contacts and the sidechain RMSD for each combination of dielectric constant and histidine treatment. The systems were alanized, therefore only polar sidechains were used in the bump and RMSD analysis. These predictions represent a "best case scenario" for predicted sidechains and are a reasonable set to use for identifying the optimum default settings for DarwinDock. However, this set doesn't represent a true real-world test because information about the ligand and sidechains are used to identify which prediction method to use for each system. Table 3 shows the analysis of bumps for different settings. It is obvious from the analysis of the bumps that there are clear cases where alanization of bulky residues dramatically decreases the number of bumps with the ligand.

For real-world testing of DarwinDock we used two additional sets of sidechain predictions. The first used a constant dielectric of 2.5. While not the best performer in terms of bumps, it represents the default settings that have been used in previous applications of DarwinDock. The set with the fewest average number of bumps used a distance-dependent dielectric of 2.5. Both sets were tested with and without alanization.

For reference only we also tested DarwinDock against the systems using crystal sidechains with and without alanization.

Table 2-3 - This table shows the number of bumps with the ligand for each type of sidechain prediction using different dielectrics and with or without alanization. The left half of the table shows results with alanization, the right without alanization. The table has been color-coded with large numbers of bumps shown in red. It is clear when comparing the left (alanized) and right (not alanized) portions of the table that alanization is key to reducing the number of bumps with the ligand. Even small numbers of bumps can make it impossible for the ligand to be placed correctly during docking.

	alanized									all sidechains								
	best	worst	2.50/flat	3.33/flat	5.00/flat	1.00/dist	2.50/dist	3.33/dist	5.00/dist	best	worst	2.50/flat	3.33/flat	5.00/flat	1.00/dist	2.50/dist	3.33/dist	5.00/dist
average			1.57	1.14	1.16	1.78	1.11	1.54	1.59			6.62	5.22	5.62	6.19	5.03	4.70	5.49
worst			26	10	10	14	10	22	29			35	21	35	19	20	24	29
ace	0	1	0	0	0	0	1	1	1	0	1	0	0	0	0	1	1	1
ache	0	0	0	0	0	0	0	0	0	7	12	7	8	7	8	11	12	12
ada	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
alr2	0	0	0	0	0	0	0	0	0	0	13	0	0	0	13	2	2	2
ampc	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ar	2	6	2	2	2	2	2	6	2	10	23	23	17	22	16	18	10	16
comt	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cox1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	1	1	1
cox2	0	10	2	10	8	4	0	0	0	0	10	2	10	8	4	0	0	0
dhfr	0	6	0	0	0	3	6	6	6	6	20	14	15	14	17	20	6	6
egfr	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	2	2	5
er_ag	0	0	0	0	0	0	0	0	0	8	11	9	8	8	8	11	8	8
er_ant	0	0	0	0	0	0	0	0	0	0	7	0	1	4	1	6	6	7
fgfr1	0	0	0	0	0	0	0	0	0	0	6	0	4	0	6	0	0	6
fxa	0	0	0	0	0	0	0	0	0	3	7	7	3	3	3	7	7	3
gart	0	4	1	0	2	1	1	4	0	0	4	1	0	2	1	1	4	0
gpb	2	10	3	3	10	10	10	2	2	2	13	10	6	13	10	10	2	2
gr	1	8	2	2	2	8	1	1	1	12	24	24	21	21	19	12	12	12
hivpr	0	8	0	0	0	8	1	1	5	0	9	2	2	0	9	2	1	6
hivrt	0	0	0	0	0	0	0	0	0	2	17	14	14	3	17	3	2	3
hmga	0	2	2	2	1	0	0	1	1	2	7	7	3	2	5	5	6	6
hsp90	0	0	0	0	0	0	0	0	0	2	13	6	6	6	2	2	2	13
inha	0	0	0	0	0	0	0	0	0	12	35	35	20	35	12	20	20	20
mr	0	1	0	0	1	0	0	0	0	1	9	3	3	1	9	2	2	3
na	2	29	26	6	10	2	5	22	29	3	29	27	6	11	3	7	24	29
p38	1	14	10	5	1	14	4	2	2	4	17	11	8	4	17	5	5	5
parp	0	1	0	0	0	1	0	0	0	0	12	0	2	2	12	0	0	0
pde5	0	0	0	0	0	0	0	0	0	0	15	15	7	4	4	1	2	0
pnp	0	1	1	1	1	0	1	1	0	1	4	1	4	4	3	4	3	3
ppar	0	2	0	0	0	0	0	0	2	0	7	2	0	2	2	3	7	2
pr	0	0	0	0	0	0	0	0	0	0	3	3	2	3	0	1	0	0
rxr	0	3	3	3	0	1	0	0	0	2	10	10	4	7	2	8	8	8
sahh	0	6	3	2	0	6	3	4	3	0	6	3	2	0	6	5	4	3
src	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0
tk	0	4	0	3	3	4	4	4	4	0	15	0	8	15	12	9	9	15
trypsin	0	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	0
vegfr2	0	0	0	0	0	0	0	0	0	4	5	5	5	5	5	5	4	5

"Best Case" Predicted Sidechain Results

This set of sidechain predictions represents a "best case scenario" for predicted sidechains. Such a set allowed us to identify the best default settings for the percent of families scored, the composition of the scoring energy, the clustering diversity, and the completeness threshold. For the percent of families scored we tested 10, 25, 33, 50, and 100%. As with the sidechain predictions, we assessed both a flat dielectric of 2.5 and a distance-dependent dielectric of 2.5. We also tried various scalings of the polar (Coulomb and hydrogen bond) component and the nonpolar (van der Waals) component of the scoring energy. The

scalings tested were 100% polar with 0-100% phobic energy in 10% increments, and 100% phobic with 0-100% polar energy in 10% increments. In order to assess the accuracy of the docking calculations we consider how many of the final 120 poses are within a 2.0Å heavy-atom RMSD of the crystal ligand. Through experience we have found that a pose within 2Å of the crystal ligand is sufficient to identify the pharmacophore. It represents both the orientation of the ligand in the binding site as well as the key interactions with the protein. While it is of course ideal to have as small an RMSD as possible, a 2Å RMSD is sufficient to have predictive value and acknowledges that the crystal structure is only a snapshot of what is really a dynamical system..

Figure 3 summarizes the average number of 2Å hits across all 37 DUD systems docked to, out of a possible 120 hits. There was a clear preference for the scoring energy using 100% polar (Coulomb + hydrogen bonding) and 10% phobic (van der Waals), regardless of the dielectric constant used or the percent of families scored, therefore this was set as the default for remaining calculations.

Figure 4 examines the impact of the percent of families scored and the dielectric constant when the 100% polar / 10% phobic scoring energy is used. The average number of 2Å hits was slightly higher for the distance-dependent 2.5 dielectric than for the constant 2.5 dielectric. As one would expect, increasing the percent of families scored increased the average number of hits; however, the impact was not significant. Increasing the percent of families scored from 10% to 100% yielded a little more than 1 extra hit on average while dramatically increasing the computational cost and calculation time. As 25, 33, and 50%

yielded even smaller increases, there is no reason to set the default percent families scored above 10%. Therefore, the default parameters for the remaining calculations were set at:

- Scoring energy: 100% polar, 10% phobic
- Distance-dependent dielectric, 2.5
- 10% families scored

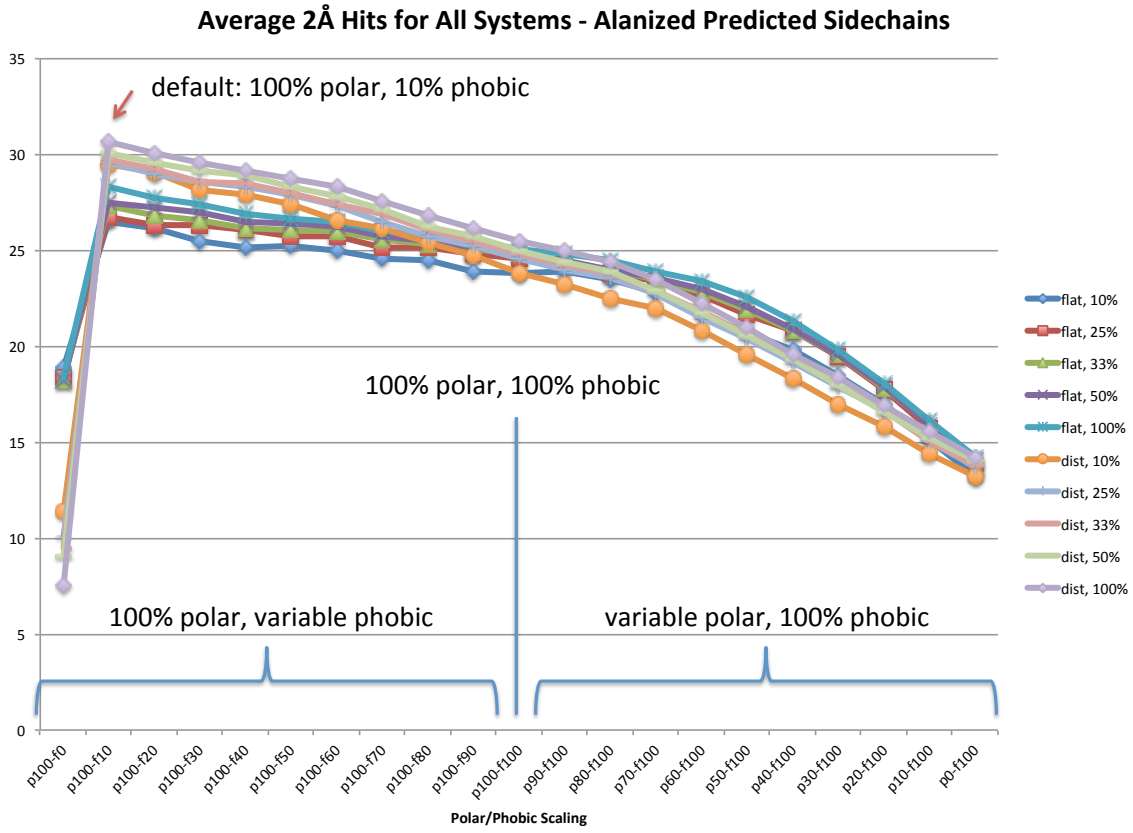

Figure 2-3 - Analysis of docking to the "best" sidechain predictions. The series show either flat or distance-dependent dielectric constants for varying percentages of families scored. The left half of the chart shows scoring energies with 100% polar and the phobic energy scaled from 0 to 90%. The right half of the chart shows scoring energies with 100% phobic energy and the polar energy scaled from 90 to 0%. The center point has 100% polar and 100% phobic scaling. There is a clear preference for 100% polar and 10% phobic regardless of dielectric and percent of families scored.

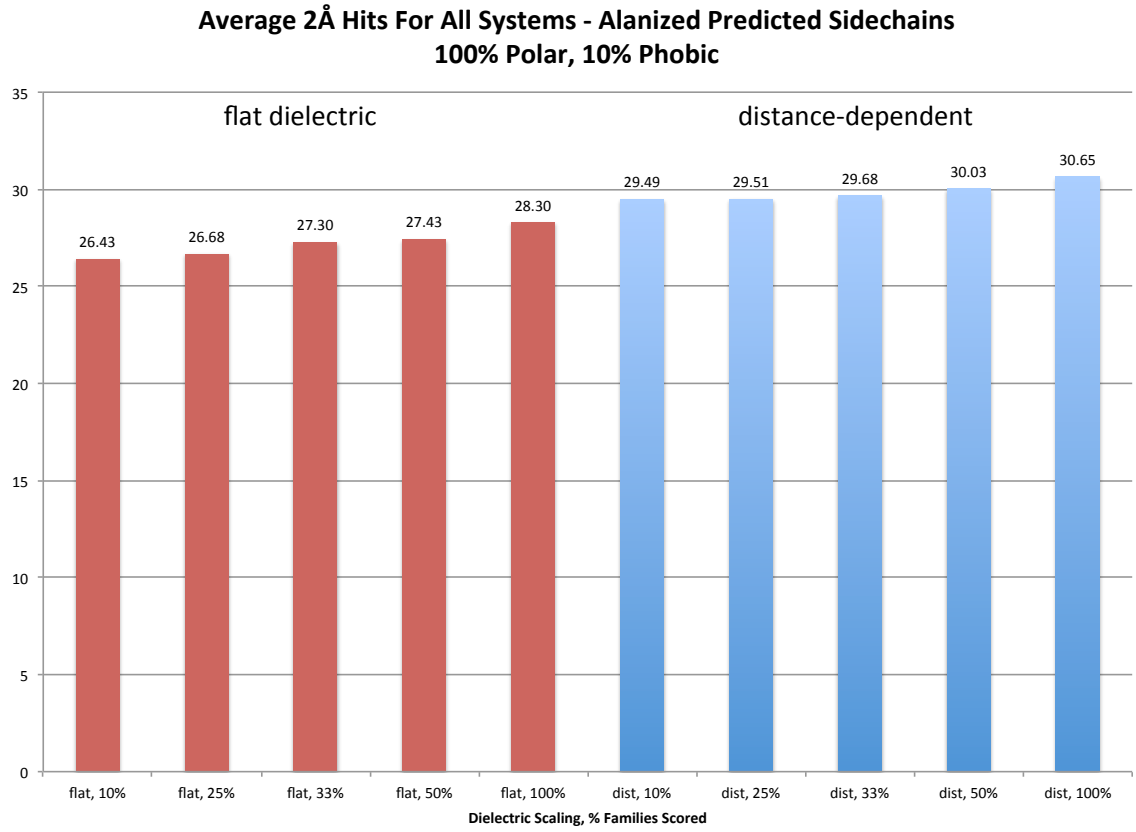

Figure 2-4 - Analysis of docking to the "best" sidechain predictions with the scoring energy set to 100% polar, 10% phobic. The red columns show a flat dielectric of 2.5 with the percent of families scored set to 10, 25, 33, 50, and 100%. The blue columns show the distance-dependent dielectric for the same percent of families scored. There is a clear, modest preference for the distance-dependent dielectric. There is also a slight, but insignificant improvement from increasing the percent of families scored. Due to the added computational cost, there is no reason to score more than 10% of the families for such insignificant improvement.

Using these parameters we assessed different possible completeness thresholds and clustering diversities. In addition to the accuracy of the calculations, these parameters can have a significant impact on the computational cost of the calculations. Using the default clustering diversity of 2Å, we tested completeness thresholds of 1, 2, 5, 10, and 25%. As expected, the accuracy of the calculations increases with the thoroughness of the calculations. That is, a higher completeness threshold represents a more complete sampling of the binding site and is more likely to produce correct poses. Conversely, a

looser completeness threshold is more likely to miss correct poses. The 1% completeness threshold produced a higher number of average hits (33.4) compared to the 2% completeness threshold (29.5), but had a much higher computational cost. The maximum number of poses used by a system increased from 60,000 to 100,000 poses and the average number of poses increased from 39,000 to 60,000. The number of systems that failed to produce any 2Å hits remained the same. Decreasing the threshold to 5% significantly reduced the computational cost, but the number of failed systems increased from 3 to 4. Therefore we feel that a completeness threshold of 2% represents a reasonable balance between accuracy and computational cost.

Using the 2% completeness threshold we assessed clustering diversities of 1, 2, and 3Å. The clustering diversity determines how closely a family head resembles the rest of the family members, and thus how accurate our assumption is that the energy of the family head can be used to eliminate the family members from further consideration. Clustering at 1Å yielded ~6.5 additional hits than the 2Å clustering and one fewer completely failed systems. However, the computational cost of clustering at 1Å was dramatically higher than that at 2Å. 1Å clustering had a maximum number of poses of 125,000 and an average of 86,000, while 2Å clustering had a maximum of 60,000 and an average of 39,000. Increasing the clustering to 3Å reduced the average number of hits by ~3 and decreased the maximum and average number of poses to 40,000 and ~27,000, respectively. However, despite the reduced number of poses, clustering at 3Å actually increased the number of energy calls. This is because the average number of children per family increased from 5.6 to 10. Our default of clustering at 2Å represents an optimum, given the other settings used.

Table 4 summarizes the results for the completeness threshold and clustering diversity tests.

Table 2-4 – Summary of results for various completeness thresholds and clustering diversities. Smaller completeness thresholds and clustering diversities yield improved results, but at significantly increased computational cost. The default settings are highlighted in green.

	Clustering Diversity: 2Å						
Threshold	Avg Hits	# Fails	% Fails	Max Poses	Avg Poses	Avg Fams	Poses/Fam
1%	33.4	3	8.1	100000	60214.3	9124.3	6.6
2%	29.5	3	8.1	60000	39082.1	6986.2	5.6
5%	25.5	4	10.8	35000	23984.4	5052.2	4.7
10%	22.8	4	10.8	20000	16416.8	3879.8	4.2
25%	18.9	4	10.8	15000	11011.4	2941.9	3.7

	Completeness Threshold: 2%						
Diversity	Avg Hits	# Fails	% Fails	Max Poses	Avg Poses	Avg Fams	Poses/Fam
1 Å	36.1	2	5.4	125000	86025.1	30701.3	2.8
2 Å	29.5	3	8.1	60000	39082.1	6986.2	5.6
3 Å	26.4	3	8.1	40000	26784.8	2675.4	10.0

% Scored	10% of Families
Scoring Energy	100% Polar / 10% Phobic
Dielectric	2.5, Distance-Dependent

Standard Predicted Sidechains

Based on the results of the "best case" predicted sidechains, we performed tests on the two sets of standard sidechain predictions. These sets used either a flat or distance-dependent dielectric of 2.5 for the sidechain predictions. The scoring energy for the docking calculations used the settings identified above: 100% polar, 10% phobic, distance-dependent dielectric of 2.5. Additionally, these sets included all of the various histidine

combinations mentioned above. These calculations represent a more authentic real-world approach to a docking problem. The "best" predicted sidechains relied on knowledge of the crystal structure in order to identify which set to use. The "standard" sidechain predictions only made use of the protein backbone from the crystal, the ligand position to identify the general binding site for sphere generation, and the crystal ligand conformation, thus making it much closer to a real-world scenario.

These docking sets produced interesting results. First, on average the systems with sidechains predicted using the distance-dependent dielectric outperformed those using the constant dielectric by about 8 hits on average, despite not having a large difference in the bindsite sidechain RMSD. The number of bumps slightly favors the distance-dependent dielectric sidechains on average. The true surprise is from the non-alanized results. The non-alanized, flat dielectric sidechains outperformed the alanized sidechains by nearly 15 hits on average. The non-alanized, distance-dependent dielectric sidechains *underperformed* the alanized sidechains by about 20 hits on average. This result is puzzling. However, the average number of hits per system is not the only important criteria. It is also important to identify the number of systems that produced no 2Å hits, which represents a complete failure of docking. Flat dielectric with alanization had the fewest systems with no hits at 4 systems, followed by the distance-dependent dielectric with alanization at 7 systems. Both non-alanized sets had 18 of 37 systems with zero 2Å hits.

As with the crystal sidechains, the loss of the van der Waals energy contribution to the binding energy and the increase in the volume to sample can clearly have a detrimental impact on some systems. However, unless one can be *supremely* confident in their sidechain predictions, alanizing the binding site is the most reliable way to ensure that some number of good poses will be produced. There are some systems where alanization is essential (e.g. "ppar"), some where it is detrimental (e.g. "pnp"), and some where it doesn't make an impact (e.g. "trypsin"). Similarly, some systems do better with the flat dielectric (e.g. "dhfr") and some do better with the distance-dependent dielectric (e.g. "rxr"). The ideal approach when working on an individual system is likely to dock to several sets of diverse sidechain predictions in order to cover multiple possibilities.

Table 2-5 – Comparison of docking results for flat-dielectric and distance-dependent dielectric predicted sidechains with and without alanization. The average number of 2Å hits per system looks encouraging for the flat-dielectric set without alanization, but *half* of the systems produce no 2Å hits.

	Avg 2Å Hits				# Sys With 0 2Å Hits			
set	flat, ala	flat, full	dist, ala	dist, full	flat, ala	flat, full	dist, ala	dist, full
33%	45.8	64.9	53.4	36.3	3	17	6	16
25%	45.8	63.5	53.2	36.4	3	17	6	17
10%	45.7	59.4	53.3	33.6	4	18	7	18

Table 2-6 – Average number of 2Å hits for flat dielectric and distance-dependent dielectric predicted sidechains, with and without alanization. There are some systems where alanization is essential (e.g. "ppar"), some where it is detrimental (e.g. "pnp"), and some where it doesn't make an impact (e.g. "trypsin"). Similarly, some systems do better with the flat dielectric (e.g. "dhfr") and some do better with the distance-dependent dielectric (e.g. "rxr"). Note: "A" refers to a histidine replaced with alanine, "B" refers to a neutral histidine changed to protonated histidine.

system	flat, ala	flat, full	dist, ala	dist, full
ace.A353_A_A513_A	31	100	42	13
ace.B353_A	74	30	15	11
ace.B353_A_B513_A	106	0	17	16
ace.B513_A	115	69	61	73
ace	106	109	36	100
ache.A440_A	1	0	0	0
ache.B440_A	0	0	0	0
ache	0	0	0	0
ada.A157_A	10	43	5	30
ada.B157_A	10	31	5	23
ada	5	50	9	11
alr2.A110_A	0	0	0	0
alr2.B110_A	7	0	0	0
alr2	7	0	0	0
ampc	1	0	0	0
ar	0	0	0	0
comt.A142_A	18	49	6	3
comt.B142_A	9	28	7	5
comt	9	28	7	5
cox1	30	120	36	120
cox2.A90_A	12	116	0	0
cox2.B90_A	2	0	9	0
cox2	3	90	2	63
dhfr.A28_A	43	0	59	22
dhfr.B28_A	51	0	4	0
dhfr	51	0	4	0
egfr	14	42	3	0
er_ag.A524_A	5	73	5	119
er_ag.B524_A	0	0	1	120
er_ag	7	77	9	119
er_ant.A524_A	2	115	10	0
er_ant.B524_A	2	96	10	0
er_ant	2	96	10	0
fgfr1	5	34	4	15
fxa.A57_A	0	0	49	0
fxa.B57_A	14	0	58	0
fxa	14	0	58	0
gart.A108_A_A137_A	0	19	0	0
gart.B108_A	0	1	0	0
gart.B108_A_B137_A	0	0	0	0
gart.B137_A	0	0	0	0
gart	0	1	0	0
gpb.A377_A	0	0	2	13
gpb.B377_A	2	0	0	0
gpb	0	0	0	0

system	flat, ala	flat, full	dist, ala	dist, full
gr	8	0	45	0
hivpr	59	0	27	6
hivrt.A235_A	17	0	27	0
hivrt.B235_A	20	0	25	0
hivrt	18	0	26	0
hmga.A752_C	1	1	1	2
hmga.B752_C	3	2	8	3
hmga	0	0	12	6
hsp90	7	3	9	0
inha	51	0	59	0
mr	14	120	33	120
na	0	0	0	0
p38.A148_A	55	0	117	0
p38.B148_A	0	0	40	0
p38	44	0	10	0
parp.A862_A	5	0	2	0
parp.B862_A	13	1	11	1
parp	12	0	4	1
pde5.A613_A	0	0	0	11
pde5.B613_A	0	0	0	0
pde5	0	0	0	0
pnp.A257_A	11	120	23	0
pnp.B257_A	2	120	15	37
pnp	16	120	37	0
ppar.A323_D_A449_D	0	0	0	0
ppar.B323_D	112	0	110	2
ppar.B323_D_B449_D	65	0	109	2
ppar.B449_D	65	0	109	2
ppar	112	0	110	2
pr	26	0	24	3
rxr.A435_A	40	0	101	0
rxr.B435_A	13	0	96	0
rxr	16	0	90	0
sahh.A55_A_A353_A	1	0	1	0
sahh.B353_A	0	21	0	0
sahh.B55_A	0	0	0	0
sahh.B55_A_B353_A	0	0	0	0
sahh	0	21	0	0
src	6	7	11	54
tk.A58_B	0	16	1	0
tk.B58_B	0	0	8	0
tk	6	118	3	0
trypsin	102	109	118	110
vegfr2	3	0	6	0

Crystal Sidechains

Tests of DarwinDock were also performed using the crystal sidechains without alanization using the same range of scoring parameters as in the "best case" tests. As it is not really possible to have crystal sidechains available in a real-world situation, these calculations are provided as a reference, not as guidance for future calculations. Compared to the "best case" predicted sidechains, the crystal sidechains showed a preference for a higher phobic scoring scaling, but surprisingly not 100%. Instead, when 100% of the families were scored, the best result was for 100% polar and 50% phobic, although 40 and 30% phobic provided nearly identical results. At 10% of families scored 100% polar and 30% phobic was the best combination. The shift toward higher phobic scoring was not surprising due to the lack of alanization. The nonpolar residues, which were removed for the predicted sidechains test, interact with ligands predominantly via van der Waals energy. What was surprising about these results is that including the full van der Waals energy was not the best. These results are summarized in Figure 5.

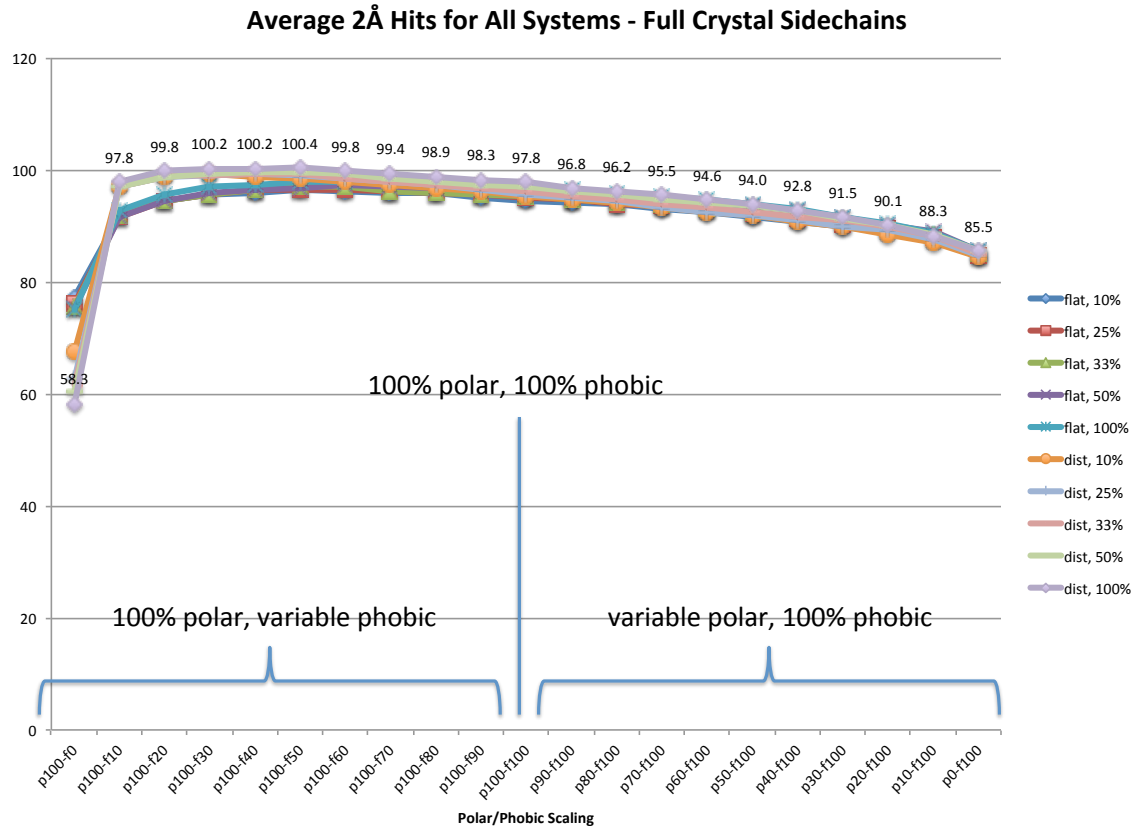

Figure 2-5 – Analysis of docking to the crystal, non-alanized sidechains. Unlike with the "best" predicted sidechains, these results show a preference for a higher phobic content, although not 100%. The best result is for distance-dependent dielectric and 100% polar, 50% phobic when scoring 100% of the families. This drops to 100% polar, 30% phobic when only scoring 10% of the families.

Docking calculations for crystal sidechains *with* alanization using 100% polar, 10% phobic and distance-dependent dielectric showed many fewer 2Å hits than the crystal, *non*-alanized results, but still more than the "best" predicted sidechains. Two factors explain the lost hits in the crystal, alanized case. First, removing the nonpolar residues obviously removed whatever van der Waals contributions those sidechains make to the binding energy. Second, alanization dramatically increased the size of the binding site for many systems. This allows ligands to make spurious interactions with polar sidechains that wouldn't be possible if they were blocked by the nonpolar residues. It also increased the

number of poses necessary for complete sampling of the binding site, meaning that the final 120 poses represented a smaller percentage of all the poses generated and scored.

Table 2-7 - Number of 2Å hits when using full crystal sidechains.
(Columns are split for compactness.)

system	xtl, full
ache	120
cox1	120
cox2	120
dhfr	120
er_ag	120
gr	120
hivrt	120
inha	120
mr	120
p38	120
pnp	120
ppar	120
pr	120
rxr	120
sahh	120
tk	120
trypsin	120
ace	119
ar	119
alr2	118
gart	118

system	xtl, full
hivpr	117
vegfr2	116
gpb	99
er_ant	98
hmga	92
pde5	92
ada	87
fxa	75
na	64
comt	54
parp	49
fgfr1	45
egfr	37
src	25
hsp90	16
ampc	9

failed	0
avg	97.0
max	120

Conclusions

The results in Table 5 and Table 6 show that DarwinDock is broadly successful when tested against the DUD set. Only two systems completely fail to produce any 2Å hits across all four sidechain predictions (flat- or distance-dependent dielectric, with or without alanization). Several other systems only have small numbers of hits. This is in contrast to

Table 7, which shows that nearly half of the systems have a full set of 120 2Å hits when the full, crystal sidechains are used. This comparison illustrates that the difficulty is not with DarwinDock, but rather the inaccuracy of sidechain predictions. Here the fundamental problem is that in the apo-protein the best sidechain conformations often invade the binding site.

The ideal starting point for a docking calculation would be a crystal structure with a related ligand already bound. In such a situation it should be possible to identify what residues are likely to move and which are not. A crystal structure without a ligand would also provide some insight. Both of those situations would relieve some of the uncertainty of the sidechain positions and yield good docking results with DarwinDock.

Of course a most interesting case is where there is no crystal structure for the protein. Indeed, most applications of DarwinDock have been for cases where the protein structure was predicted. With such *ab initio* starting structures things are more challenging. Clearly the tests discussed above show that some predicted sidechains are reliable and some are not. If one has the time to focus on a single system it may be possible to improve the odds of getting good docking results by trying multiple combinations of sidechain conformations and by using available experimental knowledge of the system.

These results show that DarwinDock is a reliable method for generating docked poses for small molecule ligands. The primary improvements necessary to the docking process are not with DarwinDock itself, but obtaining a good structure to dock to. As such,

DarwinDock is a useful tool for investigating the interactions between proteins and small molecules.

References

1. Scott, C. E., Ahn, K. H., Graf, S. T., William A Goddard, I., Kendall, D. A., & Abrol, R. (2016). Computational Prediction and Biochemical Analyses of New Inverse Agonists for the CB1 Receptor. *Journal of Chemical Information and Modeling*, *56*(1), 201–212. http://doi.org/10.1021/acs.jcim.5b00581
2. Li, Q., Kim, S.-K., Goddard, W. A., III, Chen, G., & Tan, H. (2015). Predicted Structures for Kappa Opioid G-Protein Coupled Receptor Bound to Selective Agonists. *Journal of Chemical Information and Modeling*, *55*(3), 614–627. http://doi.org/10.1021/ci500523z
3. Abrol, R., Trzaskowski, B., Goddard, W. A., III, Nesterov, A., Olave, I., & Irons, C. (2014). Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, *111*(36), 13040–13045. http://doi.org/10.1073/pnas.1413216111
4. Bray, J. K., Abrol, R., Goddard, W. A., III, Trzaskowski, B., & Scott, C. E. (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein–coupled receptor conformations. *Proceedings of the National Academy of Sciences*, *111*(1), E72–E78. http://doi.org/10.1073/pnas.1321233111
5. Kim, S.-K., & Goddard, W. A. (2014). Predicted 3D structures of olfactory receptors with details of odorant binding to OR1G1. *Journal of Computer-Aided Molecular Design*, *28*(12), 1175–1190. http://doi.org/10.1007/s10822-014-9793-4
6. Kim, S.-K., Goddard, W. A., Yi, K. Y., Lee, B. H., Lim, C. J., & Trzaskowski, B. (2014). Predicted Ligands for the Human Urotensin-II G Protein-Coupled Receptor with Some Experimental Validation. *ChemMedChem*, *9*(8), 1732–1743. http://doi.org/10.1002/cmdc.201402087
7. Tan, J., Abrol, R., Trzaskowski, B., & William A Goddard, I. (2012). 3D Structure Prediction of TAS2R38 Bitter Receptors Bound to Agonists Phenylthiocarbamide (PTC) and 6-n-Propylthiouracil (PROP). *Journal of Chemical Information and Modeling*, *52*(7), 1875–1885. http://doi.org/10.1021/ci300133a
8. Kim, S.-K., Fristrup, P., Abrol, R., & William A Goddard, I. (2011a). Structure-Based Prediction of Subtype Selectivity of Histamine H3 Receptor Selective Antagonists in Clinical Trials. *Journal of Chemical Information and Modeling*, *51*(12), 3262–3274. http://doi.org/10.1021/ci200435b
9. Kim, S.-K., Riley, L., Abrol, R., Jacobson, K. A., & Goddard, W. A. (2011b). Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins: Structure, Function, and Bioinformatics*, *79*(6), 1878–1897. http://doi.org/10.1002/prot.23012
10. Niu Huang, Brian K Shoichet, A., & Irwin, J. J. (2006). Benchmarking Sets for Molecular Docking. *Journal of Medicinal Chemistry*, *49*(23), 6789–6801. http://doi.org/10.1021/jm0608356

11. Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, *36*(15), 1132–1156. http://doi.org/10.1002/jcc.23905
12. Mayo, S. L., Olafson, B. D., & Goddard, W. A. (2002). DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, *94*(26), 8897–8909. http://doi.org/10.1021/j100389a010
13. Lim, K. T., Brunett, S., Iotov, M., & McClurg, R. B. (1997). Molecular dynamics for very large systems on massively parallel computers: the MPSim program. *Journal of Computational Chemistry*, *18*(4), 501-521
14. The PyMOL Molecular Graphics System, Version 1.3 Schrödinger, LLC.
15. Protein Preparation Wizard 2015-2; Epik version 2.4, Schrödinger, LLC, New York, NY, 2015; Impact version 5.9, Schrödinger, LLC, New York, NY, 2015; Prime version 3.2, Schrödinger, LLC, New York, NY, 2015.
16. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *Journal of Chemical Theory and Computation*, *7*(2), 525–537. http://doi.org/10.1021/ct100578z
17. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R., & Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, *27*(3), 221–234. http://doi.org/10.1007/s10822-013-9644-8
18. A D MacKerell, J., Bashford, D., Bellott, M., R L Dunbrack, J., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins †. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. http://doi.org/10.1021/jp973084f
19. MacroModel, version 10.9, Schrödinger, LLC, New York, NY, 2015.
20. Jaguar, version 8.9, Schrödinger, LLC, New York, NY, 2015.
21. Kam, V. W. T., & William A Goddard, I. (2008). Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation*, *4*(12), 2160–2169. http://doi.org/10.1021/ct800196k

Chapter 3

PREDICTING GLYCOSAMINOGLYCAN-SURFACE PROTEIN INTERACTIONS: IMPLICATIONS FOR STUDYING AXONAL GROWTH

Adam R. Griffith,[a,b,1] Claude J. Rogers,[b,c,1] Ravinder Abrol,[a,b] William A. Goddard III[a,ab,*]
Greg Miller,[b,c] and Linda C. Hsieh-Wilson,[b,c]
[a]Materials and Process Simulation Center,
California Institute of Technology, Pasadena, CA 91125
[b]Division of Chemistry and Chemical Engineering
California Institute of Technology, Pasadena, CA 91125,
[c]Howard Hughes Medical Institute,
[1]These two authors contributed equally
[*]To whom correspondence should be addressed; email wag@wag.caltech.edu

Abstract

Glycosaminoglycan (GAG)-protein interactions play important roles in the development and maintenance of the nervous system, angiogenesis, spinal cord injury, viral invasion, and immune response. Unfortunately, little structural information is available for these complexes; indeed, for such important GAGs as the highly sulfated chondroitin sulfate motifs, CS-E and CS-D, there are no structural data. This is due to the structural heterogeneity of GAGs and the difficulty of obtaining sufficient quantities of material of consistent length and sulfation pattern. Here, we describe the development and validation of the GAG-Dock computational method to accurately predict the binding poses of protein-bound GAGs. We validate that GAG-Dock accurately reproduces (< 1 Å RMSD) the crystal structure poses for four known heparin-protein structures. Further, we predict the pose of heparin and chondroitin sulfate derivatives bound to the axonal guidance proteins: protein tyrosine phosphatase σ (RPTPσ) and the Nogo receptor (NgR). Such predictions should be useful in understanding and interpreting the role of GAGs in axonal growth and other processes.

Keywords: docking | chondroitin sulfate | heparin | RPTPσ | NgR | axonal growth

Abbreviations: CS, chondroitin sulfate, GAG, glycosaminoglycan; HS, heparin sulfate; LRR, leucine-rich repeat; RMSD, root-mean-square deviation

Introduction

The glycosaminoglycans (GAGs) heparin sulfate (HS) and chondroitin sulfate (CS) are involved in a diverse array of physiological processes, such as cell proliferation, migration, differentiation, morphogenesis, angiogenesis, blood coagulation, axon guidance, and spinal cord injury through interactions with a wide variety of proteins (1-4). Despite the importance of GAG-protein interactions, there is remarkably little structural information for these complexes. This is due in part to the inherent heterogeneity of GAGs both in length and degree of sulfation, and the lack of tools required to obtain homogeneous oligosaccharides. GAGs form a family of linear polysaccharides composed of alternating uronic acid and hexosamine units. The polysaccharides can vary in length, net charge, disaccharide composition, and the pattern and degree of sulfation. The biosynthesis leads to distinct sulfation motifs for both CS and HS (Fig. 3-1). Recent studies have shown that biological activity is often dependent on the sulfation sequence, with specific, highly sulfated sequences directing the interactions of GAGs with growth factors and other signaling proteins (5-11).

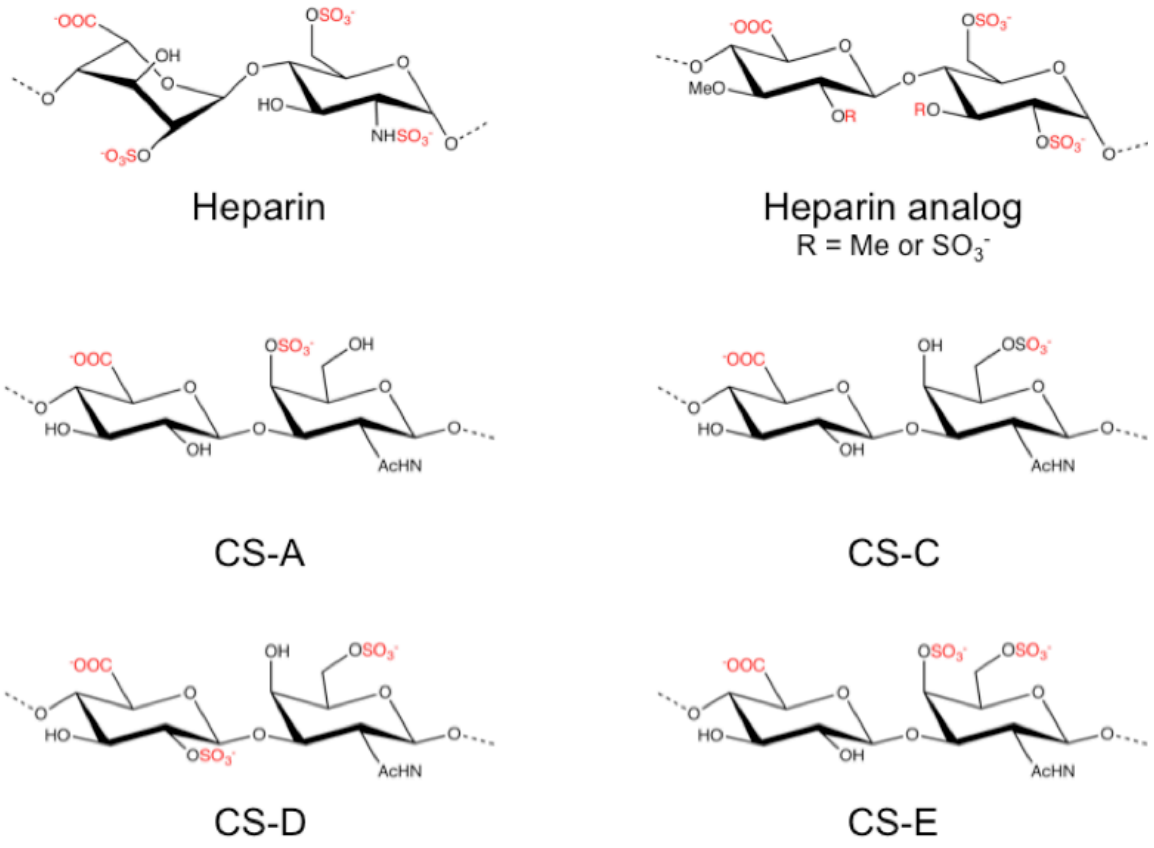

Figure 3-6 – Structures of glycosaminoglycans: heparin, heparin analog, chondroitin sulfates CS-A, CS-C, CS-D, and CS-E

Obtaining oligosaccharides with defined length and sulfation sequence is a difficult and specialized task for highly sulfated HS/heparin, and even more difficult with over-sulfated CS motifs such as CS-D and CS-E. As a result, structural data is available for only a handful of heparin-protein complexes, and no structural information is available for the CS-D and CS-E motifs. Recent work has shown that over-sulfated CS and HS interact directly with transmembrane receptors such as Nogo receptor (NgR) and type IIa receptor protein tyrosine phosphatase s (RPTPs) (11-14). However, it is unclear how GAGs engage and activate these receptors.

An alternative approach to *in vitro* structural determination is computational modeling of GAG-protein complexes. However, modeling GAG-protein interactions is extremely challenging because of the conformational flexibility of GAGs, the high charge density of GAGs and GAG-binding sites, and the weak surface complementarity of GAG-protein interactions. Despite these challenges, we (7) and others (15-18) have used molecular modeling successfully to predict the site at which GAGs engage their target proteins (7, 15-18). Some of these methods have limited accuracy in predicting the bound pose of the ligand or have limited robustness across different systems. Moreover, most of these methods have not been applied to systems other than the known heparin-protein structures.

Herein, we report the GAG-Dock method, that we developed to model accurately GAG-protein interactions, and we validate this method against known GAG-protein systems. We further apply the method to predict the protein-bound pose of various GAGs, including over-sulfated CS, to systems without known structures.

Summary of the GAG-Dock Method

Unlike small molecule ligands often docked successfully with various techniques (19-25, 44), even the truncated GAGs are large (the CS-A 4-mer has 60 heavy atoms and a net charge of -4; the CS-E 8-mer has 137 heavy atoms, a net charge of -12). Additionally, they bind to protein surfaces rather than in pockets, and engage proteins primarily through electrostatic interactions.

Our new GAG-Dock method is based on the DarwinDock and GenDock methodology (19, 20) with modifications to accommodate bulky, highly charged, surface-binding ligands

characteristic of GAGs. The GAG-binding site is generally not known; hence, it is necessary to examine systematically all possible binding regions. To do this, we complete two rounds of docking. First, we perform "coarse-level" docking to identify the best regions for further study. Second, we carry out "fine-level" docking on the best coarse regions to identify specific, strongly bound poses.

DarwinDock/GenDock

The DarwinDock/GenDock docking method applied here (19, 20) has been applied recently to predict ligand binding sites for GPCRs such as CB1 (21), GLP1-R (22), OR1G1 (23), TAS2R38 (42), AA_3R (24), and 5HT2b-R (25). Briefly it consists of four parts:

(1) System Preparation. Starting with target protein structures (usually with no hydrogen atoms), we prepare the systems as follows: (**a**) add hydrogens to various heavy atoms using standard bond distances and hydrogen binding criteria; (**b**) assign partial charges to all protein atoms based on general force field criteria and to all heteroatoms based on Mulliken charges; (**c**) optimize the protein structure using the force field to minimize the energy; (**d**) replace the 7 bulky, nonpolar residues (V/L/I/M/F/Y/W) with alanine ("alanization") to allow more complete sampling of the binding site; and (e) generate and select regions to be sampled by the ligand.

Generally the conformations of the protein side chains at the ligand binding site depend on the location and the conformation of the ligand (the pose), while the location and conformation of the ligand depends on the side chain conformations. Our solution to this "chicken-egg" problem is to alanize the bulky, nonpolar sidechains in step **d** (mentioned

above) to allow the ligand to fully sample available sites on the protein surface in the presence of the polar interactions. After selecting the best poses, the original nonpolar sidechains are replaced and reoptimized for each pose using SCREAM (45) in a process we call "dealanization". This allows a different set of protein side chains for each ligand pose.

To select poses that are close enough to the protein to interact favorably, while not too close to clash with protein atoms, we generate spheres to describe the space available for the ligand. This is done with the *sphgen* program (26), modified to work with protein surfaces. The spheres are partitioned into overlapping boxes ("sphere regions") for docking.

(2) Generation of a Complete Set of Poses. Prior to evaluating interaction energies between the ligand and protein, we want to sample the complete set of all possible poses. We do this by iteratively generating poses and then clustering them into Voronoi-like families using RMSD as the distance metric. This is done until the number of families stops changing as additional poses are added. For the cases considered here, we used an RMSD criterion of 2 Å in defining families, which generally leads to ~50,000 poses partitioned into ~2000 families, for each of which we select the "family head" as the central pose. During the pose-generation process no energies are calculated. To choose the best binding region, a quick but systematic "coarse" docking is first done using 10,000 poses without attempting the iterative, complete sampling.

(3) Scoring. To reduce computational cost, we want to minimize the number of poses for which an energy must be evaluated. Thus, scoring of the poses is broken into two steps.

First, the protein-ligand interaction energy of each family head is calculated, and the families ranked. Then, 90% of the families are eliminated based on the energy of the family head. Finally, the binding energies are calculated for all of the family members (children) in these 10% best families, and the poses are ranked with only the best 100 poses selected for further analysis. This hierarchical scoring procedure allows for a majority of the poses from the complete set (~50,000) to be eliminated without energy calculations.

(4) Optimization and Refinement. The 100 best poses from step 3 are further optimized and refined to identify the best poses. The first step is to de-alanize, i.e., replace and reoptimize the "alanized" residues with the full hydrophobic side chains. Simultaneously, all sidechains in the binding site are re-optimized (SCREAMed) using the SCREAM side-chain optimization method (45), in the presence of the specific ligand pose. Thus we end up with 100 different sets of side chain conformations, a different set for each ligand pose. Then, each of these 100 systems is energy minimized for 10 conjugate gradient steps. At this point the 100 poses are rescored and 50% eliminated. Then, another 50 steps of minimization are performed for these 50, with the poses again rescored. This final round of minimization is skipped during "coarse" docking.

GAG-Dock Modifications

The small-molecule docking methodology (DarwinDock/GenDock) was adapted to GAG structures through the following changes:

Sphere generation for flat protein surfaces requires alterations to the standard *sphgen* procedure (26). First, all spheres are generated with the 'dotlim' parameter in *sphgen* set to

-0.9, which allows spheres to be generated for flat surfaces. Second, in order to prevent the generation of deeply buried spheres that would be inaccessible to GAG ligands, a second set of spheres is generated using a probe radius of 2.8 Å instead of the normal 1.4 Å. The normal (1.4 Å probe radius) set of spheres is compared to the restricted (2.8 Å set), and only spheres within 2.8 Å of the restricted set sphere are kept. This procedure allows for spheres to be generated for the protein surface, while preventing those spheres from being so close to the surface to cause a large number of clashes with the protein during pose generation. These spheres are then partitioned into overlapping boxes/regions with 20 Å sides and 5 Å overlap.

System Preparation

All proteins studied here were prepared from PDB structures, with the exception of NgR1, NgR2, and NgR3, which required homology modeling from related systems with x-ray derived structures.

GAG Ligand Preparation

For the four validation systems, a ligand was already present in the crystal structure.

For the three systems without x-ray structures, no specific binding site is known, and hence we selected ligand structures based on the isolated ligands. Thus, the CS-A, CS-D, and CS-E structures used for docking to the non-validation systems were based on a CS-A hexasaccharide crystal structure (28), while the heparin structures for docking to the non-validation systems are based on a heparin 18-mer NMR structure. For CS-A and heparin, it

was only necessary to extend or truncate the structure to the appropriate length. We prepared CS-D and CS-E by extending the CS-A structure to a 12-mer, modifying the sulfation pattern, optimizing the sidechains, and performing Molecular Dynamics (MD) in solution. The structure closest to the average during MD was selected as the conformation for docking. This conformation was then truncated to a hexa- or octasaccharide by removing sugars from both the reducing and non-reducing ends. This step was necessary because the terminal saccharides display high variability in torsion angles during MD that are unphysical (inconsistent with possible movements) for an extended polysaccharide. Heparin and the other GAGs adopt a helical conformation that distributes charge radially along the length of the polysaccharide (31-33).

Results and Discussion

In order to validate the GAG-Dock method for such complex ligands and binding sites, we applied it to two sets of systems. The first set consists of the four validation systems for which a crystal structure including the ligand bound to the specific binding site was known. The second set of systems consists of three proteins known to bind to one or more GAG ligands, but for which the specific binding site was not known (although the general region of binding may be known). In each case, we followed the procedure of (1) coarse docking to identify the best regions, and (2) fine docking to identify the best ligand poses. In both cases the predicted binding energy was the criterion for selection.

Case 1: Validation of systems for which there are x-ray structures of the co-crystal

Five heparin-protein crystal structures have been solved, providing a means to validate our method. We applied GAG-Dock to four of these cases. We did not consider the 5th system, FGF1-FGFR2 (PDB: 1E0O (29)), because this 10-mer ligand is significantly more demanding computationally, but similar otherwise to the other validation cases. The RMSD comparisons for the predicted and crystal ligands for the validation systems are summarized in Table 3-1, showing that GAG-Dock reproduces the ligand positions with good accuracy. Figure S3-13 compares the nonbond interactions between the ligands and sidechains within the binding sites of the validation systems. As can be clearly seen from the plots in Fig. S3-13, most of the ligand-sidechain interactions were faithfully reproduced. A major source of error in the sidechain placement and interaction energies was the lack of waters in our validation systems. For structures without known binding sites, such as RPTPs and NgR, the placement of waters in an apo-protein crystal structure cannot be assumed to be correct for a ligand-bound structure, and even that information is lacking if homology modeling is used to generate the protein structure. Therefore, for a realistic assessment of the validation systems, any waters present in the crystal structures were removed. As waters often play a role in ligand binding, removing the waters allows sidechains in the protein to interact more strongly with the ligand.

FGF1

We validated our method using the crystal structure of the heparin hexasaccharide bound to two molecules of fibroblast growth factor 1 (FGF1, PDB: 2AXM (30)).

GAG-Dock correctly identified the binding site, with both molecules of FGF1 interacting with heparin at the same site, but with different specific residues interacting with the ligand for the proteins. The lowest-energy pose was within 0.70 Å root-mean-square deviation (RMSD) of the crystal structure ligand (Fig. 3-2A, S3-5, S3-6), calculated by comparing all atoms in the docked ligand to all atoms (including added hydrogen atoms) in the x-ray ligand.

Since the crystal structure is available, we docked the protein with all side chains in their experimental conformation. In this case we predict the lowest energy (strongest binding) ligand pose to have an RMSD error of 0.70 Å. Optimizing the residues for the heparin-binding site of the FGF1 molecules led to the lowest energy structure with an RMSD of 2.08 Å for the sidechains compared to the x-ray structure (Fig. S3-5). We consider that this is a success. Comparing to the x-ray pose, we find some minor differences in the energy contributions (Table S3-5). For example, K112 and K113 in chain A and K128 in chain B made stronger Coulomb and hydrogen bonding interactions with the ligand in the docked pose than in the x-ray (possibly because the water plays a role in the x-ray structure but not in ours). On the other hand, R119 was positioned farther from the ligand in the docked pose leading to weaker Coulomb interactions with the ligand. Overall the predicted energy contributions for the ligand interacting with each residue were consistent between the docked and crystal structures, indicating that these energy contributions can be used to understand the relative contributions to binding for each residue of the protein. Our conclusion is that our GAG-Dock methodology accurately predicts both the ligand pose and the relative importance of residues on the protein toward binding. Our analysis

suggests that K112, K113, K118, R122, and K128 make the most important contributions to heparin binding.

FGF2

For the complex of a heparin tetrasaccharide with FGF2 (PDB: 1BFB (34)), heparin makes contacts primarily with a single molecule of FGF2. However, in the crystal there are additional contacts with three nearby FGF2 molecules that appear to influence the conformation of the ligand. Thus, we docked the heparin tetrasaccharide to the central protein while including the three nearby FGFs to describe the conditions of the crystal structure. Again, GAG-Dock correctly predicts the binding site and the optimum pose of the crystal ligand (0.70 Å RMSD, Fig. 3-2B, S3-7).

For FGF2, the side chains of the active site differ from the x-ray structure by 2.09 Å RMSD. In particular, GAG-Dock predicts conformations of K120, R121, and K130 that lead to stronger hydrogen bond and Coulombic contributions to binding than in the x-ray structure (Table S3-6). However for FGF2, no residues had less favorable conformations in the docked structure compared to the crystal structure. Again, GAG-Dock correctly predicts the relative importance of all residues involved in binding, showing that residues K120, R121, K126, K130, and K136 contribute most strongly to heparin binding.

FGF2-FGFR1

Heparin is known to form a ternary complex with FGF and its receptor FGFR2. The crystal structure of the FGF2-FGFR1-heparin complex features a 2:2:2 stoichiometry (PDB: 1FQ9

(35)). In this structure, each heparin molecule (an 8-mer and a 6-mer) binds to the positively charged groove formed at the junction of the proteins, making contacts with an FGF2 molecule and with the D2 domains of both FGFR1 molecules. Interestingly, this structure is very similar to the FGF2-FGFR1 complex without heparin (0.37 Å RMSD (36)), suggesting that in this case GAG-Dock correctly predicts the multimeric protein-receptor-GAG complex (7). We docked both heparin molecules to regions near the FGF1 molecule and to both FGFR2 molecules. For both heparin molecules, the predicted pose correctly identifies the binding pose (with RMSD of 0.75 Å (8-mer) and 1.51 Å (6-mer); Fig. 3-2C, S3-8 – S3-11). The RMSD of the side chains in the binding site was 1.76 (8-mer) and 2.28 Å (6-mer). Again, the predicted pose accounts for the relative importance of residues involved in binding, leading to the same pharmacophore identified in the crystal structure (Table S3-7 – S3-11).

α-Antithrombin III

The interaction between heparin and α-antithrombin III (ATIII) is one of the most studied GAG-protein complexes due to its role in blood coagulation (37). The structure of ATIII bound to a heparin analog (PDB: 1E03 (38)) provided a more challenging test than the other validation cases. With no other protein species making significant contacts to the ligand, this structure lacked the constraints of the other validation systems. Even without such constraints, GAG-dock predicts the crystal structure pose with 0.60 Å RMSD (Fig. 3-2D, S3-12). The protein side chains in the binding site have an RMSD of 1.96 Å compared to the crystal structure. Again, our predicted pose accounts for the relative importance of

residues involved in binding, with residues R13 and K125 contributing more to binding in the docked pose (Table S3-12).

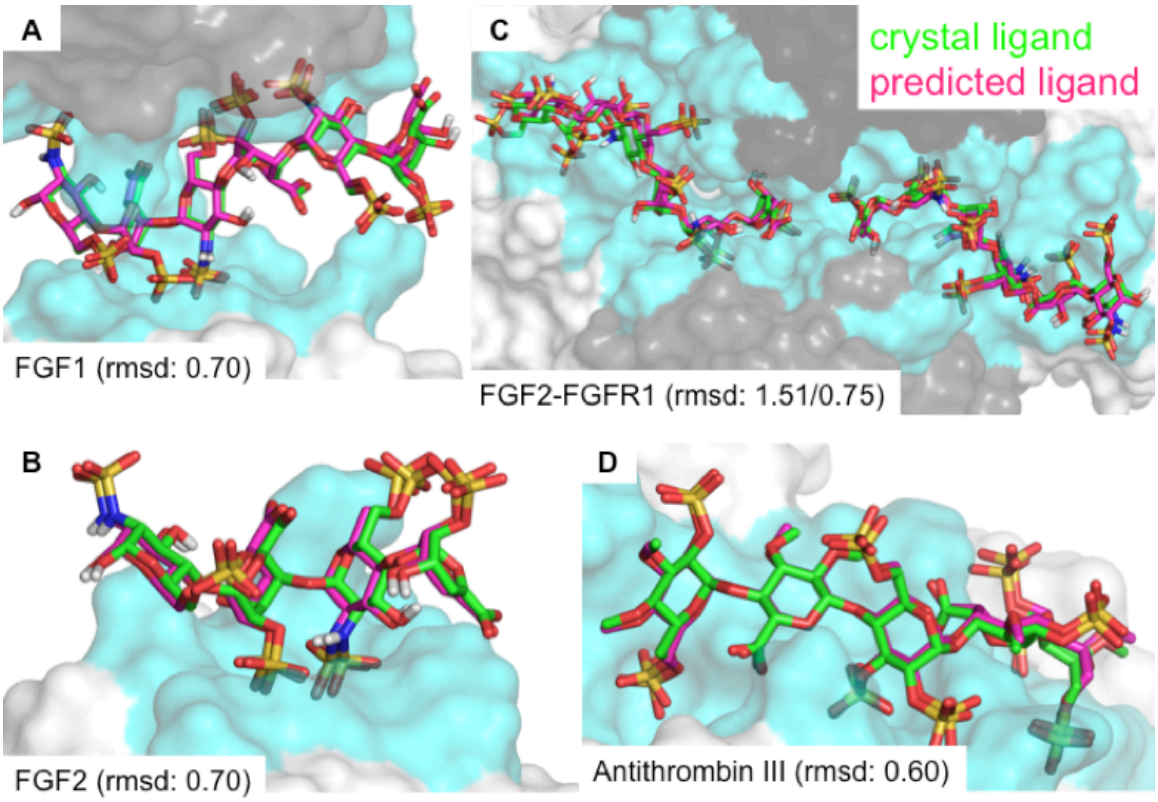

Figure 3-7 – Comparison of predicted binding sites for heparin (magenta) to the x-ray crystal ligand positions. (A) FGF1 [RMSD: 0.70Å], (B) FGF2 [RMSD: 0.70Å], (C) FGF2-FGFR1 [RMSD: 1.51Å, 0.75Å], (D) α-antithrombin III [RMSD: 0.60Å]

Table 3-7 – Summary of docking validations. The resolution of the x-ray structure is given along with the heavy-atom RMSD between the predicted and x-ray position of the ligand.

	PDB	Resolution (Å)	RMSD* (Å)
FGF1	2AXM	3.00	0.70
FGF2	1BFB	1.90	0.70
FGF2-FGFR1 A	1FQ9	3.00	1.51
FGF2-FGFR1 B			0.75
α-ATIII	1EO3	2.90	0.60

*RMSD between docked and crystal structures

Case 2: Predictions for systems for which no co-crystal structure is available

Unlike heparin, no structural information is available for over-sulfated chondroitin sulfate motifs, despite increasing evidence of their biological importance (5, 7, 11, 12, 14). This is due to the difficulty in obtaining CS molecules that are purely one type (e.g. CS-E) for use in generating crystals. The recent identification of RPTPs and NgR as mediators of CS-induced axon inhibition, and the discovery that HS and CS have opposing effects on axon morphology (13), highlight the critical need for structural data to facilitate a mechanistic understanding of GAG function. Interestingly, both RPTPs and NgR bind to polysaccharides enriched in the CS-D, CS-E, or heparin sulfated epitopes, but not the lower sulfated motifs, such as CS-A. Thus, these proteins are ideal first systems to test how consistent our docking predictions are with *in vitro* binding data. To this end, we predicted docked structures of various GAGs to RPTPs, NgR1, NgR2, and NgR3.

RPTPs

While structural data for an RPTPs-GAG complex has not been reported, the GAG binding site on the protein is well understood. A defined GAG-binding site lies on the Ig1 domain of the protein, mediated by the K67, K68, K70, K71, R96, and R99 residues (13). This region forms a shallow electropositive cavity on the surface of the protein between b strands C-D and E-F (Fig. S3-14). The quadruple mutation of K67, K68, K70, K71 to alanine has been shown to impair binding to both CS and HS (12, 39). ELISA binding data to natural GAG polysaccharides indicate that RPTPs binds strongly to the CS-D and CS-E motifs and to heparin, but *not* to CS-A (11,14). To better understand RPTPs-GAG binding, we docked CS-E, CS-D, and heparin hexasaccharides to the protein (PDB: 2YD2). We also

docked CS-A hexasaccharide but did not find significant binding, which is consistent with the lack of binding observed experimentally. The docked CS-E and heparin structures are shown in Fig. 3-3, with detailed structures shown in Fig. S3-15 (CS-E) and S3-16 (heparin).

Indeed, GAG-Dock predicts that the GAG ligands bind to the previously identified GAG-binding site. That is, CS-E and heparin both bound to K67, K68, K71, N73, Q75, R76, R96, and R99 (Table S3-14). Additionally, CS-E made contacts with K71 and S74, and heparin made contacts with T97. Motifs with lower charge density, such as CS-A, had a poor docked scoring energy with the protein compared to CS-E and heparin, suggesting a weak interaction *in vitro*. A continuous tetrasaccharide makes contacts with RPTPs in the case of CS-E, while the entire heparin hexasaccharide makes contacts. These data are consistent with a single GAG-binding site; however, CS and HS have opposing effects on axon growth in DRG neuron cultures. HS promotes axonal growth whereas CS is growth inhibitory (13). This raises the question: How is it possible for these structurally related ligands to affect such drastically different signaling outcomes?

Based on size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) using heparin fragments of various lengths and using a CS-A-enriched polysaccharide, Coles *et al.* (13) suggested that the difference between the glycans is that HS is capable of clustering RPTPs but CS is not. This is consistent with our GAG-Dock predictions (based on docked scoring energies). Later experimental studies demonstrated that CS-A has poor affinity to the protein compared to other CS sulfation motifs, especially

CS-E (11, 14). Unfortunately, CS-E oligosaccharides of suitable and defined length are not readily available to make the appropriate comparison. Nevertheless, it is plausible that CS-E polysaccharides should also be capable of simultaneously binding to multiple RPTPs molecules.

However, our docking data suggests another possibility. Because of the higher charge density and steeper helical twist of heparin/HS, our predicted pose for heparin hexasaccharide exposes several charged groups to the solvent. In contrast, the charged groups of CS-E are all engaged with the protein. Therefore, the mechanistic difference between heparin and CS-E may be that heparin is able to dimerize RPTPs, just as heparin does with FGFs, rather than merely clustering the receptor. Indeed, the SEC-MALS data in Coles *et al*. show that a decasaccharide is capable of binding two molecules of RPTPs, suggesting that bound heparin may be able to engage both proteins simultaneously (13).

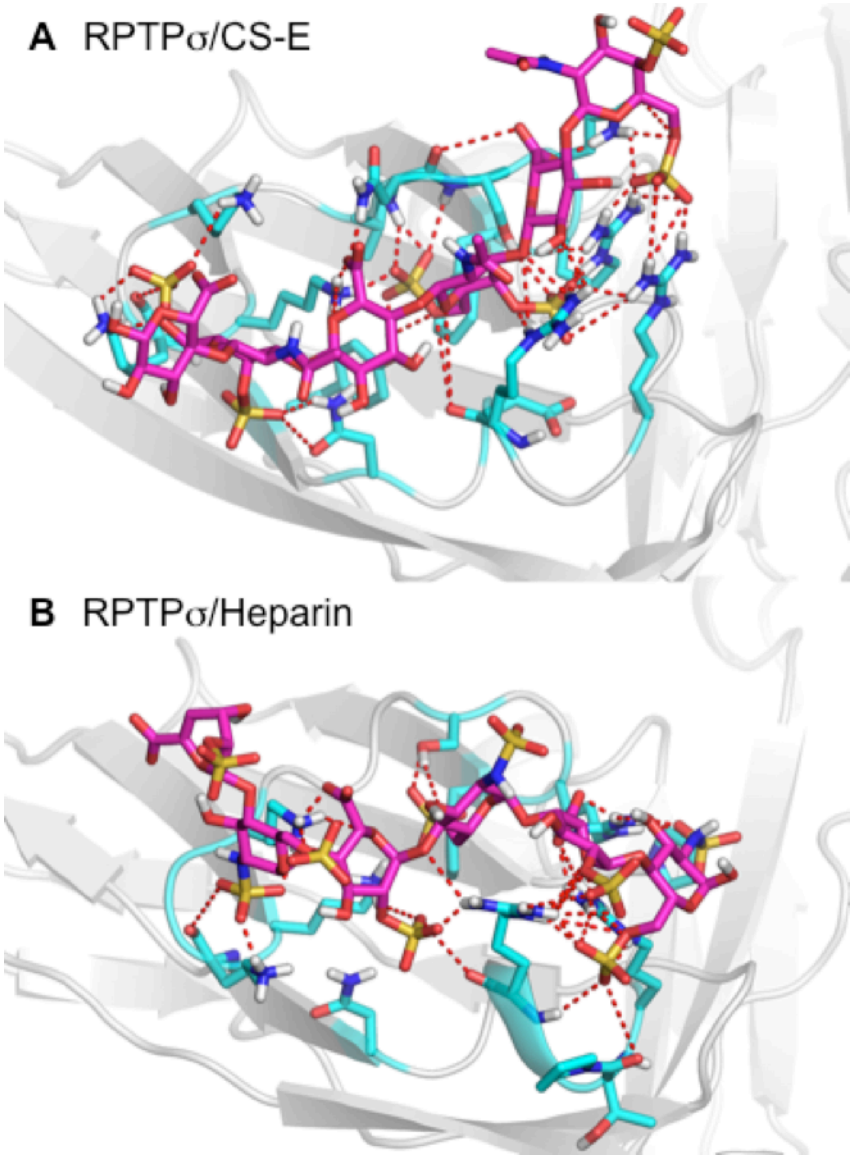

Figure 3-8 – (A) CS-E and (B) heparin bound to RPTPs. Dotted lines indicate hydrogen bonds to the protein.

NgR

The NgRs are myelin-associated inhibitors that restrict axonal growth after injury. A recent study demonstrated that NgR1 and NgR3, but not NgR2, are involved in GAG-induced axonal inhibition (14). NgRs are comprised of 8.5 leucine-rich repeat (LRR) domains

flanked by N-terminal and C-Terminal LRR capping domains and a C-terminal stalk that connects the protein to the membrane via a glycosylphosphatidylinositol (GPI) anchor (40). Compared to RPTPs, less information is known about how NgR binds to GAGs; however, domain deletion studies suggest that the C-terminal capping domain and stalk are required for CS binding (14). C-terminal regions of NgR, such as the stalk, have not been resolved in the reported crystal structures of the protein (41, 42).

To better understand the role of the C-terminal domains, we generated homology models of NgR isoforms using the ROSETTA software (43). We carried out 5 ns of MD in the presence of explicit water and counter ions to allow the 5 models per isoform to relax. We then selected the structure nearest to the average conformation for each model, minimized it, and then selected the lowest energy structure for each isoform to use in further studies. The electrostatic potential surfaces of these homology models of the extracellular domain of NgR isoforms 1–3 suggest an electrostatic basis for the difference in activity between NgR2 and NgRs 1 and 3 (Fig. S3-17). Unlike the GAG-binding isoforms, NgR2 lacks significant regions of electropositive potential. Our predicted binding energies from coarse-level docking with a CS-E tetrasaccharide to NgR2 led to much weaker interactions (–297.67 kcal/mol), relative to NgR1 and 3 (–641.27 and –985.46 kcal/mol, respectively), consistent with experimental findings.

Based on fine-level docking with CS-A, -D, -E, and heparin hexasaccharides, followed by 5 ns MD relaxation in a full water box with counterions, we predict that GAGs bind to regions of electropositive potential on the C-terminal cap of both NgR1 and NgR3 (Fig. 3-

4, S3-18, S3-23). GAG-Dock studies predict that the GAG-binding domains of NgR1 and NgR3 are on different faces of the C-terminal cap, although this could be due to the structural flexibility of this region of the protein and to discrepancies between the model and the natural state of the protein. We predict that the GAGs make polar or electrostatic contacts with residues R399, R414, R415, R416, R421, K422, R424, R426, and R430 on NgR1 (Table S3-15) and with residues R346, R350, K354, N355, N358, R360, K364, K399, R400, K401, K403 and R406 on NgR3 (Table S3-16). Many of these residues, particularly residues 414–426 on NgR1 and 399–406 on NgR3, were shown by mutagenesis studies to be important for GAG binding (14). Together, these results validate that GAG-Dock can be used both to understand the structural basis for extreme differences in GAG-binding activity between related proteins and to identify reliably the pharmacophore even in cases where the protein structure is ill defined. Detailed structures for CS-A, CS-D, CS-E, and heparin bound to NgR1 are shown in Fig. S3-19 – S3-22. Detailed structures for those ligand bound to NgR3 are shown in Fig. S3-24 – S3-27.

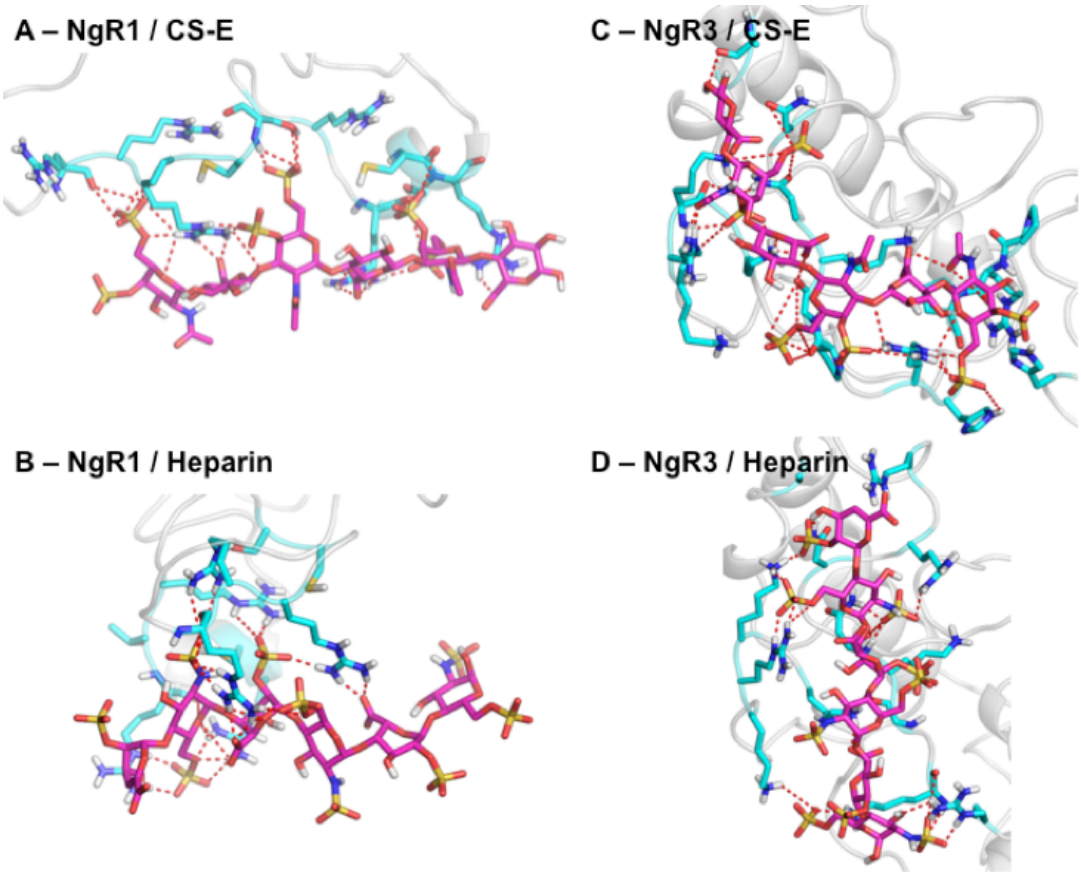

Figure 3-9 – (A) CS-E and (B) heparin bound to NgR1. (C) CS-E and (D) heparin bound to NgR3.

Suggested Post-Prediction Validations

Experimental validation of our novel RPTPs and NgR binding sites may be possible via mutation experiments. We carried out *in silico* mutations for our predicted CS-A, CS-D, CS-E, and heparin binding sites for RPTPs, NgR1, and NgR3 in order to identify reasonable suggestions for experimental validation. Rather than the more common alanine mutations, we employed a more subtle mutation to asparagine or glutamine. As noted above, the key interactions between the GAG ligands and the proteins involve arginines and lysines. Mutation of these residues to alanine represents a drastic change in character

and could result in significant disruption of the system beyond affecting binding. Mutation to asparagine or glutamine allows the possibility of maintaining some polar contact with the ligand, but without the benefit of strong charged interactions. Additionally, while the standard method is to identify mutations that decrease binding, we consider this approach to be ambiguous, as binding can be lost for many reasons. Therefore, in addition to the standard loss-of-binding mutations, we identified mutations that could potentially increase binding of the ligand to the protein.

We first employed single-residue mutations of all residues within the binding sites to either asparagine or glutamine while simultaneously optimizing the remaining sidechain conformations in the binding site using SCREAM, followed by 50 steps of conjugate gradient energy minimization. From these calculations we identified mutations that either resulted in additional or lost hydrogen bonding to the ligands. Based on these individual mutations, sets of mutations to either increase or decrease ligand binding were identified for each ligand/protein combination. It should be noted that some mutations of arginine or lysine may result in increased hydrogen bonding if the arginine or lysine was initially too constrained to make a hydrogen bond to the ligand. However, such a mutation still remains a net loss of overall binding energy due to the lost Coulomb interactions. Therefore, we only considered mutations of arginine or lysine to asparagine or glutamine for our loss-of-binding mutation sets.

For RPTPs we identified three sets of mutations that increased binding to CS-A, CS-D, or CS-E, but interestingly not to heparin. It is possible that RPTPs is already optimized for

heparin binding, but not for CS binding. Mutation set "G1" is specific for CS-E, while "G2" is specific for CS-D, and "G3" is nonspecific with the exception of decreasing heparin binding. Based on the single residue mutations we generated four sets of mutations to decrease binding. As all of these mutations affect the key arginine and lysine residues, they all unsurprisingly result in significant reductions in binding energy. The mutation sets for RPTPs are summarized in Table 3-2, and the single mutations are summarized in Table S3-17.

Table 3-8 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to RPTPs. Note that none of the sets show improved binding for heparin. Changes in binding energy are shown relative to the wildtype structures in both absolute change (kcal/mol) and in terms of percent change.

PTPS - Increased Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	-1.7	-0.3	-34.2	30.5	V73N S75N Q76N F78Q
G2	-6.2	-35.0	12.6	29.7	N74Q S75N
G3	-17.8	-23.8	-29.0	21.2	V73Q S75N Q76N F78N N103Q

PTPS - Loss of Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	565.5	719.2	709.8	1059.6	K68Q K69N K71Q R77N R97N R100N
L2	386.0	528.2	481.5	705.1	K71N R77Q R97N R100Q
L3	383.5	472.6	458.7	657.3	K68N K69Q R77N R100N
L4	468.6	629.7	603.3	891.1	K68N K71Q R77N R97N R100Q

PTPS - Increased Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	0.3	0.0	4.2	-2.9	V73N S75N Q76N F78Q
G2	1.1	4.5	-1.6	-2.8	N74Q S75N
G3	3.1	3.1	3.6	-2.0	V73Q S75N Q76N F78N N103Q

PTPS - Loss of Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	-97.3	-92.8	-87.8	-101.0	K68Q K69N K71Q R77N R97N R100N
L2	-66.4	-68.1	-59.6	-67.2	K71N R77Q R97N R100Q
L3	-66.0	-61.0	-56.7	-62.7	K68N K69Q R77N R100N
L4	-80.6	-81.2	-74.6	-84.9	K68N K71Q R77N R97N R100Q

For NgR1 we identified four sets of mutations to increase binding by building up from the single mutation information. Surprisingly one of the mutation sets ("G3") did not show any improvement in binding when tested. Set "G1" improved CS-A and CS-D binding, but not CS-E or heparin. Set "G2" improved CS-D and heparin binding. Set "G4" improved binding for every ligand except CS-E. It is again interesting that none of the mutation sets improved CS-E binding. As with RPTPs, the four loss-of-binding mutation sets were all effective in reducing ligand binding, but were non-specific for any ligand. The mutation

sets for NgR1 are summarized in Table 3-3, and single mutations are summarized in Table S3-18.

Table 3-9 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to NgR1. Note that none of the sets show improved binding for CS-E. Changes in binding energy are shown relative to the wildtype structures in both absolute change in binding energy (kcal/mol) and in terms of percent change.

NGR1 - Increased Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	-15.8	-14.0	2.7	-3.6	C395Q C405N
G2	10.6	-30.8	4.2	-38.0	S396N N399Q C405N
G3	-0.1	0.4	10.5	-1.1	C395Q S396N S403Q C405Q
G4	-20.9	-19.6	9.8	-27.1	S396Q C405N

NGR1 - Decreased Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	481.3	607.0	649.0	850.0	R390N R391N R392Q R402N R406N
L2	471.5	627.0	597.9	813.1	R390Q R391Q R400N R402N R406N
L3	382.4	463.0	580.7	610.0	R391N R392Q R402Q R406Q
L4	473.2	629.9	642.9	834.1	R390N R392Q R400N R402Q R406N

NGR1 - Increased Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	2.2	1.4	-0.3	0.3	C395Q C405N
G2	-1.5	3.1	-0.5	3.4	S396N N399Q C405N
G3	0.0	0.0	-1.2	0.1	C395Q S396N S403Q C405Q
G4	2.9	2.0	-1.1	2.4	S396Q C405N

NGR3 - Decreased Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	-65.3	-62.0	-71.3	-75.8	R390N R391N R392Q R402N R406N
L2	-64.0	-64.1	-65.7	-72.5	R390Q R391Q R400N R402N R406N
L3	-51.9	-47.3	-63.8	-54.4	R391N R392Q R402Q R406Q
L4	-64.2	-64.3	-70.6	-74.4	R390N R392Q R400N R402Q R406N

Each of the mutation sets to increase binding for NgR3 show improved binding to at least one of the ligands. None of the ligands is completely missed as heparin was for RPTPs or CS-E for NgR1. The mutation sets for NgR3 are summarized in Table 3-4, and single mutations are summarized in Table S3-19.

Table 3-10 – Predicted sets of mutations to either increase (left) or decrease (right) binding of ligands to NgR3. Changes in binding energy are shown relative to the wildtype structure in both absolute change (kcal/mol) and in terms of percent change.

NGR3 - Increased Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	-15.9	6.9	-9.9	-10.1	I345Q A348N
G2	-3.1	0.8	-10.2	-0.1	I345Q
G3	-12.9	-5.1	-5.8	-12.2	A348N
G4	-7.7	1.9	7.9	-23.7	N338Q A348N

NGR3 - Decreased Binding (Relative Energy)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	230.6	271.2	327.7	396.5	K331N K334Q R342N
L2	321.9	537.6	545.0	802.2	K331N R342N R380N R381N R383N
L3	470.9	771.7	858.4	1054.7	R330N K331N K334N R342N R380Q K381N K383N
L4	238.0	413.9	425.4	644.0	R340N R379N R380N R383N

NGR3 - Increased Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
G1	2.2	-0.6	0.8	0.7	I345Q A348N
G2	0.4	-0.1	0.8	0.0	I345Q
G3	1.8	0.4	0.5	0.8	A348N
G4	1.0	-0.2	-0.6	1.5	N338Q A348N

NGR3 - Decreased Binding (Percent Change)					
Set	CSA	CSD	CSE	HEP	Mutations
L1	-31.3	-23.6	-27.0	-25.9	K331N K334Q R342N
L2	-43.7	-46.8	-44.9	-52.4	K331N R342N R380N R381N R383N
L3	-63.9	-67.1	-70.7	-68.9	R330N K331N K334N R342N R380Q K381N K383N
L4	-32.3	-36.0	-35.0	-42.1	R340N R379N R380N R383N

Conclusions

Predicting the binding sites of highly charged GAG ligands with multiple independent charge sites and numerous possible conformations seems a formidable challenge. The very large number of charged sites on the ligands and in the binding site likely leads to redistributions of the water and ions in the solvent making polarization likely of great importance. Nevertheless, we show for eight independent systems that the simple GAG-Dock modifications of the DarwinDock general docking approach accounts well for the enormous importance of electrostatic interactions, leading to plausible structures and relative binding energies that help distinguish the strength of binding for various GAG ligands to a wide variety of receptors likely to play essential roles in axonal growth. Given the difficulty of obtaining high quality co-crystals for x-ray studies, this simple GAG-Dock computational methodology may provide the best means for predicting the structure sufficiently accurately to help design experimental probes to elucidate the issues controlling axonal growth, perhaps suggesting modified ligands that might be more selective and controllable.

Acknowledgements

ARG, RA, and WAG were supported from NIH (R01-NS071112, R01-NS073115, and R01-AI040567) and other funds donated to the Materials and Process Simulation Center. The computers used in this research were funded by grants (to WAG) from DURIP (Defense University Research Instrument Program) and from NSF (equipment part of the NSF-MRSEC-CSEM). CJR, GM, and LCH-W were supported by the National Institutes of Health (grant R01 GM084724 to L.C.H-W).

References

1. Capila I, Linhard RJ (2002) Heparin-protein interactions. *Angew Chem Int Ed* 41:390–412.
2. Raman R, Sasisekharan V, Sasisekharan R (2005) Structural insights into biological roles of protein-glycosaminoglycan interactions. *Chem Biol* 12:267–277.
3. Sugahara K et al. (2003) Recent advances in the structural biology of chondroitin sulfate and dermatan sulfate. *Curr Opin Struct Biol* 13:612–620.
4. Gama CI, Hsieh-Wilson LC (2005) Chemical approaches to deciphering the glycosaminoglycan code. *Curr Opin Chem Biol* 9:609–619.
5. Gama CI et al. (2006) Sulfation patterns of glycosaminoglycans encode molecular recognition and activity. *Nat Chem Biol* 2:467–473.
6. Shipp EL, Hsieh-Wilson LC (2007) Profiling the sulfation specificities of glycosaminoglycan interactions with growth factors and chemotactic proteins using microarrays. *Chem Biol* 14:195–208.
7. Rogers CJ et al. (2011) Elucidating glycosaminoglycan-protein-protein interactions using carbohydrate microarray and computational approaches. *Proc Natl Acad Sci USA* 108:9747–9752.
8. Hricovíni M et al. (2002) Active conformations of glycosaminoglycans. NMR determination of the conformation of heparin sequences complexed with antithrombin and fibroblast growth factors in solution. *Semin Thromb Hemost* 28:325–334.
9. Rusnati M et al. (1997) Interaction of HIV-1 Tat protein with heparin. Role of the backbone structure, sulfation, and size. *J Biol Chem* 272:11313–11320.
10. Kuschert GS et al. (1999) Glycosaminoglycans interact selectively with chemokines and modulate receptor binding and cellular responses. *Biochemistry* 38:12959–12968.

11. Brown JM et al. (2012) A sulfated carbohydrate epitope inhibits axon regeneration after injury. *Proc Natl Acad Sci USA*.

12. Shen Y et al. (2009) PTPσ is a receptor for chondroitin sulfate proteoglycan, an inhibitor of neural regeneration. *Science* 326:592–596.

13. Coles CH et al. (2011) Proteoglycan-specific molecular switch for RPTPσ clustering and neuronal extension. *Science* 332:484–488.

14. Dickendesher TL et al. (2012) NgR1 and NgR3 are receptors for chondroitin sulfate proteoglycans. *Nat Neurosci* 15:703–712.

15. Bitomsky W, Wade RC (1999) 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.

16. Bytheway I, Cochran S (2004) Validation of molecular docking calculations involving FGF-1 and FGF-2. *J Med Chem* 47:1683–1693.

17. Forster M, Mulloy B (2006) Computational approaches to the identification of heparin-binding sites on the surfaces of proteins. *Biochem Soc Trans* 34:431–434.

18. Takaoka T, Mori K, Okimoto N, Neya S, Hoshino T (2007) Prediction of the Structure of Complexes Comprised of Proteins and Glycosaminoglycans Using Docking Simulation and Cluster Analysis. *J Chem Theory Comput* 3:2347–2356.

19. Cho AE et al. (2005) The MPSim-Dock hierarchical docking algorithm: application to the eight trypsin inhibitor cocrystals. *J Comput Chem* 26:48–71.

20. Floriano WB, Vaidehi N, Zamanakos G, Goddard WA (2004) HierVLS hierarchical docking protocol for virtual ligand screening of large-molecule databases. *J Med Chem* 47:56–71.

21. Scott CE, Abrol R, Ahn KH, Kendall DA, Goddard WA (2013) Molecular basis for dramatic changes in cannabinoid CB1 G protein-coupled receptor activation upon single and double point mutations. *Protein Sci* 22:101–113.

22. Kirkpatrick A, Heo J, Abrol R, Goddard WA (2012) Predicted structure of agonist-bound glucagon-like peptide 1 receptor, a class B G protein-coupled receptor. *Proc Natl Acad Sci USA* 109:19988–19993.

23. Charlier L et al. (2012) How broadly tuned olfactory receptors equally recognize their agonists. Human OR1G1 as a test case. *Cell Mol Life Sci* 69:4205–4213.

24. Kim S-K, Riley L, Abrol R, Jacobson KA, Goddard WA (2011) Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins* 79:1878–1897.

25. Kim S-K, Li Y, Abrol R, Heo J, Goddard WA (2011) Predicted structures and dynamics for agonists and antagonists bound to serotonin 5-HT2B and 5-HT2C receptors. *J Chem Inf Model* 51:420–433.

26. Moustakas DT et al. (2006) Development and validation of a modular, extensible

docking program: DOCK 5. *J Comput Aided Mol Des* 20:601–619.

27. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci USA* 98:10037–10041.

28. Winter WT, Arnott S, Isaac DH (1978) Chondroitin 4-sulfate: The structure of a sulfated glycosaminoglycan. *J Mol Biol* 125:1–19.

29. Pellegrini L, Burke DF, Delft von F, Mulloy B, Blundell TL (2000) Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature* 407:1029–1034.

30. DiGabriele AD et al. (1998) Structure of a heparin-linked biologically active dimer of fibroblast growth factor : Abstract : Nature. *Nature* 393:812–817.

31. Tanaka K (1978) Physicochemical properties of chondroitin sulfate. I. Ion binding and secondary structure. *J Biochem* 83:647–653.

32. W D Comper OZ (1990) Hydrodynamic properties of connective-tissue polysaccharides. *Biochemical Journal* 269:561.

33. Khan S, Gor J, Mulloy B, Perkins SJ (2010) Semi-rigid solution structures of heparin by constrained X-ray scattering modelling: new insight into heparin-protein complexes. *J Mol Biol* 395:504–521.

34. Faham S, Hileman RE, Fromm JR, Linhardt RJ, Rees DC (1996) Heparin structure and interactions with basic fibroblast growth factor. *Science* 271:1116–1120.

35. Schlessinger J et al. (2000) Crystal structure of a ternary FGF-FGFR-heparin complex reveals a dual role for heparin in FGFR binding and dimerization. *Mol Cell* 6:743–750.

36. Plotnikov AN, Schlessinger J, Hubbard SR, Mohammadi M (1999) Structural basis for FGF receptor dimerization and activation. *Cell* 98:641–650.

37. Bourin MC, Lindahl U (1993) Glycosaminoglycans and the regulation of blood coagulation. *Biochem J* 289 (Pt 2):313–330.

38. McCoy AJ, Pei XY, Skinner R, Abrahams JP, Carrell RW (2003) Structure of beta-antithrombin and the effect of glycosylation on antithrombin's heparin affinity and activity. *J Mol Biol* 326:823–833.

39. Aricescu AR, McKinnell IW, Halfter W, Stoker AW (2002) Heparan sulfate proteoglycans are ligands for receptor protein tyrosine phosphatase σ. *Mol Cell Biol* 22:1881–1892.

40. Fournier AE, GrandPre T, Strittmatter SM (2001) Identification of a receptor mediating Nogo-66 inhibition of axonal regeneration. *Nature* 409:341–346.

41. Barton WA et al. (2003) Structure and axon outgrowth inhibitor binding of the Nogo-66 receptor and related proteins. *EMBO J* 22:3291–3302.

42. He XL et al. (2003) Structure of the Nogo Receptor EctodomainA Recognition

Module Implicated in Myelin Inhibition. *Neuron* 38:177–185.

43. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res* 32:W526–31.

44. Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, Rizzo RC, Case DA, James TL, Kuntz ID. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA. 2009 Jun;15(6):1219-30.

45. Kam, Victor Wai Tak (2008) *Methods in computational protein design.* Dissertation (Ph.D.), California Institute of Technology.

46. Mayo S.L., Olafson B.D., Goddard WA (1990) A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* 94, 8897.

47. Lim K-T, Brunett S, Iotov M, McClurg RB, Vaidehi N, Dasgupta S, Taylor S, Goddard WA (1997) Molecular Dynamics for Very Large Systems on Massively Parallel Computers: The MPSim Program. *J. Comp. Chem.* 18, 501.

Jaguar, version 8.9, Schrödinger, LLC, New York, NY, 2015.

Supplementary Information

System Preparation

Crystal structures & PTPσ. Protein molecules within 5 Å of the GAG ligand, if present, were selected from the PDB. Hydrogen atoms were added with tleap (S23) and CHARMM (S1) charges were assigned to each atom. The system was minimized using the DREIDING force field (S2).

NgR. Five homology models for full-length NgR1, 2, and 3 were obtained using ROSETTA (S3). Each model was minimized (5000 steps) and allowed to relax in the presence of water and counterions with 5 ns of MD. MD was performed using NAMD in four steps, as described later in the MD section. Briefly, first, a water box bounding the protein was minimized with the protein kept fixed. Second, 0.5 ns of NPT MD were performed on the water box. Third, the entire systems were minimized. Finally, 5 ns of

NPT MD were performed on the entire system. For each initial conformation, the conformation closest to the average structure from the 5 ns MD was minimized and the lowest-energy conformation was selected for each isoform.

DarwinDock

The concept behind DarwinDock (S4-S12) is (1) to generate a complete set of poses for the binding site while minimizing the number of energy evaluations, (2) then to collect these into a smaller sets containing all poses likely to be important, (3) then to evaluate the binding energy of this relatively small set to find the best poses, while ensuring that no poses are missed that might prove to be important.

Pose generation is accomplished by iteratively generating poses (but no energies) using DOCK 6 (S13) and clustering them into families using our Closest-Neighbor Seeded clustering algorithm (described below). Our usage of DOCK 6 is very simplistic, utilizing only the bump filter. We follow the default settings for generating the bump grid for DOCK 6 and set the bump cutoff to 5. Two calls to DOCK 6 are generally made. First, a request for 40,000 poses is made to determine the approximate percentage of poses that will pass the DOCK bump filter. Then a second request for poses is made, based on the percent of poses passing the bump test so that enough poses are returned to be sufficient for the iterative completeness cycle. Initially the first 5000 poses are clustered with a 2 Å diversity, then the next 5000 poses are added and reclustered, leading an increased number of families. This process is continued in increments of 5000 poses until the number of new families represents less than 5% of the total number of families at that point.

Due to the computational difficulty of dealing with GAG ligands – which are considerably larger than normal small-molecule ligands for which DarwinDock was developed – leading to correspondingly increased search volumes, we restrict the number of poses during this iterative completeness cycle to 50,000. Furthermore, it is generally not possible to request more than 15 million poses (sometimes fewer) from DOCK6 before memory limitations intercede. As a result, most regions reach the 50,000 pose limit before reaching the 5% new families threshold. Other regions may fall well short of 50,000 poses due to their geometry and memory limitations.

After generating a complete set of poses, or the largest set within our computational limits, we score all family heads (generally ~2000). For each family the central pose (based on the RMSD) is denoted as the family head. The protein-ligand interaction energy of each family head is evaluated using the DREIDING forcefield (S2) with MPSim (S14). DREIDING partitions non-bond energies into Coulomb, hydrogen-bond, and Van Der Waals (VDW). For GAG-Dock the interaction energy is the sum of all ligand-protein Coulomb and hydrogen-bond energies plus 10% of the (VDW) energy. Including only 10% of the VDW energy allows for strong polar interactions with the protein with moderate clashes that can be resolved during sidechain optimization. Not including the VDW energy results in poses with severe, unresolvable clashes with the protein, while including the full VDW energy results in poses that are too far from the protein and make poor contact.

After evaluating the interaction energy for the family heads, we eliminate the worst 90% of the families. Next, we evaluate the interaction energy for all children in the remaining 10% of families. From these children we select best 100 based on binding energy. Eliminating

90% of the families without evaluating all of their child poses allows for a large fraction of the complete set of poses to be eliminated without the time-consuming energy evaluation.

The 100 selected poses are then further refined with sidechain optimization using SCREAM (S15). Any sidechain that was alanized prior to docking is now restored and optimized ("de-alanized") by SCREAM. Simultaneously, any polar or charged sidechain in the binding site is also optimized by SCREAM, resulting in 100 unique sets of sidechain conformations each adapted to a specific ligand pose. Each complex is then energy optimized for 10 steps of conjugate gradient minimization. The minimized complexes are then scored using the "snap" binding energy, which is the total energy of the protein and the total energy of the ligand subtracted from the total energy of the complex, all calculated using DREIDING and MPSim. We then eliminated half of these complexes based on these energies. The remaining half was optimized with an additional 50 steps of conjugate gradient minimization. These fully-minimized complexes were rescored again, and the top one or two poses identified for analysis.

Closest-Neighbor Seeded Ligand Clustering

The Closest-Neighbor Seeded (CNS) ligand clustering algorithm uses a RMSD-based metric to cluster ligands into families and to assign family heads. First, all pairwise ligand RMSDs were calculated (ignoring hydrogen atoms). These pairwise RMSDs were placed in a list ordered from smallest RMSD to largest. The pair of ligands with the smallest RMSD constitutes the seed for the first family/cluster. Proceeding down the list of pairs *i* and *j*, the following operations were carried out:

1. If pose *i* and pose *j* do not belong to a pre-existing family, then a new family is seeded
2. If pose *i* belongs to family *A* and pose *j* does not belong to a family (or *vice versa*): If the RMSD of pose *j* to all members of family *A* is less than the diversity RMSD, then pose *j* is added to family *A*
3. If pose *i* belongs to family *A* and pose *j* belongs to family *B*: If the RMSD of pose *i* to all members of family *B* is less than the diversity RMSD, and if the RMSD of pose *j* to all members of family *A* is less than the diversity RMSD, then the two families are merged into a single family

The pose with the lowest RMSD to the rest of the members is designated as the family head. If a family only has two members then the family head is chosen randomly.

Forcefield

All forcefield calculations during docking – with the exception of sidechain optimizations – were performed using the DREIDING (S2) forcefield and the MPSim (S14) molecular dynamics code. DREIDING uses a three body hydrogen bond term that allows a more precise analysis of the energetics. It also eliminates the need of SHAKE constraints that must be used with the 2-body hydrogen bonds used in most force fields

Sidechain Optimization

Sidechain optimization was performed using the SCREAM program (using the DREIDING forcefield.

Sphere Generation

Spheres were generated using a modified *sphgen* program (S13). Specifically, two sets of spheres were generated for each protein:

The "normal" set:

- Use a 1.4 Å probe radius in the *dms* molecular surface program (44)
- Use dotlim=–0.9 in *sphgen*
- Use 1.4 Å minimum and 10 Å maximum sphere radii in *sphgen*

The "restriction" set:

- Use 2.8 Å probe radius in *dms*
- Use dotlim=–0.9 ' in *sphgen*
- Use 2.8 Å minimum and 10 Å maximum sphere radii in *sphgen*

The final set of spheres is taken from the "normal" set with the criteria that a sphere must be within 2.8 Å of a sphere from the "restriction" set.

The final set of spheres was partitioned into overlapping boxes having 20 Å sides and allowing 5 Å overlap.

As mentioned above, we assign electrostatic potential values to the spheres. The electrostatic potential for the protein is generated using APBS (S16-S18) and mapped onto the generated spheres. The electrostatic potential for a given sphere is taken from the value from the nearest APBS grid point.

Sphere Clustering

In order to reduce the number of spheres in each region to a computationally-manageable number, the spheres are clustered using the CNS algorithm, with each sphere treated as a single-atom ligand. The clustering diversity is set at 0.25 and increased until the total

number of families is less than 150, or until the diversity is 3.00. For sphere families with 3 or more spheres, the family head is kept to represent the family. For sphere families of 2 spheres, the coordinates are averaged.

Ligand Preparation

All crystal structure ligands were prepared by identifying the appropriate DREIDING atom types and assigning Mulliken charges from Density Functional Theory (DFT) calculations using the B3LYP level of theory and the 6-311G** basis set in Jaguar (S19, S27).

Heparin and CS ligands for the predicted systems (RPTP**s**, NgR, EphB2, and EphB3) were generated from available 18-mer heparin NMR structures (S20) and a 6-mer CS-A (S21) crystal structure as mentioned above. The heparin and CS-A structures were truncated or extended as needed for docking. Generating CS-D and CS-E required modifying the sulfation pattern of an extended 12-mer CS-A structure.

The sulfation pattern was modified using the Maestro software, Mulliken charges were calculated, and the MacroModel (S22) Conformational Search tool was used to sample the sidechain torsions (the sugar backbone was kept fixed). The resulting conformations were minimized using DREIDING and MPSim with Surface Generalized Born (SGB) solvation. The lowest-energy conformation was then selected for MD.

The AMBER (S23) package was used to place the 12-mer in a water box with a number of sodium ions added to neutralize the ligand charge. Dynamics was performed using NAMD (S24) in four steps as described in the next section. Briefly, first, the water box was minimized with the ligand kept fixed. Second 0.5 ns of MD were performed on the water box. Third, the entire system was minimized. Finally, 5 ns of MD were performed on the

entire system. The final ligand conformation for docking was the conformation closest to the average structure from the 5 ns MD. The 12-mer was truncated for docking by removing the terminal sugars.

Molecular Dynamics (MD)

The MD simulations were carried out using NAMD (S24), a parallel MD code designed for computationally demanding biomolecular systems. The CHARMM (S1) force field was used for the protein and ligands. The TIP3P (S25) force field was used for water. NAMD employs periodic boundary conditions to remove surface effects. The full electrostatic interactions within this periodic system is calculated using the particle-mesh Ewald summation method (S26). The long-range electrostatic and van der Waals interactions were cut off at 12 Å (with spline smoothing).

The calculations were performed under isothermal-isobaric conditions (NPT) at 310 K and 1 atm. The temperature was controlled using Langevin dynamics (with a coupling coefficient of 5 ps^{-1}) and the pressure is maintained using a Langevin-Hoover barostat. A time step of 1 fs was used throughout this study.

Simulations. The MD is carried out in 4 steps:

a) The water atoms and counter-ions were conjugate gradient minimized for 5000 steps while keeping the protein and ligand atoms fixed. This allows for the water and counter ions to remove any bad contacts with each other and the protein or the ligand, prior to MD.

b) Then the water and counter-ion atoms were equilibrated under NPT conditions (310 K and 1 atm) for 0.5 ns, while keeping the protein and ligand fixed. This allows the lipids

and waters to equilibrate in the presence of the protein and to fill any gaps around the protein created due to system setup.

c) Next, the full system (protein-ligand-water) was minimized for 5000 steps, allowing the protein and ligand to adjust to the equilibrated water and counter ions.

d) Finally, the full system is equilibrated for at least 5 ns under NPT conditions, of which only the last 5 ns is used for dynamical analysis. Snapshots are saved every 1 ps.

Supplemental References

S1. A D MacKerell, J., Bashford, D., Bellott, M., R L Dunbrack, J., Evanseck, J. D., Field, M. J., et al. (1998). All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *The Journal of Physical Chemistry B*, *102*(18), 3586–3616. http://doi.org/10.1021/jp973084f

S2. Mayo, S. L., Olafson, B. D., & Goddard, W. A. (2002). DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, *94*(26), 8897–8909. http://doi.org/10.1021/j100389a010

S3. Kim, D. E., Chivian, D., & Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, *32*(suppl 2), W526–W531. http://doi.org/10.1093/nar/gkh468

S4. Scott, C. E., Ahn, K. H., Graf, S. T., William A Goddard, I., Kendall, D. A., & Abrol, R. (2016). Computational Prediction and Biochemical Analyses of New Inverse Agonists for the CB1 Receptor. *Journal of Chemical Information and Modeling*, *56*(1), 201–212. http://doi.org/10.1021/acs.jcim.5b00581

S5. Li, Q., Kim, S.-K., Goddard, W. A., III, Chen, G., & Tan, H. (2015). Predicted Structures for Kappa Opioid G-Protein Coupled Receptor Bound to Selective Agonists. *Journal of Chemical Information and Modeling*, *55*(3), 614–627. http://doi.org/10.1021/ci500523z

S6. Abrol, R., Trzaskowski, B., Goddard, W. A., III, Nesterov, A., Olave, I., & Irons, C. (2014). Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, *111*(36), 13040–13045. http://doi.org/10.1073/pnas.1413216111

S7. Bray, J. K., Abrol, R., Goddard, W. A., III, Trzaskowski, B., & Scott, C. E. (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein–coupled receptor conformations. *Proceedings of the National Academy of Sciences*, *111*(1), E72–E78. http://doi.org/10.1073/pnas.1321233111

S8. Kim, S.-K., & Goddard, W. A. (2014). Predicted 3D structures of olfactory receptors with details of odorant binding to OR1G1. *Journal of Computer-Aided Molecular Design*, *28*(12), 1175–1190. http://doi.org/10.1007/s10822-014-9793-4

S9. Kim, S.-K., Goddard, W. A., Yi, K. Y., Lee, B. H., Lim, C. J., & Trzaskowski, B. (2014). Predicted Ligands for the Human Urotensin-II G Protein-Coupled Receptor with Some Experimental Validation. *ChemMedChem*, *9*(8), 1732–1743. http://doi.org/10.1002/cmdc.201402087

S10. Tan, J., Abrol, R., Trzaskowski, B., & William A Goddard, I. (2012). 3D Structure Prediction of TAS2R38 Bitter Receptors Bound to Agonists Phenylthiocarbamide (PTC) and 6-n-Propylthiouracil (PROP). *Journal of Chemical Information and Modeling*, *52*(7), 1875–1885. http://doi.org/10.1021/ci300133a

S11. Kim, S.-K., Fristrup, P., Abrol, R., & William A Goddard, I. (2011a). Structure-Based Prediction of Subtype Selectivity of Histamine H3 Receptor Selective Antagonists in Clinical Trials. *Journal of Chemical Information and Modeling*, *51*(12), 3262–3274. http://doi.org/10.1021/ci200435b

S12. Kim, S.-K., Riley, L., Abrol, R., Jacobson, K. A., & Goddard, W. A. (2011b). Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins: Structure, Function, and Bioinformatics*, *79*(6), 1878–1897. http://doi.org/10.1002/prot.23012

S13. Allen, W. J., Balius, T. E., Mukherjee, S., Brozell, S. R., Moustakas, D. T., Lang, P. T., et al. (2015). DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, *36*(15), 1132–1156. http://doi.org/10.1002/jcc.23905

S14. Lim, K. T., Brunett, S., Iotov, M., & McClurg, R. B. (1997). Molecular dynamics for very large systems on massively parallel computers: the MPSim program. *Journal of Computational Chemistry*, *18*(4), 501-521

S15. Kam, V. W. T., & William A Goddard, I. (2008). Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation*, *4*(12), 2160–2169. http://doi.org/10.1021/ct800196k

S16. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, *98*(18), 10037–10041. http://doi.org/10.1073/pnas.181342398

S17. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., & Baker, N. A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, *35*(suppl 2), W522–W525. http://doi.org/10.1093/nar/gkm276

S18. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, *32*(suppl 2), W665–W667. http://doi.org/10.1093/nar/gkh381

S19. Jaguar, version 8.9, Schrödinger, LLC, New York, NY, 2015.

S20. Khan, S., Gor, J., Mulloy, B., & Perkins, S. J. (2010). Semi-Rigid Solution Structures of Heparin by Constrained X-ray Scattering Modelling: New Insight into Heparin–Protein Complexes. *Journal of Molecular Biology*, *395*(3), 504–521. http://doi.org/10.1016/j.jmb.2009.10.064

S21. Winter, W. T., Arnott, S., Isaac, D. H., & Atkins, E. D. T. (1978). Chondroitin 4-sulfate: The structure of a sulfated glycosaminoglycan. *Journal of Molecular Biology*, *125*(1), 1–19. http://doi.org/10.1016/0022-2836(78)90251-6

S22. MacroModel, version 10.9, Schrödinger, LLC, New York, NY, 2015.

S23. D.A. Case, R.M. Betz, W. Botello-Smith, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I. Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2016), AMBER 2016, University of California, San Francisco.

S24. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16), 1781–1802. http://doi.org/10.1002/jcc.20289.

S25. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, *79*(2), 926–935. http://doi.org/10.1063/1.445869

S26. Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092. http://doi.org/10.1063/1.464397

S27. Bochevarov, A. D., Harder, E., Hughes, T. F., Greenwood, J. R., Braden, D. A., Philipp, D. M., et al. (2013). Jaguar: A high-performance quantum chemistry software program with strengths in life and materials sciences. *International Journal of Quantum Chemistry*, *113*(18), 2110–2142. http://doi.org/10.1002/qua.24481

Supplemental Figures & Tables

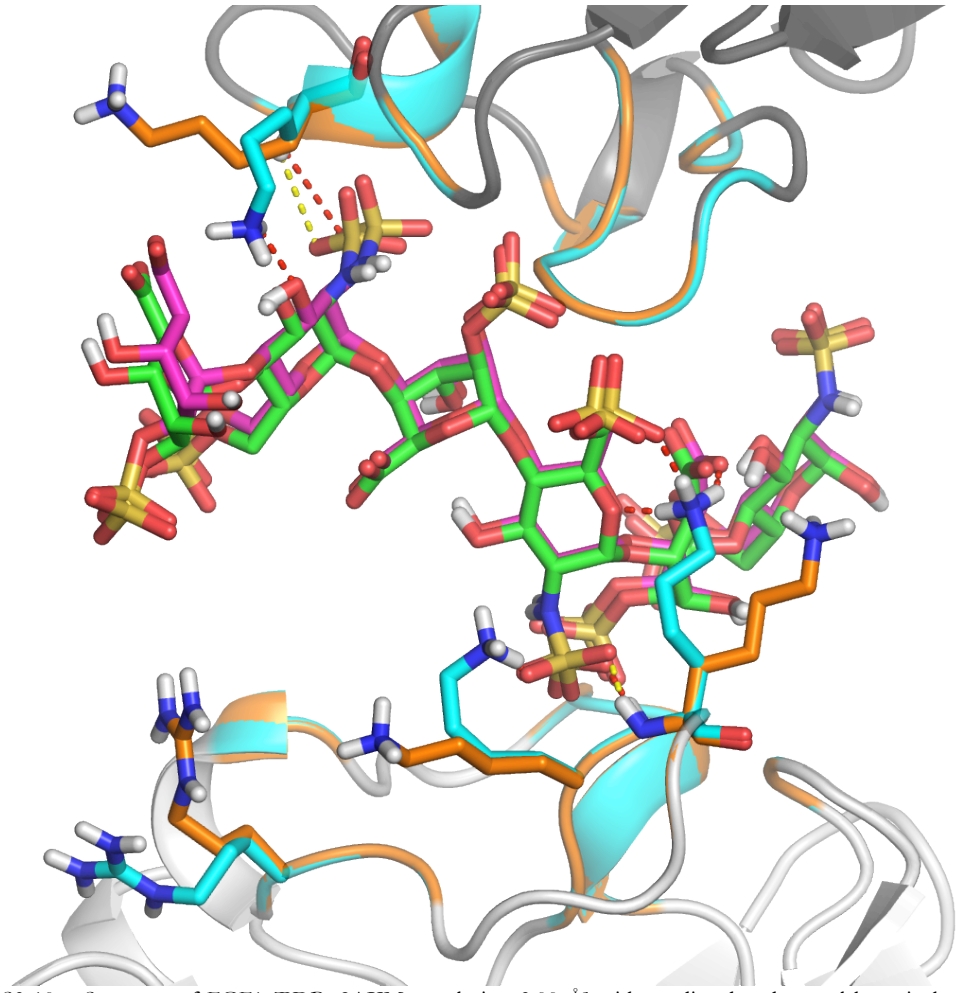

Figure S3-10 – Structure of FGF1 [PDB: 2AXM, resolution 3.00 Å] with predicted and crystal heparin hexamer ligands (magenta: predicted, green: crystal). Residues in the binding site with significant deviations from the crystal are labeled (cyan: predicted, orange: crystal). Ligand RMSD is 0.70 Å.

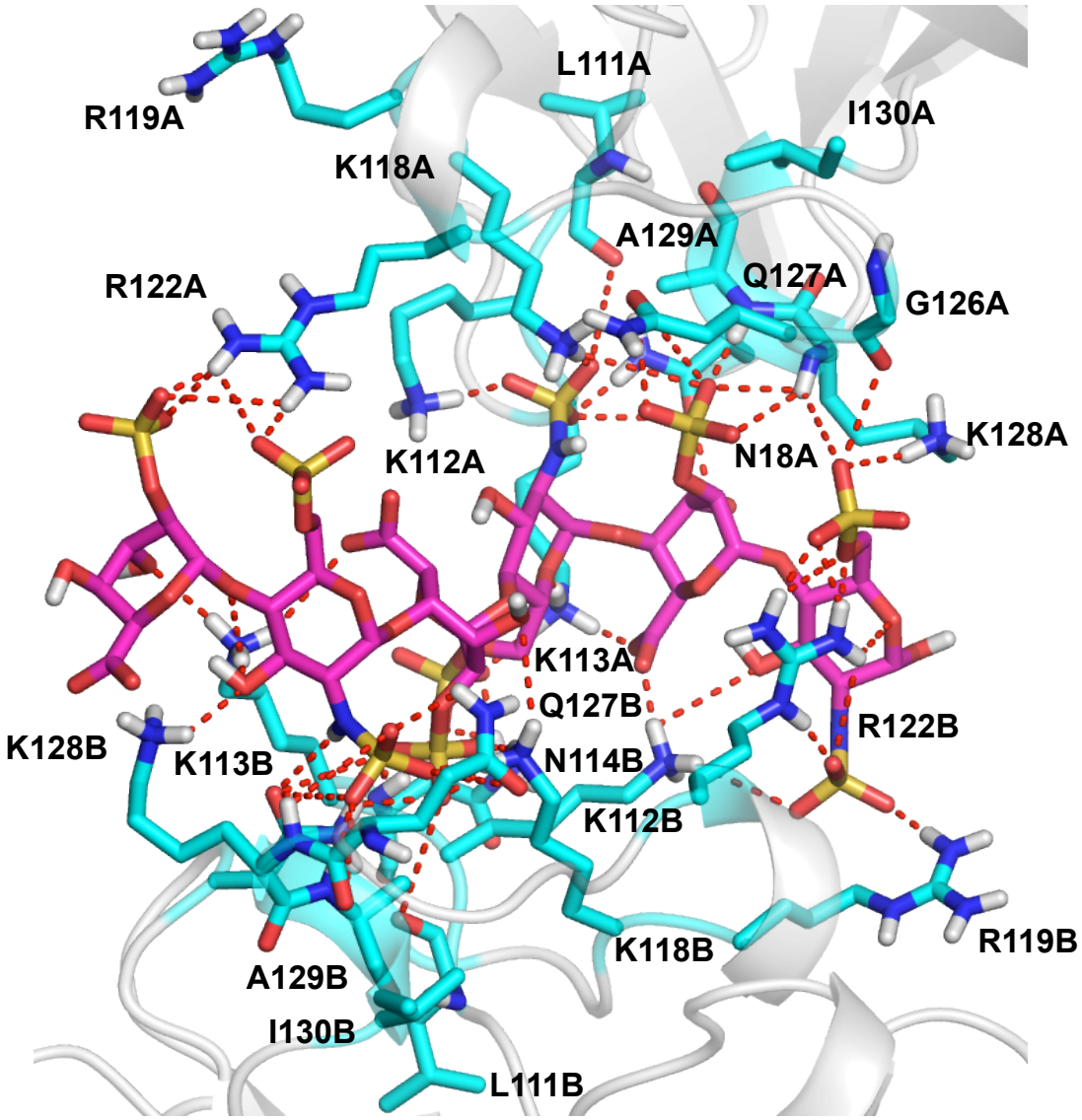

Figure S3-11– Structure of FGF1 [PDB: 2AXM, res. 3.00 Å] with predicted heparin hexamer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 0.70 Å.

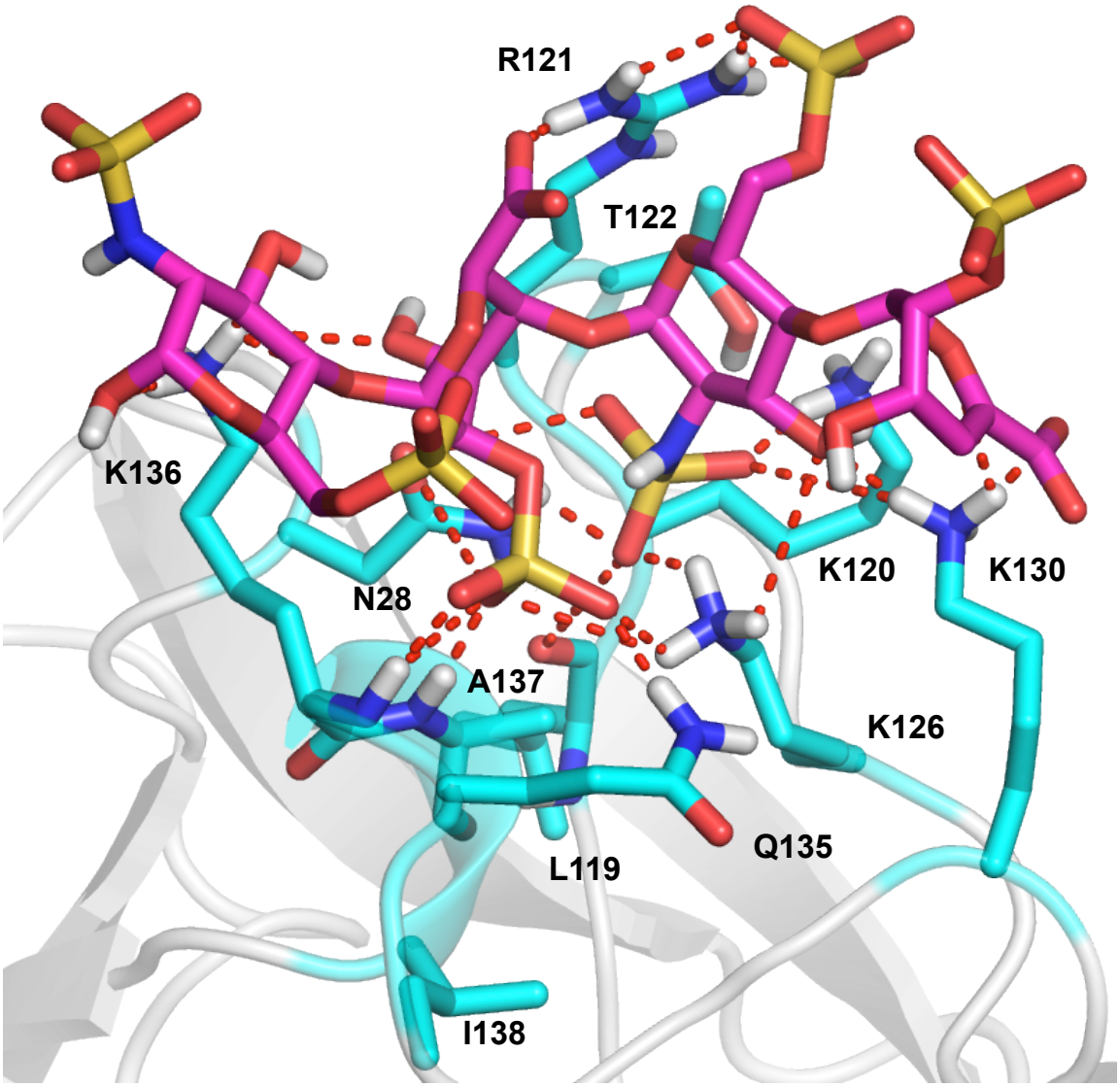

Figure S3-12 – Structure of FGF2 [PDB: 1BFB, res. 1.90 Å] with predicted heparin tetramer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 0.70 Å.

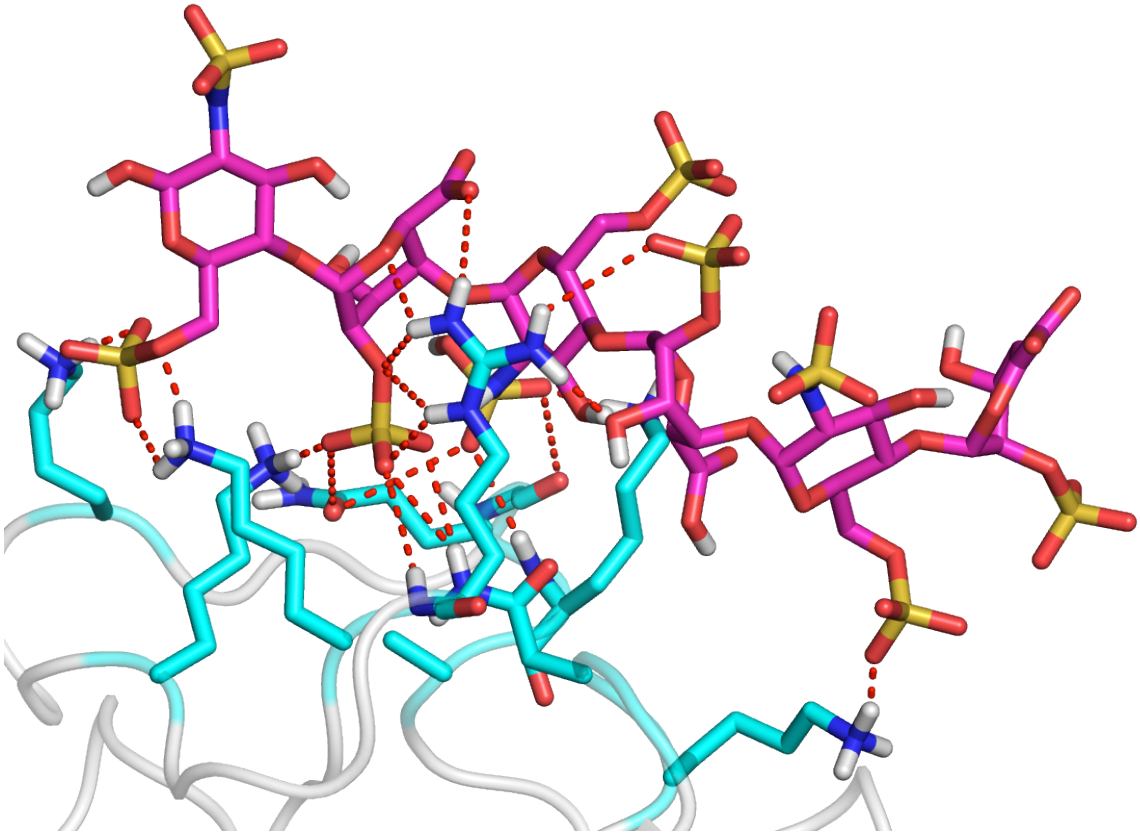

Figure S3-13 – Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain A with predicted heparin hexamer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å.

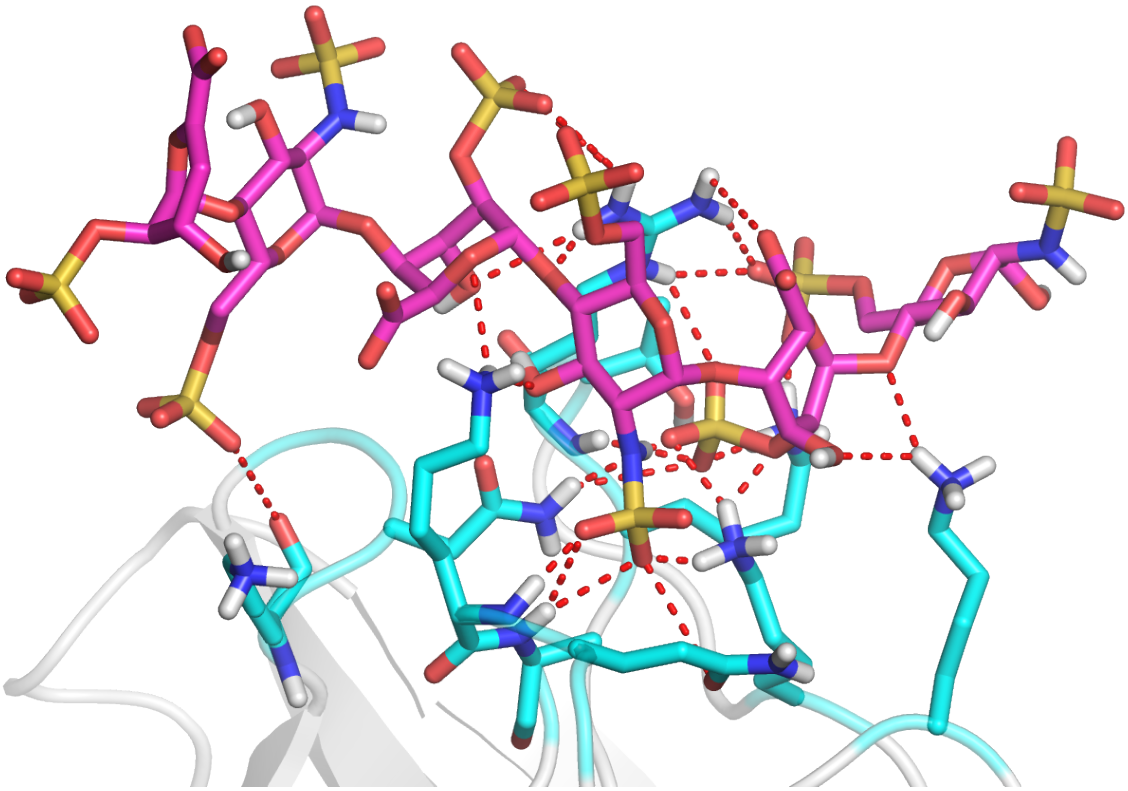

Figure S3-14 – Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain B with predicted heparin hexamer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å.

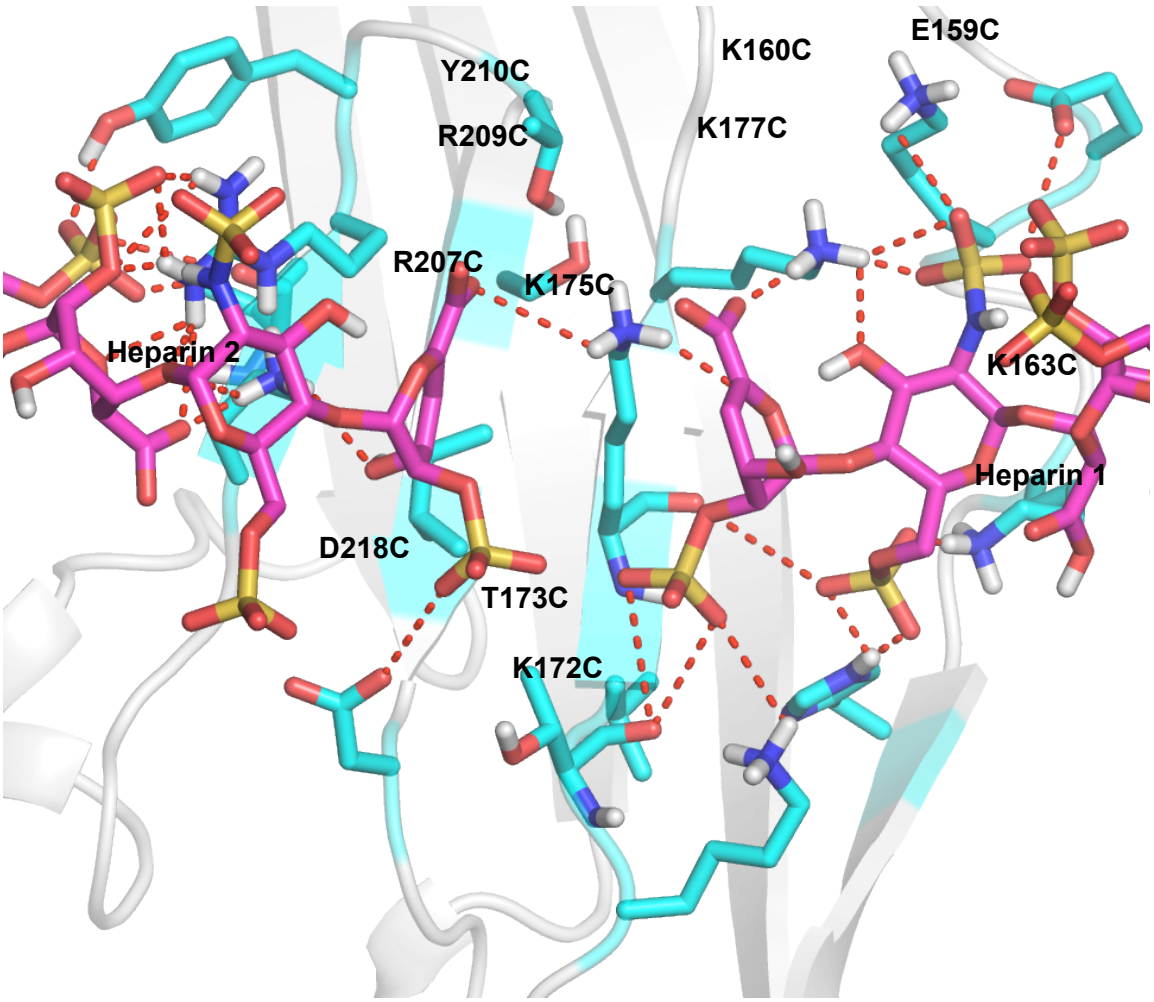

Figure S3-15 – Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain C with predicted heparin hexamer and octamer ligands (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å.

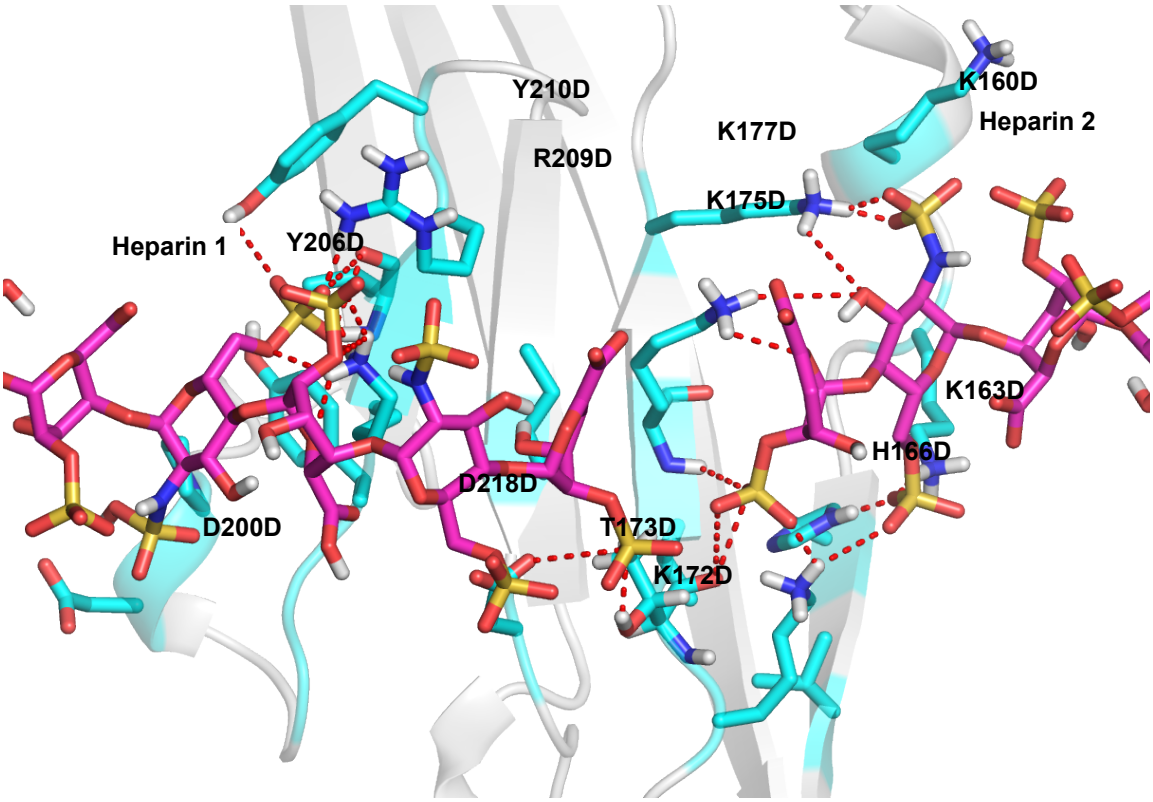

Figure S3-16 – Structure of FGF2-FGFR1 [PDB: 1FQ9, res. 3.00 Å] chain C with predicted heparin hexamer and octamer ligands (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 1.51/0.75 Å.

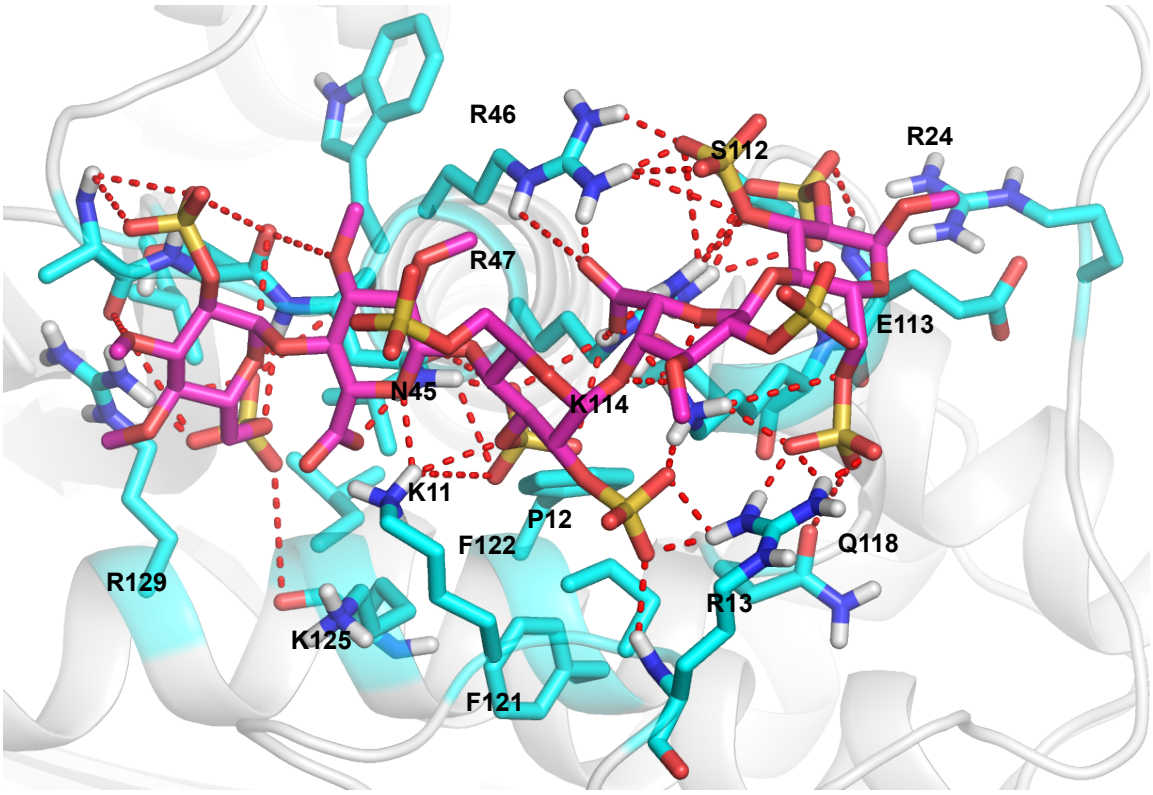

Figure S3-17 – Structure of α-Antithrombin-III [PDB: 1E03, res. 2.90 Å] with predicted heparin analog pentamer ligand (magenta) and 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. The predicted ligand has excellent agreement with the crystal ligand, RMSD: 0.60 Å.

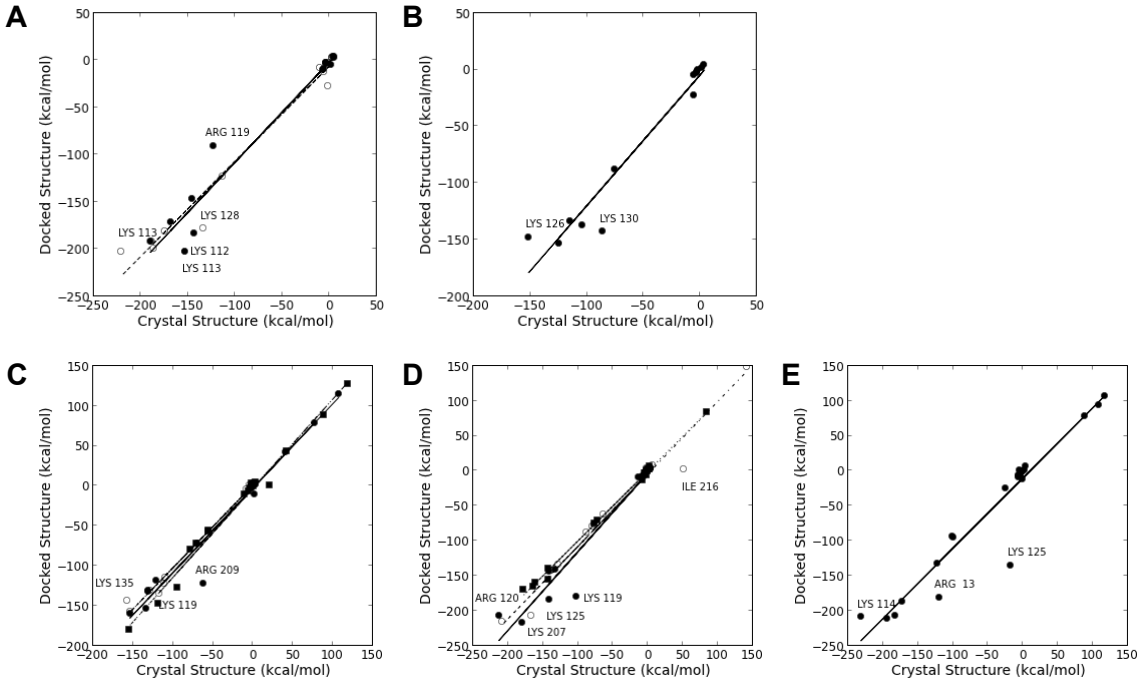

Figure S3-18 – Plots of nonbond energies for residues in the (A) FGF1, (B) FGF2, (C) FGF2-FGFR1 Chain A complex, (D) FGF2-FGFR1 Chain B complex, and (E) α-Antithrombin-III binding sites in complex with a heparin ligand in the crystal versus docked structure. Residues with significant deviations from the trend are labeled.

Table S3-11 – Per-residue energetic contributions in the FGF1/heparin predicted (left) and crystal (right) structures. [PDB: 2AXM, res. 3.00 Å, RMSD: 0.70 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain A	LYS 113	17.56	-204.76	-15.75	-202.95	2.77	1.19	-2.00	1.96	-204.91
	LYS 118	3.22	-189.23	-5.96	-191.97	-0.56	4.09	0.00	3.54	-195.51
	LYS 112	3.29	-183.50	-3.81	-184.01	-0.33	-143.67	0.00	-144.00	-40.01
	ARG 122	7.16	-169.14	-9.33	-171.31	3.14	-153.18	-3.12	-153.17	-18.14
	LYS 128	3.06	-148.57	-1.85	-147.37	-0.41	4.77	0.00	4.37	-151.74
	ARG 119	-0.11	-91.00	0.00	-91.11	-0.06	-3.21	0.00	-3.27	-87.84
	GLN 127	-2.00	-8.03	0.00	-10.03	-1.99	-186.44	-1.59	-190.03	180.00
	ALA 129	5.47	-10.00	-4.60	-9.14	-0.38	-122.47	0.00	-122.86	113.72
	ASN 18	2.01	-1.44	-5.71	-5.13	-1.52	-165.54	-1.67	-168.72	163.59
	ILE 130	-0.32	-3.07	0.00	-3.39	-0.88	5.58	0.00	4.70	-8.08
	GLY 115	-0.04	-3.22	0.00	-3.27	-0.13	-6.95	0.00	-7.08	3.81
	GLY 126	-0.87	3.54	0.00	2.66	-4.08	-140.24	-1.57	-145.89	148.55
	LEU 111	-0.48	3.51	0.00	3.03	3.28	-8.41	-1.05	-6.18	9.21
	ASN 114	-0.23	3.33	0.00	3.10	-0.30	-2.97	0.00	-3.28	6.38
Chain B	LYS 113	4.89	-197.61	-9.67	-202.39	1.11	-215.76	-5.84	-220.49	18.10
	LYS 112	4.24	-198.73	-5.46	-199.96	2.69	-190.94	-0.44	-188.69	-11.26
	LYS 118	11.00	-197.81	-5.73	-192.53	-4.61	-181.97	0.00	-186.58	-5.96
	ARG 122	4.88	-176.18	-10.38	-181.68	2.69	-175.17	-1.89	-174.37	-7.31
	LYS 128	10.38	-181.92	-6.98	-178.52	2.06	-135.42	0.00	-133.36	-45.16
	ARG 119	7.42	-125.16	-5.83	-123.58	-2.11	-111.44	0.00	-113.55	-10.03
	ASN 114	5.11	-22.15	-10.47	-27.51	-0.34	-7.68	-1.92	-9.94	-17.57
	GLN 127	2.94	-13.43	-1.56	-12.05	-0.60	-5.77	0.00	-6.37	-5.68
	ALA 129	6.73	-10.22	-4.49	-7.99	-0.14	-3.90	0.00	-4.04	-3.94
	GLY 115	-0.14	-4.03	0.00	-4.17	-0.27	-2.84	0.00	-3.11	-1.07
	ILE 130	-0.37	-3.13	0.00	-3.50	-1.63	2.88	-3.24	-1.99	-1.50
	ASN 18	0.24	2.05	-5.05	-2.76	-2.38	1.24	-0.18	-1.32	-1.44
	GLY 126	-0.24	2.42	0.00	2.18	-0.56	2.92	0.00	2.36	-0.17
	LEU 111	-0.72	2.97	0.00	2.25	-0.25	2.86	0.00	2.61	-0.36

Table S3-12 – Per-residue energetic contributions in the FGF2/heparin predicted (left) and crystal (right) structures. [PDB: 1BFB, res. 1.90 Å, RMSD: 0.70 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain A	ARG 121	17.30	-158.68	-12.19	-153.56	-0.27	-150.23	-0.96	-151.45	-2.11
	LYS 126	9.10	-156.65	-0.36	-147.92	-3.44	-119.39	-2.31	-125.14	-22.78
	LYS 130	10.15	-144.73	-8.56	-143.15	-5.00	-107.74	-1.75	-114.49	-28.65
	LYS 120	8.16	-140.09	-5.58	-137.52	-2.88	-101.57	0.00	-104.45	-33.07
	LYS 136	6.73	-135.14	-5.78	-134.19	-0.75	-85.16	0.00	-85.91	-48.28
	LYS 27	-0.42	-87.54	0.00	-87.96	-0.28	-75.32	0.00	-75.60	-12.37
	GLN 135	2.17	-18.79	-6.09	-22.72	-2.77	-2.99	0.00	-5.76	-16.96
	ALA 137	7.94	-9.51	-3.59	-5.16	5.61	-8.31	-2.66	-5.37	0.21
	ILE 138	-0.28	-2.34	0.00	-2.62	-2.44	1.25	-1.53	-2.72	0.10
	ASN 28	3.07	-1.79	-3.76	-2.47	-0.28	-2.29	0.00	-2.57	0.10
	THR 122	-0.40	0.02	0.00	-0.37	-0.31	-1.32	0.00	-1.63	1.26
	GLY 134	-0.37	2.18	0.00	1.81	-0.43	2.34	0.00	1.92	-0.11
	LEU 119	-0.62	4.44	0.00	3.82	-0.45	4.03	0.00	3.57	0.25

Table S3-13 – Per-residue energetic contributions in the FGF2-FGFR1/heparin predicted (left) and crystal (right) structures for chains A and B. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain A	ARG 120	-2.28	-149.49	-5.26	-157.04	1.52	-152.06	-7.27	-157.81	0.77
	LYS 135	1.54	-139.58	-5.91	-143.95	4.60	-153.99	-3.79	-153.18	9.24
	LYS 119	0.67	-131.00	-4.78	-135.12	3.14	-130.93	-0.25	-128.04	-7.08
	LYS 125	5.84	-131.09	-5.87	-131.11	0.62	-116.95	-0.65	-116.98	-14.13
	LYS 26	4.39	-120.65	-5.68	-121.94	-1.02	-109.45	0.00	-110.47	-11.47
	LYS 129	4.99	-114.65	-5.58	-115.25	-1.24	-107.92	-0.68	-109.84	-5.41
	ALA 136	-0.53	-3.71	0.00	-4.24	2.86	-6.00	-4.60	-7.74	3.51
	GLY 28	-0.37	-2.08	0.00	-2.45	-2.02	1.97	-3.62	-3.67	1.22
	ILE 137	-0.12	-1.95	0.00	-2.07	-0.66	-1.95	0.00	-2.61	0.53
	GLN 134	-1.91	0.68	0.00	-1.23	-0.29	-2.08	0.00	-2.36	1.14
	ASN 27	-1.89	2.20	0.00	0.31	-0.48	-1.59	0.00	-2.08	2.39
	THR 121	-0.20	0.54	0.00	0.34	-0.17	-1.53	0.00	-1.70	2.04
	TYR 24	-0.14	1.63	0.00	1.49	4.63	-4.77	0.00	-0.14	1.63
	LEU 118	-0.15	2.35	0.00	2.20	-0.31	2.16	0.00	1.85	0.35
	GLY 133	-0.32	2.55	0.00	2.23	-0.50	2.69	0.00	2.19	0.04
	LEU 126	-0.14	2.56	0.00	2.42	-0.28	2.62	0.00	2.34	0.08
Chain B	LYS 135	1.19	-212.03	-6.46	-217.30	-8.49	-203.03	-1.48	-212.99	-4.31
	ARG 120	-6.58	-199.88	-1.36	-207.81	31.40	-208.27	-3.53	-180.39	-27.41
	LYS 125	5.60	-182.56	-6.84	-183.80	38.12	-175.56	-3.61	-141.05	-42.75
	LYS 119	12.36	-181.40	-10.45	-179.49	-2.14	-138.68	0.00	-140.82	-38.67
	LYS 129	-1.92	-141.31	-0.25	-143.48	-1.59	-130.86	0.00	-132.44	-11.04
	LYS 26	-2.18	-138.89	0.00	-141.07	37.15	-138.94	0.00	-101.78	-39.29
	GLN 134	-2.27	-6.73	0.00	-9.00	9.71	-20.50	-3.13	-13.92	4.92
	ALA 136	-1.24	-7.10	0.00	-8.34	-1.58	-6.76	0.00	-8.33	-0.01
	GLY 28	-0.34	-3.40	0.00	-3.74	-0.40	-2.89	0.00	-3.29	-0.45
	THR 121	-1.70	2.18	0.00	0.48	-0.16	-1.81	0.00	-1.97	2.44
	ASN 27	-2.39	4.25	0.00	1.86	3.82	-3.11	-0.03	0.68	1.18
	TYR 24	-0.15	2.25	0.00	2.10	-2.71	5.94	-0.05	3.18	-1.08
	GLY 133	-0.21	3.30	0.00	3.09	-0.25	3.55	0.00	3.30	-0.22
	LEU 118	-0.33	3.65	0.00	3.32	-0.52	3.96	0.00	3.44	-0.12

Table S3-14 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-A predicted (left) and crystal (right) structures for chain C. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain C	LYS 177	7.03	-159.17	-8.16	-160.30	5.72	-158.41	-1.26	-153.95	-6.35
	LYS 175	0.73	-149.86	-4.69	-153.82	2.31	-130.03	-5.88	-133.60	-20.22
	LYS 163	6.37	-132.42	-6.11	-132.16	1.23	-132.09	-0.39	-131.25	-0.91
	LYS 160	0.37	-131.48	0.00	-131.12	-1.30	-129.72	-0.02	-131.03	-0.08
	ARG 209	-0.44	-117.34	-4.40	-122.18	-0.82	-120.14	0.00	-120.97	-1.21
	LYS 172	-0.78	-118.38	0.00	-119.16	-0.04	-70.04	0.00	-70.08	-49.08
	LYS 207	-0.04	-71.70	0.00	-71.75	-0.04	-61.48	0.00	-61.53	-10.22
	HSE 166	5.52	-12.11	-4.86	-11.44	-1.33	-3.41	0.00	-4.74	-6.70
	VAL 174	-0.99	-3.05	0.00	-4.04	-0.41	-2.90	0.00	-3.30	-0.74
	ILE 216	-0.52	-3.02	0.00	-3.54	-0.02	-1.90	0.00	-1.92	-1.62
	SER 219	-0.02	-1.90	0.00	-1.92	-0.02	-1.66	0.00	-1.68	-0.24
	TYR 206	-0.02	-1.58	0.00	-1.60	-0.01	-0.82	0.00	-0.83	-0.76
	VAL 208	-0.02	-0.86	0.00	-0.87	0.00	0.29	0.00	0.29	-1.16
	PRO 199	0.00	0.24	0.00	0.24	0.00	1.09	0.00	1.09	-0.85
	THR 173	-1.59	2.37	0.00	0.79	0.00	1.64	0.00	1.64	-0.85
	GLY 204	0.00	1.13	0.00	1.13	-1.28	3.00	0.00	1.72	-0.59
	GLY 205	0.00	1.63	0.00	1.63	-1.73	3.49	0.00	1.76	-0.13
	TYR 210	-0.02	2.19	0.00	2.18	-0.01	1.86	0.00	1.84	0.33
	ASP 200	0.00	41.39	0.00	41.39	0.00	41.11	0.00	41.10	0.28
	ASP 218	-0.09	78.09	0.00	78.01	-0.08	77.16	0.00	77.08	0.93
	GLU 159	-0.71	115.33	0.00	114.62	-0.45	107.49	0.00	107.04	7.59

Table S3-15 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-A predicted (left) and crystal (right) structures for chain D. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain D	LYS 207	7.45	-180.81	-6.56	-179.92	17.13	-171.90	-0.35	-155.12	-24.79
	ARG 209	3.65	-146.18	-5.18	-147.70	-0.83	-117.34	0.00	-118.16	-29.54
	LYS 175	5.98	-126.60	-6.24	-126.87	-0.40	-94.29	0.00	-94.69	-32.18
	LYS 172	-0.12	-79.28	0.00	-79.40	-0.12	-78.00	0.00	-78.11	-1.29
	LYS 177	-0.04	-72.16	0.00	-72.20	-0.03	-70.11	0.00	-70.14	-2.06
	LYS 160	-0.01	-56.66	0.00	-56.66	0.00	-55.42	0.00	-55.42	-1.25
	LYS 163	-0.01	-56.16	0.00	-56.16	-0.01	-55.17	0.00	-55.17	-0.99
	THR 173	3.72	-10.80	-3.54	-10.63	5.19	-10.69	-4.75	-10.26	-0.37
	TYR 210	-0.96	-5.72	0.00	-6.69	-0.25	-5.06	0.00	-5.30	-1.38
	PRO 199	-1.08	-2.70	0.00	-3.77	-0.45	-2.77	0.00	-3.23	-0.55
	HSE 166	-0.01	-2.21	0.00	-2.22	-0.01	-2.10	0.00	-2.11	-0.12
	GLY 205	-0.32	0.09	0.00	-0.23	-0.12	-1.83	0.00	-1.95	1.72
	ILE 216	0.51	-0.44	0.00	0.07	-0.22	0.36	0.00	0.14	-0.07
	TYR 206	-0.82	1.99	0.00	1.18	-0.69	3.14	0.00	2.45	-1.28
	SER 219	-0.10	2.54	0.00	2.44	-0.30	2.83	0.00	2.53	-0.08
	GLY 204	-0.30	2.82	0.00	2.52	-0.14	3.05	0.00	2.91	-0.39
	VAL 174	-0.15	3.01	0.00	2.86	-1.07	4.05	0.00	2.98	-0.12
	VAL 208	-1.09	5.42	-0.55	3.78	20.87	-0.03	0.00	20.85	-17.07
	GLU 159	0.00	43.02	0.00	43.02	0.00	42.22	0.00	42.22	0.80
	ASP 200	-0.29	88.12	0.00	87.83	-0.18	89.15	0.00	88.97	-1.14
	ASP 218	-2.17	129.70	0.00	127.53	-1.74	120.80	0.00	119.06	8.47

Table S3-16 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-B predicted (left) and crystal (right) structures for chain C. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain C	ARG 209	1.18	-211.75	-4.98	-215.55	-5.88	-198.45	-3.76	-208.09	-7.46
	LYS 207	12.57	-210.11	-9.85	-207.39	38.55	-204.68	-1.08	-167.22	-40.17
	LYS 175	7.82	-136.79	-5.80	-134.77	-0.95	-127.70	0.00	-128.65	-6.12
	LYS 177	-0.07	-88.11	0.00	-88.18	-0.06	-87.73	0.00	-87.79	-0.39
	LYS 172	-0.05	-79.78	0.00	-79.83	-0.04	-79.56	0.00	-79.60	-0.23
	LYS 160	-0.01	-76.48	0.00	-76.49	-0.01	-76.64	0.00	-76.65	0.16
	LYS 163	-0.01	-63.13	0.00	-63.14	-0.01	-63.12	0.00	-63.12	-0.02
	TYR 210	1.90	-7.94	-3.75	-9.79	-0.36	-6.01	0.00	-6.37	-3.42
	THR 173	-0.99	-7.60	0.00	-8.59	-0.78	-5.25	0.00	-6.03	-2.56
	VAL 168	-0.01	-1.12	0.00	-1.13	-0.01	-1.10	0.00	-1.11	-0.02
	SER 214	-0.31	0.98	0.00	0.68	-0.25	-1.54	0.00	-1.79	2.47
	HSE 166	-0.01	1.28	0.00	1.28	-0.01	1.35	0.00	1.34	-0.07
	ILE 216	2.12	-0.30	0.00	1.82	51.25	-0.18	0.00	51.07	-49.25
	VAL 174	-0.12	3.36	0.00	3.24	-0.10	3.33	0.00	3.24	0.01
	THR 212	-0.65	4.14	0.00	3.49	-0.55	-1.10	0.00	-1.65	5.14
	VAL 208	-1.18	8.37	0.00	7.19	-0.79	8.04	0.00	7.24	-0.05
	ASP 218	-0.99	148.97	0.00	147.98	-0.74	142.16	0.00	141.42	6.56

Table S3-17 – Per-residue energetic contributions in the FGF2-FGFR1/heparin-B predicted (left) and crystal (right) structures for chain D. [PDB: 1FQ9, res. 3.00 Å, RMSD: 1.51/0.75 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain D	LYS 177	8.65	-171.98	-6.19	-169.52	0.20	-174.51	-3.33	-177.64	8.12
	LYS 175	1.59	-162.42	-4.93	-165.76	-2.60	-161.79	-0.16	-164.55	-1.22
	LYS 172	-0.28	-159.18	-0.78	-160.24	-1.62	-158.79	-0.63	-161.05	0.80
	LYS 163	6.54	-157.81	-4.80	-156.06	-1.33	-141.73	-0.02	-143.09	-12.98
	LYS 160	-1.22	-138.51	0.00	-139.73	3.52	-146.18	0.00	-142.67	2.94
	LYS 207	-0.03	-75.10	0.00	-75.13	-0.04	-75.73	0.00	-75.77	0.64
	ARG 209	-0.04	-71.55	0.00	-71.58	-0.05	-72.21	0.00	-72.25	0.67
	HSE 166	5.13	-14.03	-4.53	-13.43	-0.83	-6.97	0.00	-7.80	-5.63
	SER 214	-0.07	-5.85	0.00	-5.92	-0.10	-2.00	0.00	-2.09	-3.82
	VAL 174	-0.95	-3.15	0.00	-4.10	-1.09	-3.17	0.00	-4.26	0.17
	ILE 216	-0.31	-3.34	0.00	-3.64	-0.44	-3.31	0.00	-3.75	0.11
	VAL 168	-0.27	-2.77	0.00	-3.03	-0.33	-2.64	0.00	-2.97	-0.06
	VAL 208	-0.01	-0.83	0.00	-0.84	-0.02	-0.93	0.00	-0.94	0.10
	THR 212	-0.10	1.83	0.00	1.74	-0.13	0.41	0.00	0.28	1.46
	TYR 210	-0.01	1.78	0.00	1.77	-0.01	2.20	0.00	2.19	-0.41
	THR 173	-1.69	7.34	0.00	5.65	-1.89	3.99	0.00	2.10	3.56
	ASP 218	-0.06	83.76	0.00	83.70	-0.07	83.74	0.00	83.68	0.02

Table S3-18 – Per-residue energetic contributions in the Antithrombin-III/heparin analog predicted (left) and crystal (right) structures. [PDB: 1E03, res. 2.90 Å, RMSD: 0.60 Å].

		Docked				Crystal				
	Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**$\Delta_{NonBond}$**
Chain I	ARG 47	5.37	-206.85	-10.73	-212.21	5.88	-194.99	-4.73	-193.85	-18.36
	LYS 114	12.38	-211.46	-10.33	-209.40	-5.92	-219.34	-6.42	-231.68	22.28
	LYS 11	3.13	-202.41	-7.42	-206.70	-6.64	-175.36	0.00	-182.00	-24.70
	ARG 46	9.21	-187.90	-8.44	-187.12	0.31	-171.42	-1.86	-172.98	-14.15
	ARG 13	9.47	-177.67	-12.91	-181.11	-0.70	-117.01	-1.69	-119.39	-61.71
	LYS 125	-1.21	-134.22	0.00	-135.43	176.01	-192.82	-0.26	-17.07	-118.36
	ARG 129	4.95	-129.62	-8.03	-132.71	-0.61	-120.34	-2.08	-123.03	-9.68
	ARG 24	-0.29	-95.38	0.00	-95.67	-0.16	-99.79	0.00	-99.95	4.29
	ARG 132	-0.15	-93.62	0.00	-93.77	-0.36	-100.85	0.00	-101.21	7.44
	ASN 45	-2.51	-17.11	-5.46	-25.08	-4.27	-15.33	-4.67	-24.27	-0.81
	SER 112	-1.18	-11.59	0.00	-12.77	-1.47	1.62	0.00	0.16	-12.93
	PRO 12	-1.66	-8.21	0.00	-9.87	3.69	-10.49	0.00	-6.80	-3.07
	THR 44	-4.47	-3.49	0.00	-7.96	-2.79	0.12	0.00	-2.67	-5.29
	ALA 43	-3.12	-4.65	0.00	-7.77	-3.92	-1.74	0.00	-5.66	-2.12
	VAL 48	-1.15	-5.49	0.00	-6.64	0.26	-6.05	0.00	-5.79	-0.86
	PHE 122	-1.04	-1.68	0.00	-2.72	0.26	-1.98	0.00	-1.71	-1.01
	ILE 40	-0.40	-1.93	0.00	-2.33	-0.29	-2.29	0.00	-2.57	0.25
	PHE 121	-0.37	-1.23	0.00	-1.60	-0.72	-1.42	0.00	-2.14	0.54
	LEU 126	-0.17	-0.39	0.00	-0.56	-0.31	0.04	0.00	-0.27	-0.29
	LEU 417	-0.20	-0.32	0.00	-0.53	-0.24	-0.48	0.00	-0.72	0.20
	GLN 118	-0.23	0.24	0.00	0.01	-0.50	-4.03	0.00	-4.53	4.55
	THR 115	-0.42	0.54	0.00	0.12	-0.42	3.03	0.00	2.61	-2.49
	PRO 41	-0.26	5.91	0.00	5.65	-0.17	4.70	0.00	4.53	1.12
	GLU 42	-0.54	79.06	0.00	78.53	-0.39	89.43	0.00	89.04	-10.51
	GLU 113	1.90	92.68	-1.04	93.55	11.78	97.20	0.00	108.98	-15.43
	ASP 14	-0.11	107.49	0.00	107.38	-0.30	116.97	0.00	116.66	-9.29

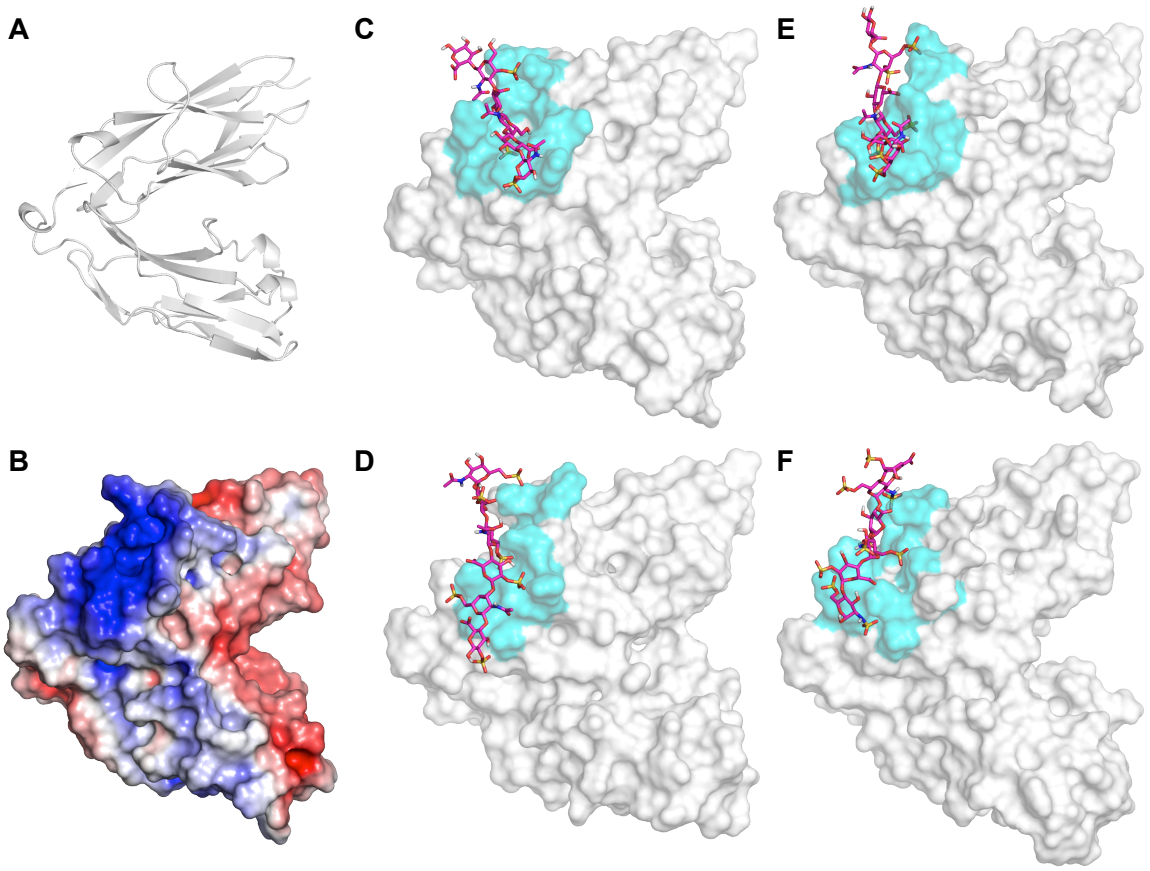

Figure S3-19 – RPTPσ. (A) Ig1 and Ig2 domains of RPTPσ. (B) Electrostatic potential surface. (C-F) Predicted structures of CS-A, CS-D, CS-E, and heparin after docking and molecular dynamics.

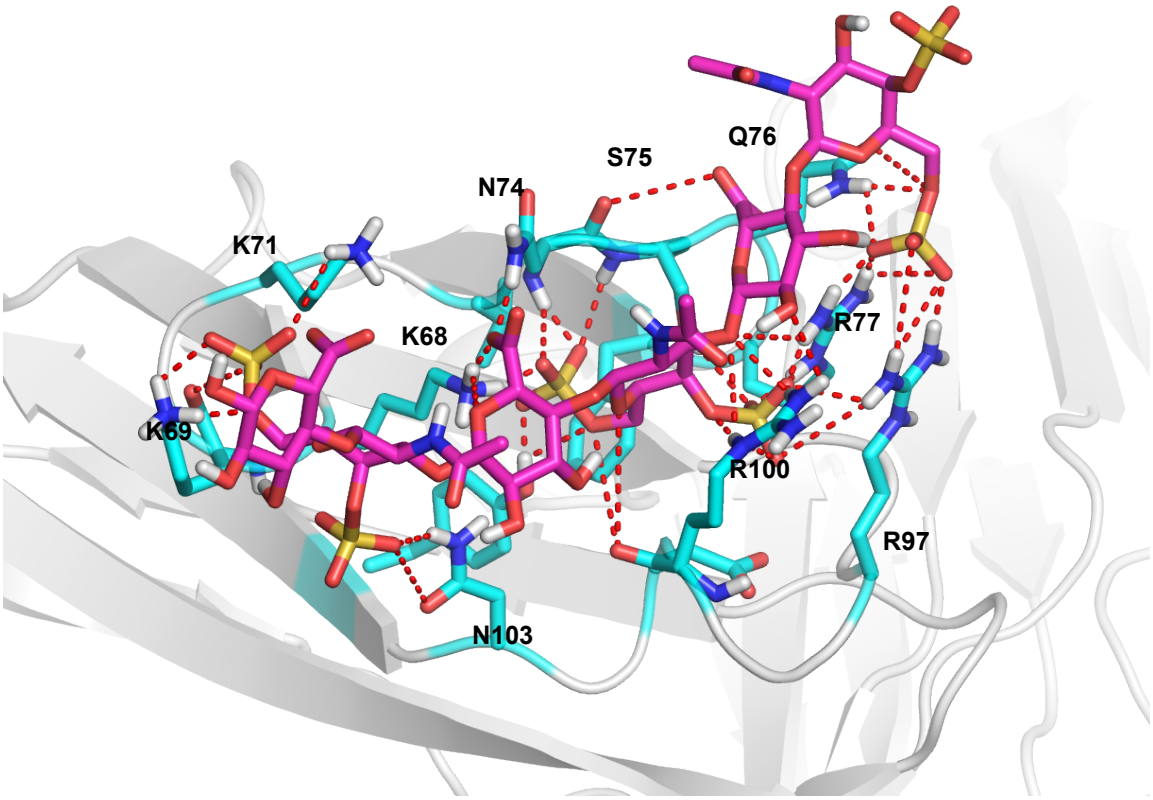

Figure S3-20 – Predicted structure of CS-E hexamer (magenta) bound to RPTPσ with 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

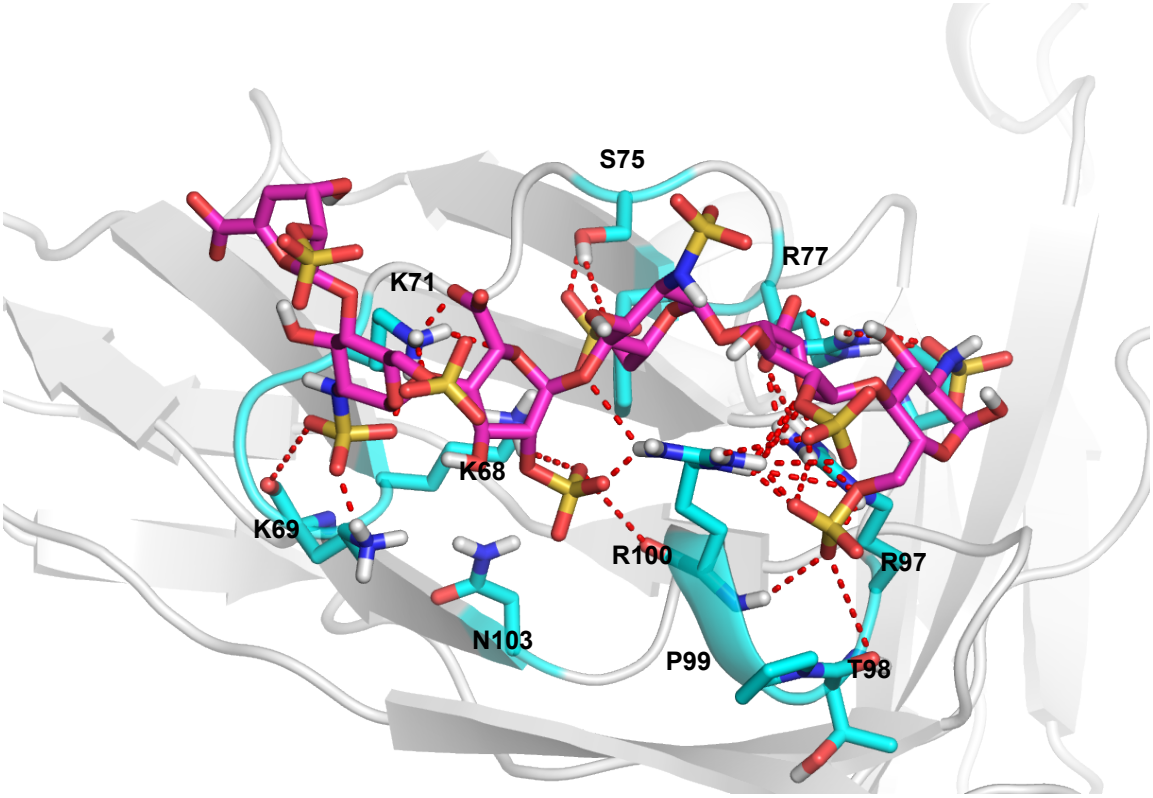

Figure S3-21 – Predicted structure of heparin hexamer (magenta) bound to RPTPσ with 5 Å binding site shown (cyan). Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

Table S3-19 – Per-residue energetic contributions in the predicted RPTPσ/CS-A (left) and RPTPσ/CS-D (right) structures.

	CS-A				CS-D			
Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**
ARG 77	2.59	-109.82	-8.26	-115.50	2.85	-152.95	-10.80	-160.91
LYS 68	9.11	-108.48	-12.32	-111.69	-0.54	-100.52	0.00	-101.06
ARG 100	-2.41	-90.30	-2.04	-94.75	11.00	-174.16	-25.17	-188.33
ARG 97	9.71	-89.82	-11.85	-91.97	-0.68	-134.41	-4.42	-139.50
LYS 71	-1.67	-85.86	0.00	-87.54	14.25	-141.64	-11.80	-139.20
LYS 69	2.01	-99.90	-6.62	-104.51	-0.19	-76.14	0.00	-76.33
LYS 72	--	--	--	--	-0.46	-75.57	0.00	-76.03
GLN 76	-1.67	-7.74	-4.48	-13.89	1.86	-16.06	-8.45	-22.65
ASN 74	-0.86	2.13	0.00	1.27	-0.30	-0.29	-5.86	-6.45
SER 75	4.90	-12.36	-4.75	-12.21	-2.31	-6.93	0.00	-9.24
ASN 103	-1.20	-2.24	0.00	-3.44	-0.15	-1.06	0.00	-1.20
GLU 102	-0.23	48.03	0.00	47.80	--	--	--	--
GLY 70	--	--	--	--	--	--	--	--
PRO 95	--	--	--	--	-1.05	1.32	0.00	0.27
PRO 99	--	--	--	--	--	--	--	--
THR 98	--	--	--	--	-0.19	1.83	0.00	1.65
VAL 73	-0.82	-2.94	0.00	-3.77	-0.30	-1.35	0.00	-1.65
TYR 105	--	--	--	--	-0.09	-2.46	0.00	-2.55
PHE 78	-3.66	-1.37	0.00	-5.03	--	--	--	--
ASP 101	-0.63	75.50	0.00	74.88	--	--	--	--

Table S3-20 – Per-residue energetic contributions in the predicted RPTPσ/CS-E (left) and RPTPσ/heparin (right) structures.

	CS-E				heparin			
Residue	**VdW**	**Coulomb**	**H Bond**	**NonBond**	**VdW**	**Coulomb**	**H Bond**	**NonBond**
ARG 77	11.74	-158.46	-17.21	-163.92	0.89	-180.82	-7.26	-187.18
LYS 68	10.98	-158.86	-10.24	-158.12	-0.60	-170.34	-1.68	-172.61
ARG 100	-0.66	-139.54	-9.43	-149.63	3.44	-208.70	-9.97	-215.23
ARG 97	1.84	-137.33	-5.15	-140.64	2.01	-180.78	-8.27	-187.04
LYS 71	-2.11	-131.74	0.00	-133.86	4.10	-223.34	-3.40	-222.65
LYS 69	3.46	-115.00	-5.67	-117.21	-1.69	-127.16	0.00	-128.85
LYS 72	-0.37	-65.91	0.00	-66.28	--	--	--	--
GLN 76	5.83	-20.07	-12.97	-27.22	-0.17	-5.26	0.00	-5.43
ASN 74	2.57	-15.72	-10.33	-23.48	-0.22	-1.46	0.00	-1.68
SER 75	-1.95	-10.28	-0.59	-12.83	-0.85	-7.68	-0.18	-8.71
ASN 103	-1.49	-9.18	0.00	-10.67	-0.26	-9.19	0.00	-9.45
GLU 102	--	--	--	--	--	--	--	--
GLY 70	--	--	--	--	-0.15	0.74	0.00	0.59
PRO 95	--	--	--	--	-0.54	4.62	0.00	4.08
PRO 99	--	--	--	--	-1.04	-2.44	0.00	-3.47
THR 98	--	--	--	--	-0.28	3.73	0.00	3.45
VAL 73	-2.38	-5.93	0.00	-8.31	--	--	--	--
TYR 105	3.55	-6.81	-4.76	-8.02	--	--	--	--
PHE 78	-3.37	1.18	0.00	-2.19	-1.31	-0.46	0.00	-1.77
ASP 101	-1.75	97.79	0.00	96.04	--	--	--	--

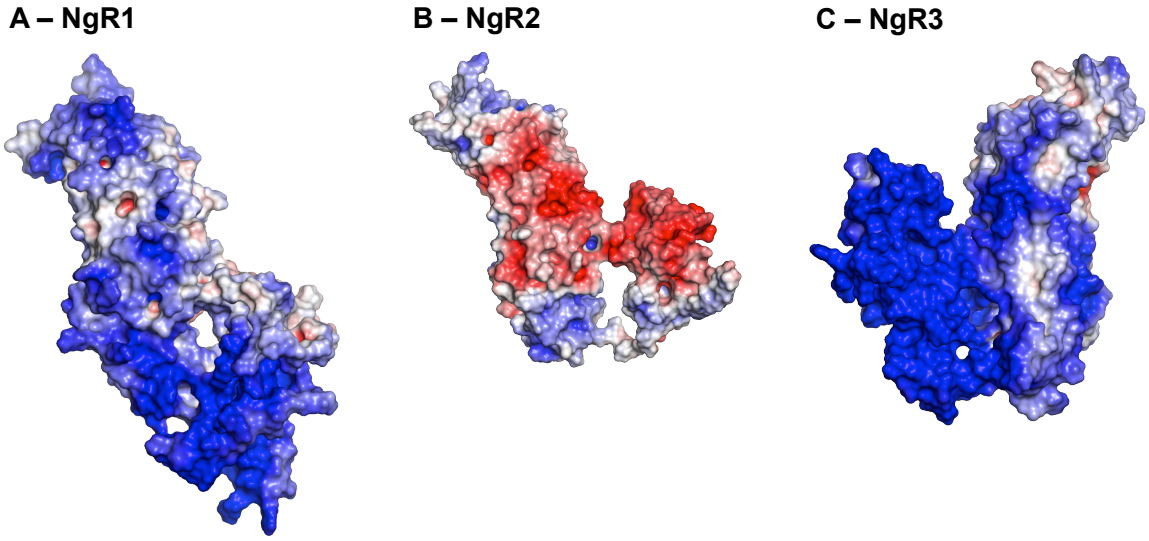

Figure S3-22 – Electrostatic potential surfaces of (A) NgR1, (B) NgR2, and (C) NgR3. Note the lack of positive charge on NgR2, but strong positive charge on NgR1 and NgR3.

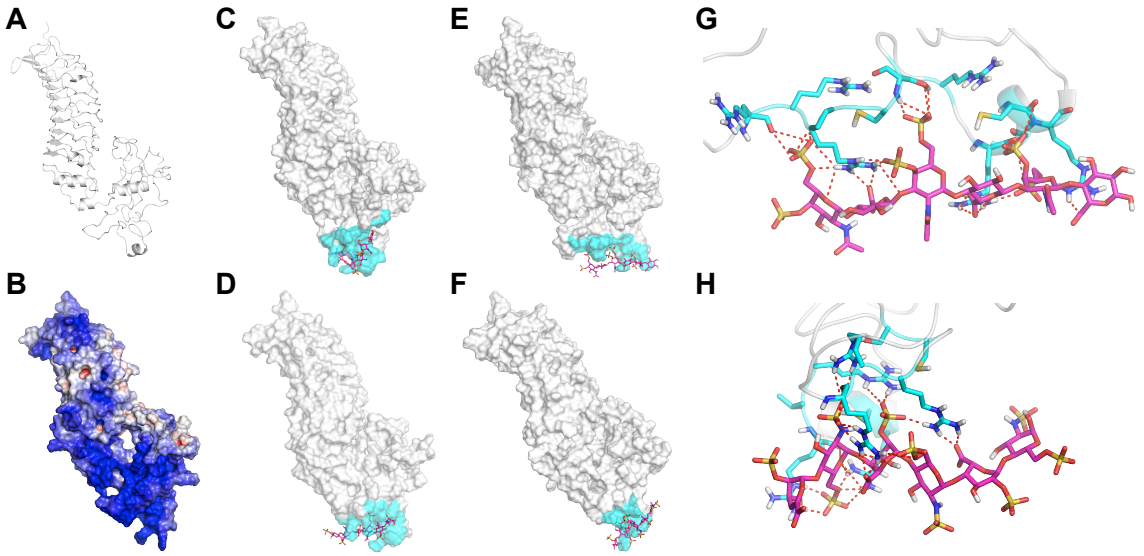

Figure S3-23 – NgR1. (A) Structure of NgR1. (B) Electrostatic potential surface showing strong positive charge. (C-F) Predicted structures of CS-A, CS-D, CS-E, and heparin after docking and molecular dynamics. (G-H) Detailed view of CS-E and heparin predicted structures.

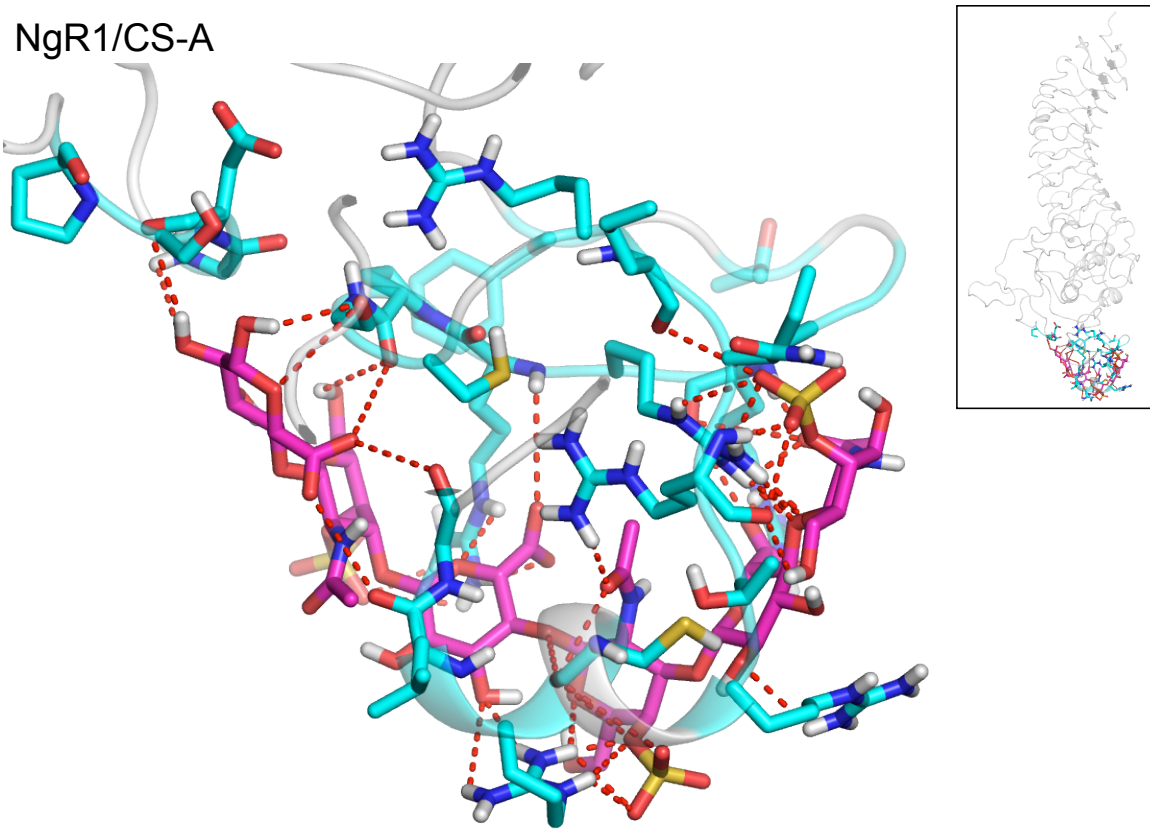

Figure S3-24 – Detail of predicted NgR1/CS-A structure after docking and dynamics with CS-A hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein. Overall placement on protein shown in inset.

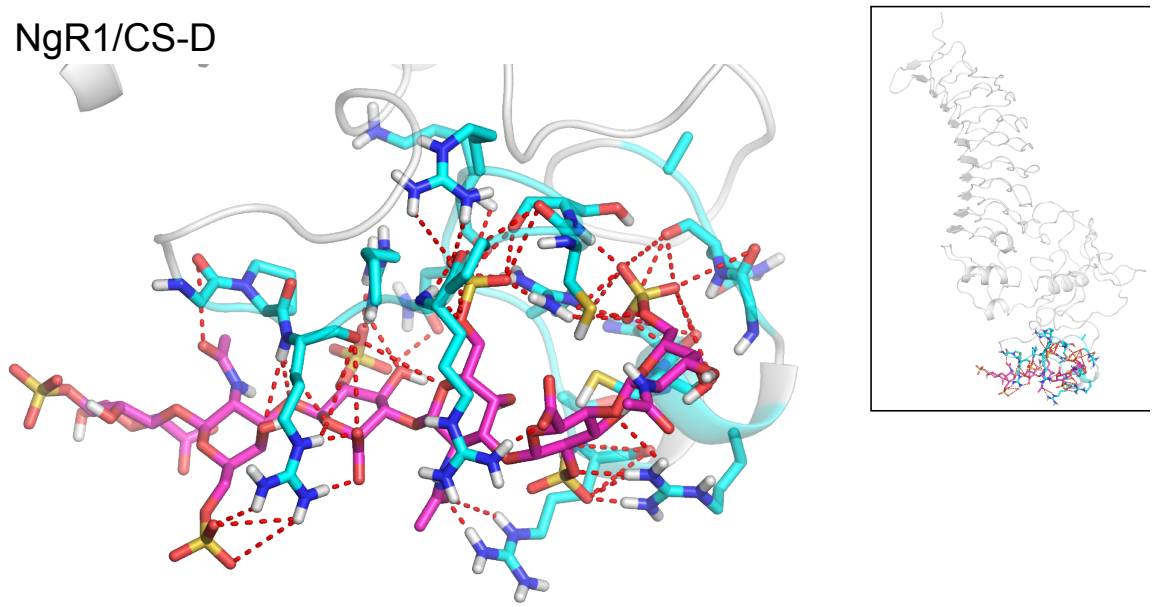

Figure S3-25 – Detail of predicted NgR1/CS-D structure after docking and dynamics with CS-D hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

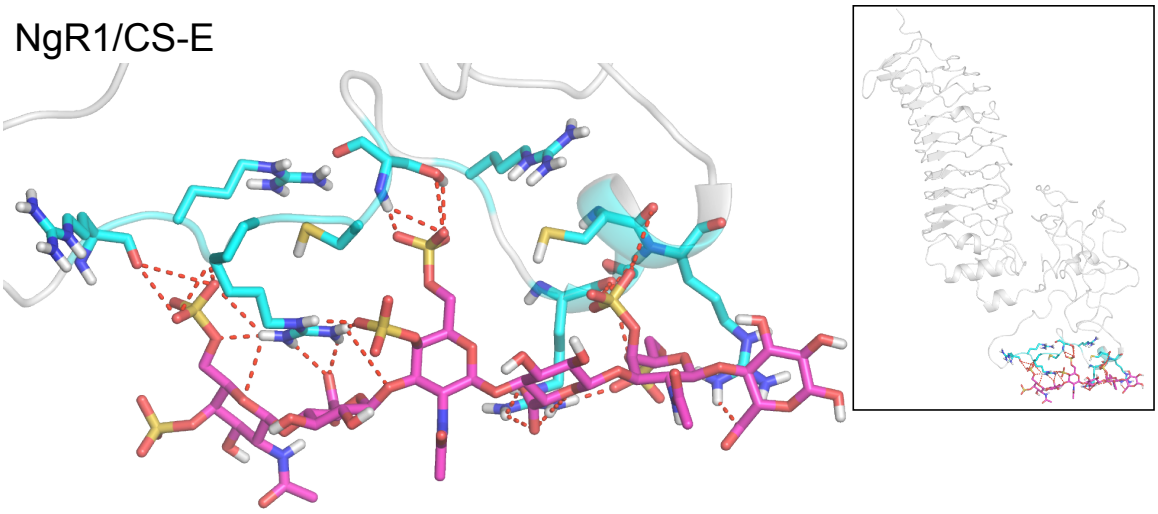

Figure S3-26 – Detail of predicted NgR1/CS-E structure after docking and dynamics with CS-E hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

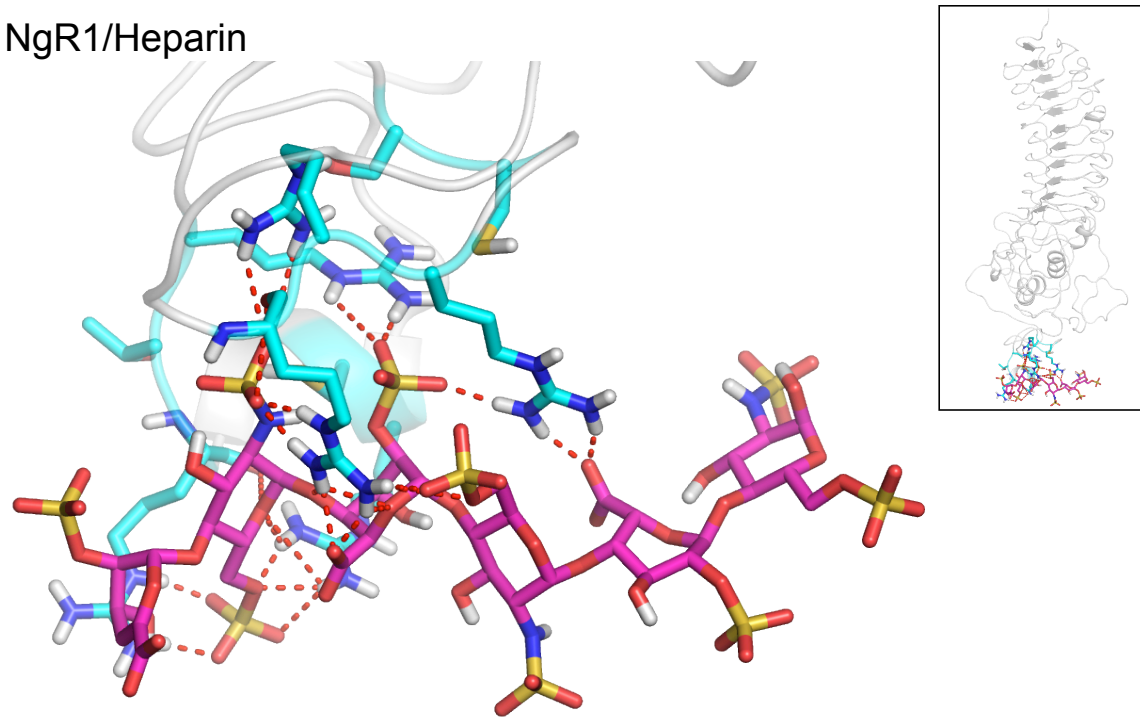

Figure S3-27 – Detail of predicted NgR1/heparin structure after docking and dynamics with heparin hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

Table S3-21 – Per-residue energetic contributions in the predicted NgR1 structures for CS-A, CS-D, CS-E, and heparin.

		CS-A				CS-D				CS-E				Heparin			
Residue		VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond
ARG	390	6.94	-90.88	-13.23	-97.17	2.15	-136.95	-11.48	-146.28	-1.11	-61.48	0.00	-62.59	5.12	-213.83	-16.07	-224.78
ARG	392	-2.73	-107.51	-5.79	-116.04	-2.26	-126.71	-8.96	-137.93	-0.19	-158.71	-14.92	-173.82	2.77	-191.60	-10.23	-199.06
ARG	406	-2.67	-96.20	-7.45	-106.33	4.79	-116.69	-8.19	-120.09	2.55	-131.51	-12.53	-141.50	-1.21	-176.87	-6.87	-184.95
ARG	402	-3.57	-69.97	0.00	-73.53	0.76	-120.34	-4.53	-124.10	3.02	-147.77	-12.85	-157.59	0.82	-151.54	-9.09	-159.81
ARG	400	-0.48	-96.72	-5.19	-102.39	8.97	-144.72	-15.76	-151.51	-0.73	-98.90	0.00	-99.63	4.04	-142.10	-5.53	-143.59
ARG	391	-2.06	-111.17	-8.33	-121.56	1.17	-142.38	-7.89	-149.09	-1.38	-105.45	0.00	-106.83	-1.06	-134.39	0.00	-135.44
ARG	397	-0.25	-56.79	0.00	-57.04	-1.74	-101.89	-0.03	-103.66	-0.15	-72.85	0.00	-73.00	-0.06	-96.30	0.00	-96.36
LYS	398	-0.40	-40.49	0.00	-40.89	-0.78	-81.71	0.00	-82.49	-0.14	-58.40	0.00	-58.54	-0.06	-68.64	0.00	-68.70
PRO	389	-4.08	-5.91	0.00	-9.99	-4.13	-4.22	0.00	-8.34	-0.33	-3.98	0.00	-4.31	-0.36	-4.12	0.00	-4.48
GLY	408	-1.82	-0.24	0.00	-2.06	-0.81	-5.51	0.00	-6.32	-0.09	-4.24	0.00	-4.33	-0.07	-4.17	0.00	-4.24
GLY	388	-0.62	-3.62	0.00	-4.24	-0.96	-6.68	0.00	-7.64	-0.04	-3.15	0.00	-3.19	-0.03	-4.10	0.00	-4.13
CYS	405	-1.71	0.31	-2.35	-3.75	-0.20	-5.66	-2.02	-7.88	-0.01	-8.79	0.00	-8.80	-3.89	-0.16	0.00	-4.05
ASN	399	-1.35	-5.54	0.00	-6.89	-1.93	1.27	0.00	-0.65	-0.19	-4.50	0.00	-4.70	-0.31	-3.64	0.00	-3.95
PHE	384	-0.69	-1.25	0.00	-1.94	-0.23	-2.07	0.00	-2.30	-0.03	-1.51	0.00	-1.55	-0.07	-2.90	0.00	-2.97
LEU	407	-0.98	0.71	0.00	-0.27	-0.25	-1.87	0.00	-2.12	-0.27	-1.86	0.00	-2.14	-0.07	-1.97	0.00	-2.04
GLY	394	1.03	-2.84	0.00	-1.81	-0.56	0.65	0.00	0.09	-0.50	0.68	0.00	0.18	-0.05	-1.33	0.00	-1.38
CYS	395	-2.02	-1.44	0.00	-3.46	-0.43	-9.15	0.00	-9.58	-2.79	-12.58	0.00	-15.37	-0.65	-0.26	0.00	-0.91
SER	344	-0.83	-0.23	0.00	-1.07	-0.11	-0.55	0.00	-0.66	-0.01	-1.48	0.00	-1.49	-0.02	-0.19	0.00	-0.21
THR	444	0.00	-0.53	0.00	-0.53	-0.01	1.78	0.00	1.78	0.00	-0.46	0.00	-0.46	0.00	0.64	0.00	0.64
ALA	410	-0.17	-0.76	0.00	-0.93	-0.24	0.61	0.00	0.37	-0.02	-0.05	0.00	-0.07	-0.03	1.06	0.00	1.03
THR	386	-0.20	0.59	0.00	0.39	-0.06	1.06	0.00	1.00	-0.01	1.22	0.00	1.21	-0.01	1.16	0.00	1.15
HSE	404	-0.48	1.27	0.00	0.79	-0.60	2.33	0.00	1.73	-0.29	0.25	0.00	-0.04	-0.23	1.87	0.00	1.63
GLN	409	-0.49	1.79	0.00	1.30	-1.17	9.18	0.00	8.01	-0.07	3.92	0.00	3.85	-0.07	2.02	0.00	1.95
PRO	345	-0.13	0.77	0.00	0.64	-0.04	1.85	0.00	1.82	0.00	1.29	0.00	1.29	-0.01	2.08	0.00	2.08
GLY	342	-0.99	-1.11	0.00	-2.10	-0.04	2.58	0.00	2.55	-0.01	1.93	0.00	1.93	-0.02	2.20	0.00	2.18
THR	401	-0.66	2.74	0.00	2.08	-0.46	4.98	0.00	4.52	-0.37	3.92	0.00	3.55	-0.72	3.17	0.00	2.45
SER	396	-0.18	3.43	0.00	3.25	0.59	5.47	-3.85	2.21	3.30	-2.21	-4.89	-3.80	-0.18	2.69	0.00	2.51
PRO	393	-3.03	4.64	0.00	1.61	-0.79	7.05	0.00	6.26	-0.40	7.33	0.00	6.92	-0.24	5.85	0.00	5.61
SER	403	-0.18	3.43	0.00	3.25	-0.19	2.81	0.00	2.61	-1.02	1.80	0.00	0.78	-0.35	6.34	0.00	5.99
ASP	343	-1.62	48.13	0.00	46.51	-0.07	71.82	0.00	71.74	-0.01	61.99	0.00	61.98	-0.02	82.58	0.00	82.56

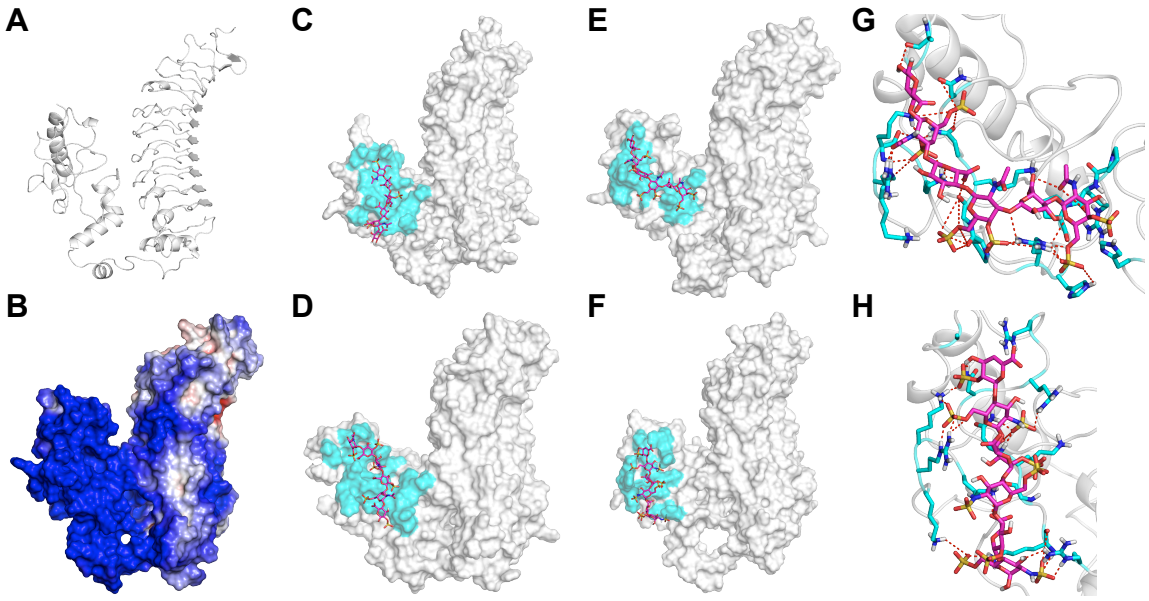

Figure S3-28 – NgR3. (A) Structure of NgR3. (B) Electrostatic potential surface. (C-F) Predicted structures of CS-A, CS-D, CS-E, and heparin after docking and molecular dynamics. (G-H) Detailed view of CS-E and heparin predicted structures.

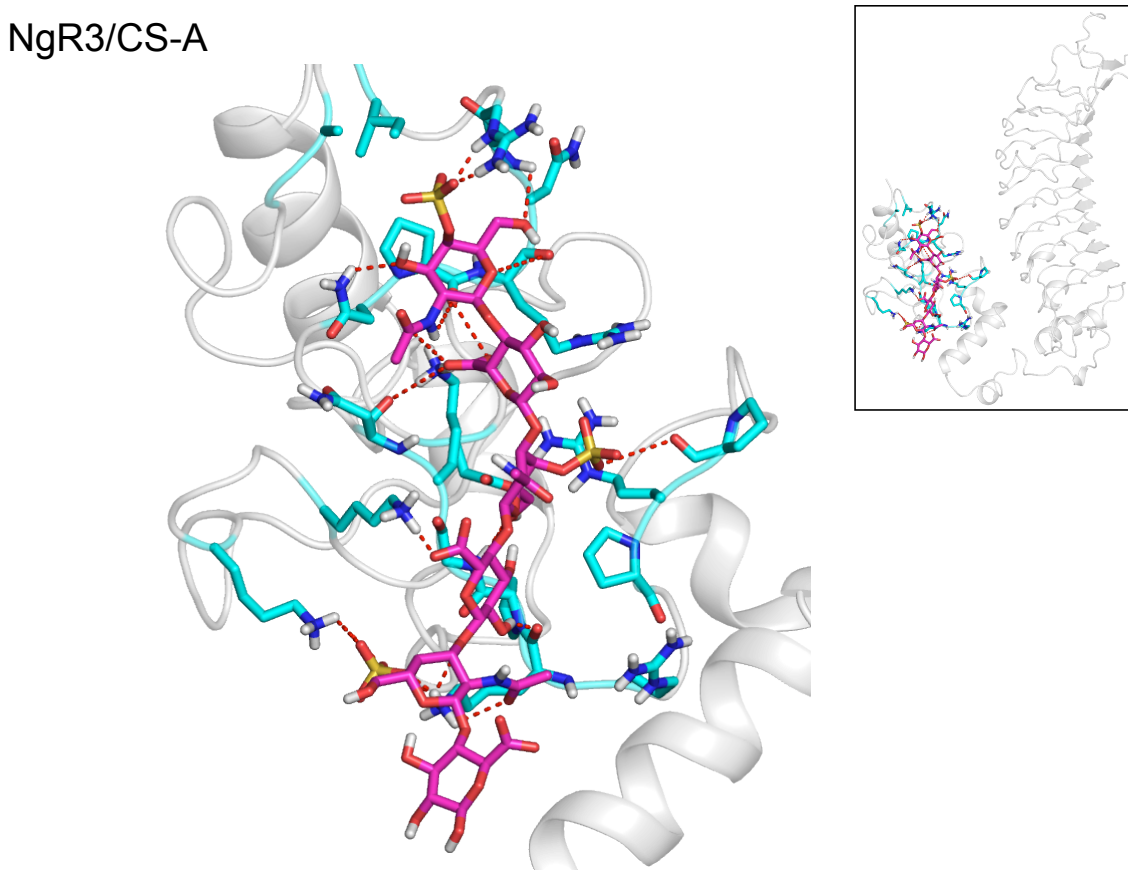

Figure S3-29 – Detail of predicted NgR3/CS-A structure after docking and dynamics with CS-A hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

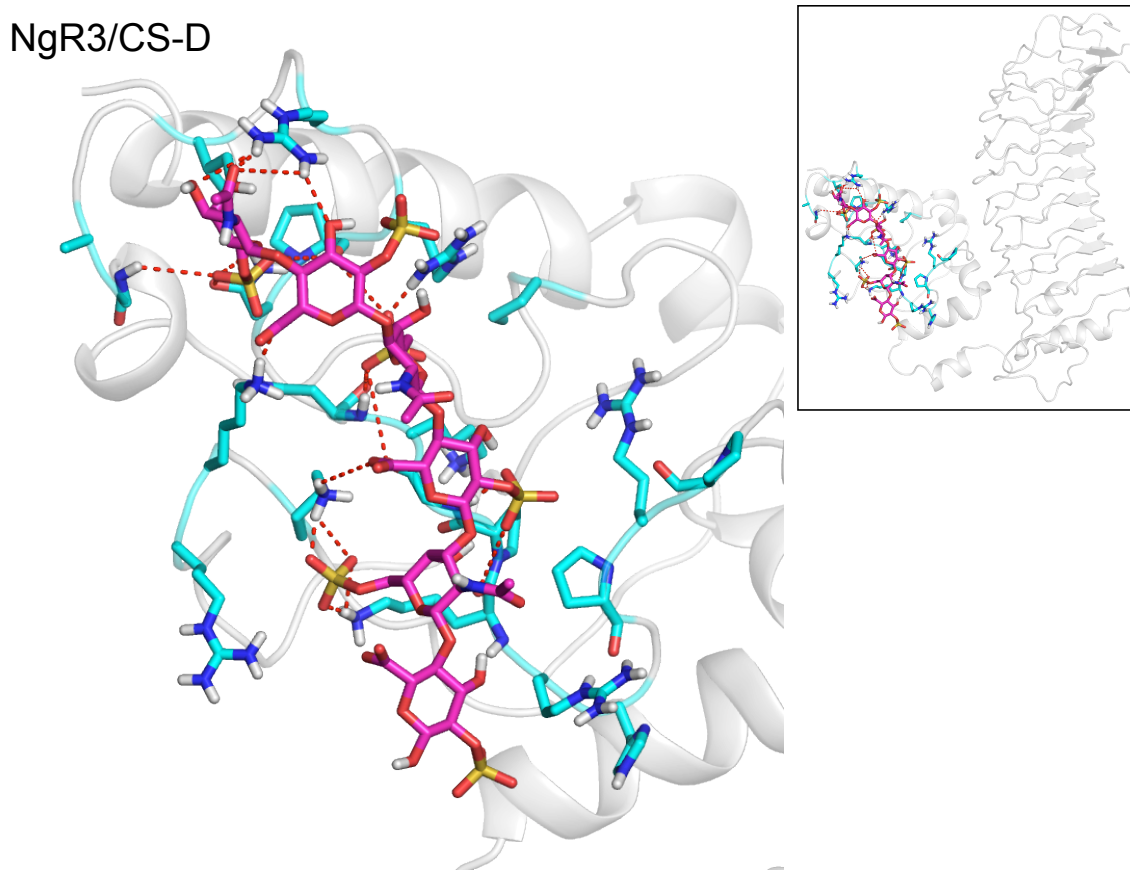

Figure S3-30 – Detail of predicted NgR3/CS-D structure after docking and dynamics with CS-D hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

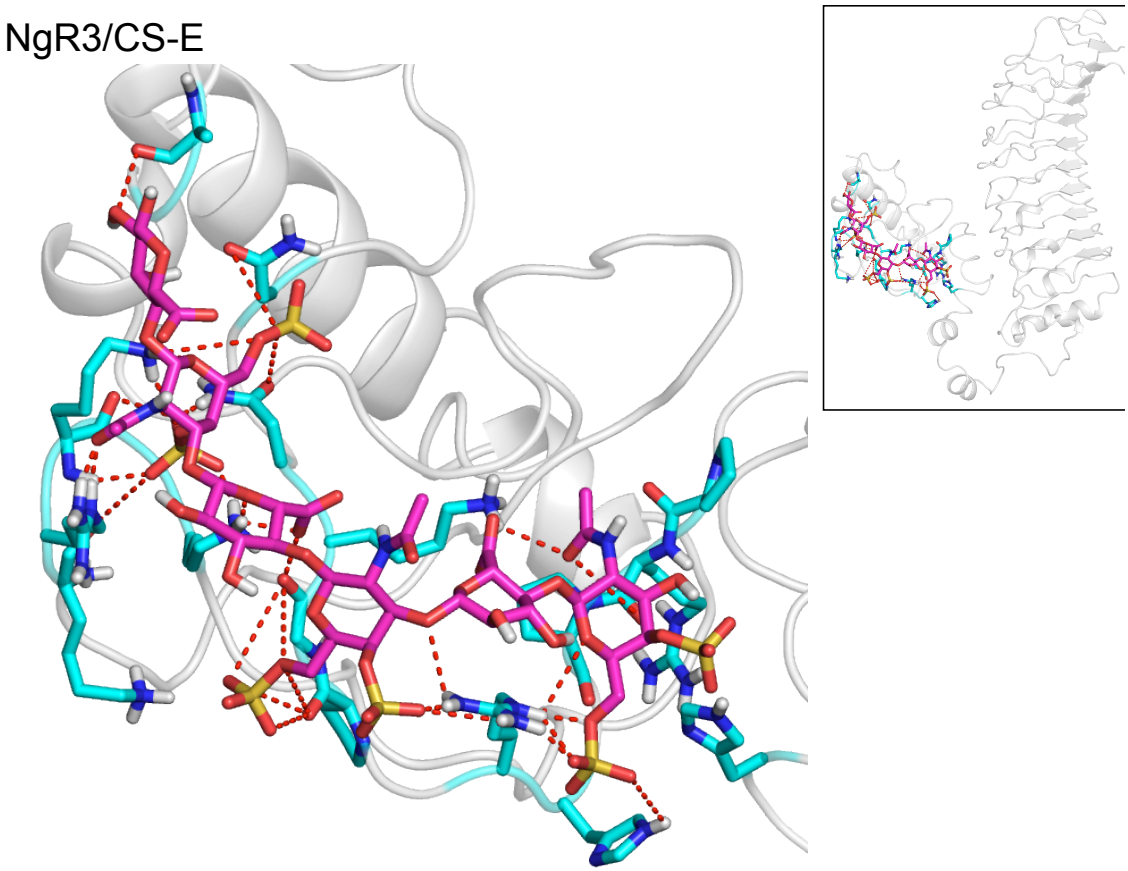

Figure S3-31 – Detail of predicted NgR3/CS-E structure after docking and dynamics with CS-E hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

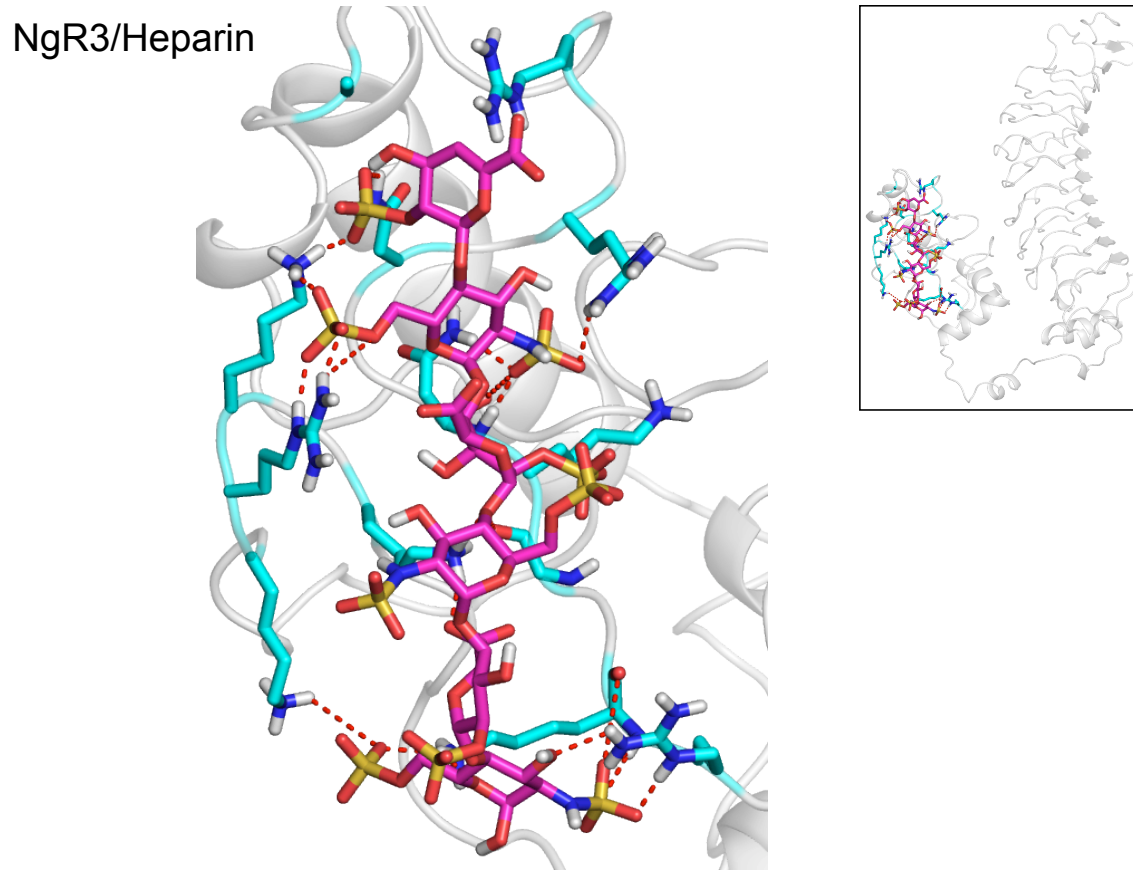

Figure S3-32 – Detail of predicted NgR3/heparin structure after docking and dynamics with heparin hexamer (magenta) and 5 Å binding site (cyan) shown. Dashed lines indicate hydrogen bonding and salt bridges between ligand and protein.

Table S3-22 – Per-residue energetic contributions in the predicted NgR3 structures for CS-A, CS-D, CS-E, and heparin.

		CS-A				CS-D				CS-E				Heparin			
Residue		VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond	VdW	Coulomb	H Bond	NonBond
LYS	383	4.11	-102.12	-2.70	-100.71	0.45	-152.91	-4.15	-156.61	6.52	-162.33	-9.44	-165.24	7.58	-206.62	-11.23	-210.27
ARG	380	-0.04	-40.12	0.00	-40.16	-1.06	-102.20	0.00	-103.26	-1.20	-119.85	-5.80	-126.85	3.48	-182.45	-7.59	-186.56
LYS	381	-0.17	-52.67	0.00	-52.84	0.75	-138.39	-4.12	-141.76	4.02	-139.02	-10.71	-145.71	7.81	-180.38	-6.68	-179.25
ARG	330	-0.84	-63.93	0.00	-64.77	-2.80	-102.67	0.00	-105.47	-0.31	-153.39	-11.01	-164.70	2.79	-158.26	-5.44	-160.90
ARG	340	-2.65	-68.57	0.00	-71.22	1.84	-118.19	-5.24	-121.59	-0.29	-83.03	0.00	-83.32	-1.34	-156.72	-2.65	-160.71
LYS	331	-0.27	-98.70	-4.75	-103.72	3.13	-123.36	-5.79	-126.02	-0.63	-74.84	0.00	-75.48	-0.71	-153.93	-0.76	-155.40
LYS	334	-1.13	-89.69	-4.68	-95.50	-2.39	-97.83	0.00	-100.22	-2.53	-141.48	-1.46	-145.47	-3.30	-134.35	0.00	-137.65
LYS	379	-0.84	-81.40	-0.02	-82.25	-0.17	-67.58	0.00	-67.74	-0.37	-84.61	0.00	-84.99	-1.05	-130.60	0.00	-131.66
ARG	342	2.01	-72.80	-6.37	-77.17	-2.08	-104.32	-3.28	-109.68	-0.09	-68.84	0.00	-68.93	-0.81	-118.99	0.00	-119.80
ARG	326	-1.18	-62.89	0.00	-64.07	-0.84	-92.99	0.00	-93.83	-2.41	-72.84	0.00	-75.25	-0.08	-97.25	0.00	-97.33
ASN	335	-1.84	0.07	0.00	-1.77	2.32	-9.47	-4.60	-11.75	-1.99	-9.78	-3.64	-15.41	-0.19	-12.30	-6.78	-19.27
ASN	338	-1.39	-2.82	-0.16	-4.37	-1.61	-2.72	-0.08	-4.41	-1.12	-2.24	0.00	-3.37	-1.17	-3.49	0.00	-4.66
PRO	362	-0.25	-0.75	0.00	-1.00	-0.20	-2.05	0.00	-2.25	-0.02	-1.54	0.00	-1.57	-0.02	-2.42	0.00	-2.44
PRO	327	-0.88	-3.14	0.00	-4.01	-0.87	-3.83	0.00	-4.70	-3.57	0.38	0.00	-3.18	-0.22	-1.95	0.00	-2.17
GLY	333	0.34	-0.67	-4.29	-4.63	-1.59	-1.25	-3.92	-6.77	-2.38	4.27	0.00	1.89	-1.51	-0.23	0.00	-1.74
GLY	382	-0.09	0.75	0.00	0.66	-0.12	2.17	0.00	2.05	-1.28	1.34	0.00	0.06	-0.25	-1.36	0.00	-1.61
ILE	345	-0.85	-0.11	0.00	-0.96	-1.88	-0.42	0.00	-2.30	-0.12	-0.55	0.00	-0.67	-0.08	-0.78	0.00	-0.86
ALA	350	-0.30	-2.38	0.00	-2.68	-0.12	-1.27	0.00	-1.39	-0.13	-0.71	0.00	-0.84	-0.02	-0.79	0.00	-0.81
ASN	341	-1.36	0.73	0.00	-0.63	-0.15	2.42	0.00	2.27	-0.04	0.79	0.00	0.75	-0.06	0.00	0.00	-0.07
GLY	349	-0.16	-2.75	0.00	-2.91	-0.45	-4.02	0.00	-4.47	-1.21	-2.14	0.00	-3.35	-0.06	0.43	0.00	0.37
PRO	332	-2.68	-0.11	0.00	-2.79	-1.95	-2.37	0.00	-4.32	-0.92	4.48	0.00	3.55	-0.59	1.27	0.00	0.68
PRO	339	-4.19	5.10	0.00	0.92	-0.55	8.20	0.00	7.65	-0.28	5.86	0.00	5.59	-0.26	1.69	0.00	1.43
ALA	348	-0.46	-3.28	0.00	-3.74	-0.29	-2.64	0.00	-2.93	-1.34	-2.64	0.00	-3.97	-0.44	2.12	0.00	1.69
HSE	309	-0.02	-0.37	0.00	-0.40	-0.02	0.60	0.00	0.58	-0.65	-1.57	0.00	-2.21	-0.02	1.89	0.00	1.87
HSE	329	-0.06	0.65	0.00	0.58	-0.27	0.00	0.00	-0.26	-1.25	-1.38	0.00	-2.63	-0.26	3.43	0.00	3.18
PRO	325	-0.44	3.13	0.00	2.69	-0.19	3.21	0.00	3.03	-2.53	-1.17	0.00	-3.70	-0.03	3.36	0.00	3.33
ASP	359	-0.30	64.07	0.00	63.78	-0.04	79.36	0.00	79.32	-0.03	69.11	0.00	69.09	-0.08	100.87	0.00	100.79

Table S3-23 – Single residue mutation data for RPTPs. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen bonding + van der Waals or Coulomb energy.

PTPS Increased Binding																		
PTPS Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
73	V	N	-1.48	0.00	-6.17	0.00	1.56	-6.53	8.09	5.03	7.96	-2.93	-28.66	3.08	-31.75	-1.41	-2.47	1.07
73	V	Q	-1.82	0.00	0.00	0.00	-2.36	-5.15	2.78	-3.85	5.70	-9.56	-32.20	-6.47	-25.73	-5.87	-0.20	-5.66
74	N	Q	0.00	-4.29	5.91	0.00	-11.00	-3.44	-7.56	13.75	1.60	12.14	14.66	10.45	4.21	2.97	-3.06	6.03
75	S	N	-1.22	-4.66	-1.17	0.31	-7.17	7.22	-14.40	-35.76	-4.34	-31.41	11.00	1.23	9.78	24.41	0.73	23.69
75	S	Q	4.75	-1.11	4.44	0.31	26.31	29.85	-3.54	-28.87	-6.92	-21.96	21.53	24.46	-2.94	21.00	2.33	18.68
76	Q	N	-1.61	1.19	-1.50	0.00	-9.37	-2.33	-7.04	0.44	-0.59	1.02	2.28	5.24	-2.97	6.07	-1.82	7.89
78	F	N	-4.58	0.00	0.00	0.00	3.34	0.46	2.89	5.48	5.98	-0.50	-3.23	3.50	-6.73	6.58	1.98	4.60
78	F	Q	0.00	0.00	-3.07	0.00	-11.05	-4.91	-6.15	-8.99	4.79	-13.79	-6.42	6.54	-12.97	2.47	-2.04	4.51
103	N	Q	-1.14	0.00	5.15	-0.04	-13.39	-0.59	-12.80	-22.49	-8.02	-14.47	-6.56	0.38	-6.94	-6.77	-3.88	-2.88

PTPS Decreased Binding																		
PTPS Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
68	K	N	7.58	0.00	5.88	9.87	95.67	8.45	87.22	79.75	-5.99	85.73	119.99	-3.28	123.26	137.68	1.17	136.51
68	K	Q	7.60	0.00	0.43	9.87	103.11	7.80	95.31	77.89	-4.12	82.01	116.16	-0.89	117.05	160.96	3.13	157.83
69	K	N	8.27	0.00	-0.43	0.41	109.55	12.56	96.99	87.95	5.61	82.35	84.90	-3.08	87.97	123.59	-1.29	124.88
69	K	Q	8.27	0.00	5.12	0.41	97.91	5.28	92.63	75.40	4.98	70.41	90.73	-5.27	95.99	135.29	-6.80	142.10
71	K	N	4.58	4.29	1.68	5.71	78.37	4.92	73.44	84.49	-3.90	88.39	74.38	-1.00	75.38	172.64	-0.70	173.34
71	K	Q	6.15	-2.56	0.20	11.24	89.01	13.20	75.81	74.24	-11.41	85.65	72.17	-1.89	74.06	185.03	13.95	171.08
77	R	Q	6.96	6.37	6.84	14.14	100.48	5.10	95.38	128.38	3.40	124.98	107.04	-9.98	117.03	163.36	0.66	162.70
77	R	N	7.01	0.96	11.82	14.14	94.89	2.33	92.57	130.30	-1.43	131.72	108.53	-10.16	118.69	165.41	0.61	164.81
97	R	N	4.59	9.78	-0.96	12.05	100.76	0.66	100.09	123.99	1.22	122.76	121.08	11.91	109.17	189.16	-0.56	189.72
97	R	Q	4.59	9.78	-0.05	6.35	91.62	-0.55	92.18	116.85	2.75	114.09	114.11	1.33	112.78	188.49	0.66	187.84
100	R	N	12.01	18.70	3.81	15.24	98.76	2.89	95.88	150.09	5.51	144.58	139.31	0.52	138.79	219.62	-1.39	221.02
100	R	Q	12.01	23.01	0.71	15.53	91.91	2.23	89.69	132.35	2.97	129.37	112.37	3.65	108.72	221.74	1.56	220.19

Table S3-24 – Single residue mutation data for NgR1. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen bonding + van der Waals or Coulomb energy.

NGR1 Increased Binding																		
NGR1 Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
395	C	Q	-5.33	0.00	-4.87	0.00	-7.24	2.09	-9.33	3.60	-1.74	5.34	8.95	0.38	8.57	6.30	-3.63	9.93
396	S	N	0.00	-1.73	-6.06	0.00	-3.19	3.57	-6.76	-15.44	-3.64	-11.80	9.88	3.95	5.93	10.94	-3.36	14.31
396	S	Q	0.00	4.38	-0.05	-4.85	-6.51	3.69	-10.21	-0.97	1.83	-2.81	15.97	5.70	10.27	-13.35	-5.15	-8.19
399	N	Q	4.67	-3.26	0.00	0.00	2.86	2.29	0.58	-7.57	-2.71	-4.86	2.18	-0.67	2.85	-47.06	-1.26	-45.80
403	S	Q	0.00	0.00	-5.22	0.00	4.39	4.40	0.00	-3.76	-2.31	-1.45	5.82	2.99	2.84	-1.31	-4.93	3.62
405	C	N	-6.52	-4.36	-0.79	-3.27	-8.64	5.99	-14.62	-24.39	-5.79	-18.61	-0.64	-0.16	-0.47	-0.03	-7.67	7.65
405	C	Q	0.00	-1.60	-6.10	0.00	-4.76	8.05	-12.81	-4.87	5.03	-9.90	-3.39	-0.36	-3.04	-0.92	-5.56	4.64

NGR1 Decreased Binding																		
NGR1 Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
390	R	Q	0.86	8.12	0.00	14.17	83.82	3.55	80.28	127.85	4.00	123.84	74.90	-2.34	77.24	216.92	-6.24	223.17
390	R	N	3.65	7.85	0.00	20.17	91.45	4.43	87.03	129.28	7.46	121.83	103.46	1.07	102.40	214.66	5.94	208.73
391	R	N	12.39	1.33	4.80	0.00	111.02	12.01	99.03	117.11	-3.64	120.74	133.74	3.22	130.51	104.52	-5.12	109.65
391	R	Q	12.39	5.69	4.80	0.00	81.87	1.91	79.96	108.81	6.44	102.37	130.43	2.71	127.72	116.57	-6.14	122.71
392	R	N	5.47	-0.86	10.62	6.61	87.18	6.33	80.86	120.25	11.02	109.24	165.18	11.93	153.25	176.91	4.80	172.12
392	R	Q	5.93	-0.64	10.64	7.72	88.87	5.06	83.81	111.70	3.80	107.91	165.48	10.73	154.75	175.74	3.43	172.31
400	R	N	0.88	15.38	0.00	5.59	75.35	3.35	72.00	140.99	10.79	130.20	95.87	-1.00	96.87	141.97	-4.20	146.17
400	R	Q	0.95	11.95	0.00	5.57	88.95	9.99	78.97	135.65	8.85	126.80	93.43	-0.12	93.56	108.39	-2.58	110.97
402	R	Q	7.41	5.08	12.90	4.13	78.45	8.42	70.03	107.96	-4.30	112.26	147.99	4.21	143.78	142.76	-4.02	146.78
402	R	N	7.47	6.67	0.70	2.92	84.77	9.26	75.52	110.69	-4.12	114.81	144.72	3.64	141.08	149.66	-3.21	152.87
406	R	Q	6.16	2.32	13.46	5.97	91.65	9.39	82.27	86.91	-5.04	91.95	136.94	2.12	134.82	167.79	2.68	165.11
406	R	N	7.34	4.44	12.49	5.97	103.23	11.66	91.57	97.02	3.78	93.24	135.69	1.72	133.98	181.99	-1.47	183.46

Table S3-25 – Single residue mutation data for NgR3. Values show change in binding energy (kcal/mol) relative to wildtype structures. Values are shown both for the change in hydrogen bonding for the specific mutated residue as well as the overall change in the full cavity binding energy. The cavity binding energy is further separated into hydrogen bonding + van der Waals or Coulomb energy.

NGR3 Increased Binding																		
NGR3 Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
338	N	Q	1.43	0.06	3.53	-5.31	-4.84	2.39	-7.23	6.70	0.74	5.96	18.28	-2.09	20.37	0.10	1.74	-1.64
345	I	Q	-4.24	-5.59	0.00	0.00	-3.12	-0.93	-2.20	0.80	-3.97	4.78	-10.18	-7.76	-2.41	-0.10	-0.19	0.08
348	A	N	-4.60	0.00	-0.78	-2.15	-12.91	-2.87	-10.04	-5.07	-2.65	-2.42	-5.76	-1.24	-4.51	-12.19	-3.71	-8.48

NGR3 Decreased Binding																		
NGR3 Residue			Per-Res ΔHBond				CSA Cavity			CSD Cavity			CSE Cavity			HEP Cavity		
num	from	to	CSA	CSD	CSE	HEP	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou	ΔCav	Δ(HB+VDW)	ΔCou
330	R	N	0.74	-1.09	9.03	0.08	84.09	9.33	74.77	49.41	-4.51	53.92	149.89	2.24	147.65	125.52	-3.79	129.32
330	R	Q	0.74	-2.10	7.54	-0.17	84.81	10.32	74.49	28.03	-5.77	33.80	152.89	2.27	150.63	127.01	-5.33	132.33
331	K	Q	4.31	5.31	1.87	-2.08	92.34	0.99	91.35	25.97	-7.63	33.60	105.28	6.51	98.78	155.28	-2.14	157.42
331	K	N	7.53	5.75	1.87	-6.82	101.31	7.32	94.00	31.23	-5.72	36.95	100.70	6.15	94.55	148.88	-1.89	150.77
334	K	N	2.01	-6.09	9.00	0.00	70.93	-2.08	73.02	70.93	-6.27	77.20	154.48	-2.05	156.53	119.55	-2.40	121.96
334	K	Q	6.57	0.00	3.57	0.00	53.70	1.07	52.63	93.72	-3.94	97.67	102.98	-2.46	105.44	130.15	-0.69	130.84
340	R	N	0.00	-1.29	0.00	5.17	64.44	1.69	62.74	117.17	-4.79	121.96	73.59	-5.51	79.09	112.55	-10.32	122.87
340	R	Q	0.00	-1.28	0.00	5.17	64.45	0.34	64.11	72.83	-7.89	80.72	77.98	-11.87	89.85	120.72	-7.44	128.16
342	R	N	8.86	16.64	5.66	0.00	71.30	4.52	66.77	128.14	0.11	128.03	95.01	-5.88	100.89	108.88	-8.34	117.22
342	R	Q	8.86	16.07	5.66	0.00	63.84	3.86	59.98	126.70	-0.15	126.85	92.95	-4.70	97.65	108.36	-6.29	114.65
379	K	N	0.00	0.00	0.00	5.58	3.64	-4.34	7.99	76.86	-2.33	79.19	82.12	-1.28	83.39	158.63	2.23	156.40
379	K	Q	0.00	0.00	0.00	5.58	11.27	2.03	9.24	72.12	-1.96	74.08	84.91	-0.05	84.96	158.55	-1.94	160.48
380	R	N	0.00	5.43	0.59	3.93	42.61	-0.06	42.68	51.23	0.39	50.85	66.95	-3.91	70.86	188.74	-3.93	192.67
380	R	Q	0.00	5.43	2.81	1.21	35.33	0.26	35.08	45.62	-0.61	46.23	69.83	-10.08	79.91	196.39	-3.00	199.40
381	K	N	0.00	4.69	8.58	1.68	56.14	2.87	53.27	84.96	-2.92	87.89	119.97	-5.89	125.86	146.31	-14.85	161.17
381	K	Q	0.00	4.69	8.27	-7.65	47.50	0.85	46.65	110.48	-0.09	110.57	120.14	-4.00	124.14	137.24	-17.81	155.05
383	K	Q	-5.14	6.10	2.09	6.77	42.26	2.69	39.58	63.75	-7.34	71.09	95.05	-3.91	98.96	154.07	-9.80	163.87
383	K	N	0.00	6.10	10.92	7.95	42.16	1.41	40.75	60.25	-2.41	62.66	92.99	-2.73	95.71	176.79	-5.68	182.47

Chapter 4

EXPLORING NOVEL INTERACTIONS BETWEEN CHONDROITIN SULFATE AND THE EPHB3 RECEPTOR

Adam R. Griffith,[a,b,1] Claude J. Rogers,[b,c,1] Ravinder Abrol,[a,b] William A. Goddard III[a,ab,*]
Greg Miller,[b,c] and Linda C. Hsieh-Wilson,[b,c]
[a]Materials and Process Simulation Center,
California Institute of Technology, Pasadena, CA 91125
[b]Division of Chemistry and Chemical Engineering
California Institute of Technology, Pasadena, CA 91125,
[c]Howard Hughes Medical Institute,
[1] These two authors contributed equally
[*] To whom correspondence should be addressed; email wag@wag.caltech.edu

Introduction

Recent work in the Hsieh-Wilson group has identified chondroitin sulfate-E (CS-E) binds to the EphB3 receptor with high specificity and physiologically relevant affinity and is responsible for the direction of retinal neuron growth.[1] However, CS-A, which is less sulfated than CS-E, does not bind to EphB3, nor does it direct neuronal growth. Furthermore, while EphB2 shares ~60% sequence identity (extracellular region) with EphB3, experiments show that it does not bind any glycosaminoglycans (GAGs). We have previously applied computational methods (GAG-Dock) to similar GAG-protein systems with great success, both for systems with known crystal structures and for identifying novel interactions with the protein tyrosine phosphatase σ (RPTPs) and Nogo receptors (NgR).[2] We believe that predictions of the interaction between CS-E and EphB3 will be useful in studying and understanding the role of this interaction in neuron growth.

Using the GAG-Dock method, we explain the differential binding of CS ligands to the EphB3 and EphB2 receptors. Our results identify the previously unknown binding site for

CS-E on the EphB3 receptor and suggest experiments that can be used to validate our predictions.

GAG-Dock Overview

GAG-Dock[2] is a docking method based on the DarwinDock[3-11] and GenDock methodology that has been accommodated to work with large, highly charged, surface-binding ligands characteristic of GAGs. Because the binding sites for proteins that bind GAGs are typically not known, it is necessary to sample the entire surface of the protein. The surface of the protein is broken into regions, which are then evaluated using a "coarse" level of docking, which generates 10,000 ligand poses for each region. Based on the ranking of these regions by energy, a subset is docked to using a "fine" level of docking. The "fine" docking is carried out to a completeness threshold of 5%; however, due to the computational difficulty of these systems, a limit of 50,000 ligand poses is placed on the completeness.

Modifications to GAG-Dock

GAG-Dock is used almost identically to the way that it was used in our work on RPTPs and NgR. The key difference has to do with the way in which regions of the protein were sampled. The extracellular domains of EphB3 and EphB2 are very large and the location of the CS binding site was not previously known. It was therefore necessary to sample the majority of the protein surface. As in our previous work, spheres were generated that cover the entire surface of the protein. These spheres were divided into overlapping boxes/regions, however at a smaller size: 15Å/side (instead of 20Å) with 3Å overlap

(instead of 5Å). This was done to reduce the computational cost of working with an octasaccharide ligand. However, this resulted in an excessive number of boxes to test. Knowing that the CS binding site must be positively charged to match the negative charge of the ligand, we used electrostatics to eliminate most of the sphere regions. Specifically, we calculated the electrostatic potential for the proteins (Fig. 4-1A) using the Adaptive Poisson-Boltzmann Solver (APBS[12-14]) method and mapped the potential onto the spheres. Regions were ordered based on the number of positively charged spheres, and the 25% with the largest number of positive spheres were kept for docking. This resulted in 45 regions (238-1108 positive spheres) for EphB3 and 47 regions (180-1211 positive spheres) for EphB2. (Fig. S4-11)

All other parts of the GAG-Dock procedure were the same. Because CS ligands of sufficient size had already been prepared for our prior work, we used the same CS-A, CS-D, and CS-E octasaccharides.

"Coarse" docking was applied using CS-A, CS-D, and CS-E to the 45 EphB3 regions and 47 EphB2 regions. The top 13 EphB3 regions for CS-E binding were reexamined using "fine" docking for CS-A, CS-D, and CS-E.

EphB2 and EphB3 Models

Because no crystal structures of the full EphB3 or EphB2 extracellular regions exist, it was necessary to use homology modeling to generate the protein structures. EphB3 (PDB: 3P1I[15]) and EphB2 (PDB: 2QBX[16]) ephrin ligand binding domain crystal structures were

used with a crystal structure of the EphA2 ectodomain (PDB: 2X11[17]) to generate the homology models for EphB3 and EphB2 using SWISS-MODEL[18-21].

The human EphB2 model was constructed for the sequence corresponding to protein residues 20-529 by using the 2.3 Å resolution structure for human EphB2 (PDB: 2QBX) for protein residues 20-194 and combining it with a homology structure for residues 195-529 based on a lower resolution (4.3 Å) human Ephrin type-A receptor 2 (EphA2) structure (PDB: 2X11). This required aligning 2QBX structure to the full 2X11 homology structure and extracting residues 195-529 to attach to 2QBX structure. This was followed by minimizing hinge residues 192-197 using the DREIDING[22] force field in MPSIM[23] while keeping all other residues fixed and then minimizing all the residues.

The human EphB3 model was constructed for the sequence corresponding to protein residues 39-544 by using the 2.1 Å resolution structure for human EphB3 (PDB: 3P1I) for protein residues 39-209 and combining it with a homology structure for residues 210-544 based on human Ephrin type-A receptor 2 (EphA2) structure (PDB: 2X11). This required aligning 3P1I structure to the full 2X11 homology structure and extracting residues 210-544 to attach to 3P1I structure. This was followed by minimizing hinge residues 207-212 using the DREIDING force field in MPSim, while keeping all other residues fixed and then minimizing all the residues. A schematic of the domains present in our EphB2 and EphB3 models is shown in Fig. 4-1A.

Results

We observed significant differences in the amount and placement of positive charge on the electrostatic potential surfaces of the EphB2 and EphB3 models (Fig. 4-1B, S4-11). Since CS-E is a highly sulfated GAG, we expected this to provide a structural basis for the selectivity of CS-E toward EphB3. This was verified by the GAG-Dock predictions for the CS-E octa-saccharide bound to each of the two proteins. We found from coarse docking that CS-E bound to EphB3 (–345 kcal/mol) more strongly than to EphB2 (–119 kcal/mol). We also docked two other CS octa-saccharides, CS-A and CS-D, to EphB3 and EphB2. The binding energies from coarse binding for these ligands also indicated better binding to EphB3 than to EphB2 (Fig. 4-2).

Comparisons of the binding energies from fine docking of the three CS octasaccharides (Fig. 4-3) showed that CS-E bound strongly to EphB3 (–381 kcal/mol) while CS-A did not (–280 kcal/mol). This is in agreement with experimental results for CS-E and CS-A binding to EphB3 found by the Hsieh-Wilson group. In our calculations CS-D (–374 kcal/mol) bound comparably to CS-E; however, there are no experimental results for CS-D binding as it is difficult to obtain pure molecules of CS-D for ligand binding experiments.

Overall GAG-Dock predicts binding sites and energies that correspond well with the known experimental data for CS binding to EphB2 and EphB3. The predicted CS-E binding region on EphB3 contains eight arginines (R309, R344, R363, R391, R408, R420, R440, and R478) as well as two lysines (K378, K434). However, no single binding pose can access more than six of these attractive positive residues. Furthermore, distinct binding

motifs were apparent in the docking output. Therefore, for CS-E bound to EphB3, we identified five different binding motifs (Modes 1-5), all in the area of the first fibronectin III domain (Fig 4-1E, S4-10). Modes 1 and 2 (Fig 4-1C, 4-1D) are predicted to have comparable binding energy (-377.5 kcal/mol and -381 kcal/mol, respectively) and each is found 10 times in the best 25 poses making them the most likely candidates for the actual ligand binding site. Detailed images for Mode 1 and Mode 2 are shown in Fig. 4-1D/E. Mode 3 has an energy of -380.5 kcal/mol, making it comparable in energy to Modes 1 and 2, but is only represented by 3 poses. Modes 4 and 5 each have one pose, with energies of -351.9 kcal/mol, and -318.7 kcal/mol, respectively. Given the presence of multiple competing binding sites and the inability of any one pose to interact with all of the charged residues in the region, it is possible that more than one position of the ligand is biologically relevant. Furthermore, it is possible that the less represented binding modes might leave available charged sites that could allow dimerization of two EphB3 proteins, which is a possible mechanism for activation.

Per-residue nonbond energies for each of the five binding modes is shown in Table S4-1. Table S4-2 focuses on the arginine and lysine residues in the binding sites and clearly shows that, while each of these charge residues contributes to the binding energy, the pattern of interactions with these residues differs between the binding modes.

While the best pose from Mode 1 is ~3 kcal/mol worse in energy than the best pose from Mode 2, the poses from Mode 1 are very consistent in their placement (Fig S4-14) and make very good contact with the protein (Fig S4-13). A detailed image of the best Mode 1

pose is shown in Fig S4-12. The poses for Mode 2 are less consistent in their placement (Fig S4-15B). The best pose for Mode 2 (detailed, Fig S4-15A) shows that while the mode generally fits to the protein well (Fig S4-15C), the middle part of the ligand loses contact with the protein (Fig S4-15D). Similar analysis for Modes 3-5 are shown in Fig S4-16 – S4-20. Pharmacophores for all five binding modes are shown in Fig 4-4 – 4-8.

Suggested Post-Prediction Validations

To provide a means for experimentally validating our novel CS-E/EphB3 binding site, we propose several targeted mutations of key residues involved in CS-E binding. The most significant contributions for ligand binding come from eight arginines (R309, R344, R363, R391, R408, R420, R440, R478) and two lysines (K378, K434) in the binding site, as expected for a highly negatively charged ligand. (Table S4-2) We suggest that mutation of these residues to glutamine (or asparagine) should dramatically reduce the binding while minimizing the risk of large structural changes that more severe mutations (e.g., to alanine) could cause. Our methodology in determining suggested mutations is described in the supplemental information.

The differences in the orientations of our five predicted binding modes suggests that specific residues may play a larger role in binding, leaving others to play a lesser role. Since Modes 1 and 2 represented 80% of the top 25 poses, we will focus our results on mutations for these two modes. While all ten positively charged residues contributed to the overall binding energies, the strongest five contributions for Mode 1 were R440 (-174.5 kcal/mol), R363 (-137.7 kcal/mol), R309 (-128.0 kcal/mol), K434 (-125.7 kcal/mol), and

R344 (-120.9 kcal/mol). The strongest six contributions for Mode 2 were R440 (-167.1 kcal/mol), K434 (-142.9 kcal/mol), K378 (-130.6 kcal/mol), R363 (-120.6 kcal/mol), R420 (-115.5 kcal/mol), and R309 (-98.4 kcal/mol). R440 was the strongest contributor for both Mode 1 and Mode 2, suggesting that it should be the first target for specific mutations. R309 and R344 both contributed more strongly to Mode 1, and K378 and R420 both contributed more strongly to Mode 2. Mutation of these residues may be able to provide experimental evidence for which Mode is best. Since R391, R408, and R478 did not contribute strongly to either Mode 1 or Mode 2 mutations of these residues could provide experimental information on whether Modes 3-5 are relevant. Contributions for all residues are presented in Table S4-1.

We carried out *in silico* mutations of the key Mode 1 residues to glutamine, which led to the following changes to the binding energy (positive indicates weaker interactions): R440Q +165.9 kcal/mol, R363Q +131.0 kcal/mol, R309Q +122.7 kcal/mol, R344Q +120.6 kcal/mol, K434Q +114.5 kcal/mol. For Mode 2 the changes to binding energy were: R440Q +160.7 kcal/mol, K434Q +133.8 kcal/mol, K378Q +111.2 kcal/mol, R363Q +100.4 kcal/mol, R420Q +95.2 kcal/mol, R309Q +94.0 kcal/mol.

We recommend that numerous simultaneous mutations be done for tests of our predictions. The reason is that because a large number of charged residues contribute to the binding, mutation to a single residue may be insufficient to significantly alter binding. Moreover since other positive residues are available in the same regions, the ligand might find new interactions in the absence of just one or two key residues.

A more rigorous validation of our predicted binding modes would be to perform mutations that unambiguously increase binding affinity. Consequently we identified mutations of several residues that GAG-Dock suggests should increase binding affinity. We selected these mutations to allow additional contacts with the charged and polar groups on CS-E. Again we considered mutations to glutamine, since the mutated structures may be more likely to fold to the proper structure, than say mutations to alanine. Eight individual mutations for Mode 1 predicted to make new contacts with the ligand are (negative indicates stronger binding): T448Q (-18.77 kcal/mol), V339Q (-13.65 kcal/mol), I446Q (-12.48 kcal/mol), A442Q (-11.97 kcal/mol), N445Q (-11.42 kcal/mol), T319Q (-11.20 kcal/mol), N323Q (-4.23 kcal/mol), and N322Q (-0.78 kcal/mol). Seven individual mutations for mode 2 predicted to make new contacts are: E424Q (-117.08 kcal/mol), V339Q (-16.14 kcal/mol), T422Q (-16.14 kcal/mol), T338Q (-14.13 kcal/mol), N445Q (-7.44 kcal/mol), N323Q (-2.97 kcal/mol), and S341Q (-0.92 kcal/mol). These single residue mutations are summarized in Table S4-3 for Mode 1 and Table S4-5 for Mode 2. (Modes 3-5 in Tables S4-7, -8, -10)

Based on their individual predicted contributions to binding, we suggest the following set of 7 mutations for the first experiments to test Mode 1: T319Q, N322Q, V339Q, A442Q, A443N, I446Q, and T448Q. We predicted that this set of mutations for Mode 1 improves binding energy by 66.03 kcal/mol, or 16.5% better than binding to the wild-type. The predicted binding site for this set of mutations for Mode 1 is shown in Fig 4-9A. Energies for all sets of mutations tested for Mode 1 are shown in Table S4-4.

The presence of E424 in the neighborhood of Mode 2 is puzzling, since it has a repulsive interaction with the ligand. Mutating E424 to glutamine resulted in a significant increase in binding energy, but might also modify the binding site. Therefore, we propose two sets of mutations for Mode 2. The first set is: N323Q, T338N, V339Q, S341Q, T422Q, and N445Q. This improved binding energy by 46.64 kcal/mol or 11.4% better than the wild-type. The second set for mode 2 adds the E424Q mutant, resulting in an improvement of 163.84 kcal/mol or 40%. The predicted binding site for the non-E424Q set for Mode 2 is shown in Fig 4-9B. Energies for all sets of mutations tested for Mode 2 are shown in Table S4-6.

We applied this same procedure also to Modes 3-5, with the results reported in Tables S4-7, S4-9, and S4-11. Detailed images and pharmacophores of the predicted mutant binding sites for the selected sets of mutations are shown in Figures S4-21 to S4-28.

Since we found five competitive binding modes for CS-E/EphB3, it may be that CS-E binding recognizes a binding region or ensemble of binding sites rather than a specific binding site that is typical for binding of small molecules. We selected CS-E octasaccharide as a representative of the natural, extended polysaccharide. The experimental system may well be more complicated with interactions beyond a single octasaccharide binding mode. Indeed none of our five predicted binding modes interacts with all 10 positively charged residues within the binding region. We suggest that these additional charged residues may serve two purposes. First, the extra, non-shared residues could allow for a single polysaccharide to bind to two proteins using one mode for the first

protein and a different mode for the second protein, possibly allowing for dimerization and activation of the proteins. Second, the presence of extra positive residues could allow for the ligand to migrate within the binding region without losing adhesion to the protein. To test this second possibility we suggest mutations expected to increase binding affinity. A single mutation from arginine or lysine to glutamine or asparagine might not change the binding as much as we predict, because the CS-E might move its preferred binding region slightly to account for the reduced arginines. This suggests that validation be done with multiple simultaneous mutations. Of course, mutating multiple residues simultaneously may increase the likelihood of misfolding, rendering the study useless. For a single beneficial mutation, such misfolding is less likely, although the change in binding affinity may be less dramatic.

Conclusions

Studying the CS-E/EphB3 system computationally was a difficult challenge: a large, highly negatively charged ligand, and a protein with a completely unknown binding site. Furthermore, the related CS-A ligand was shown not to bind experimentally, and neither CS-A nor CS-E bound to the similar EphB2 protein. Our goal was to identify structural explanations for these differences. In both cases we successfully identified the cause to be related to the charges on the ligand and/or the protein. EphB2 lacks the positively-charged region of EphB3 and thus cannot bind the negatively-charged CS ligands. Similarly, the reduced negative charge of CS-A relative to CS-E means that it does not bind with sufficient strength to EphB3. The *pattern* of sulfation does not appear to be a significant factor, as CS-D binds comparably to CS-E. This is likely due to flexibility of the sulfate

groups on the ligand and the arginine and lysine sidechains on the protein. Specific patterns are not needed for such a general interaction.

We have further used our predicted structural information to suggest mutation experiments that would validate one or more of our binding modes for CS-E. As mentioned previously, we consider mutations from arginine/lysine to alanine to lack subtlety. Loss of binding from such mutations could be due to larger structural changes than simple binding site modification. Instead, we have suggested sets of mutations to *improve* binding, which would validate our binding modes with much less ambiguity. We encourage our experimental colleagues to attempt these sets of mutations:

- T319Q, N322Q, V339Q, A442Q, A443N, I446Q, and T448Q
- N323Q, T338N, V339Q, S341Q, T422Q, and N445Q (optionally E424Q)

The first set should increase binding affinity for CS-E if our predicted Mode 1 is the correct binding pose, and the second set should increase binding affinity for Mode 2.

This project highlights the role that computation can have in studying complicated biological systems, and in complementing and directing experiment. The specificity of the binding site predictions suggests clear follow-up experiments to further understanding of the role of CS-E in EphB3 activation, which, hopefully, will suggest new directions for computation.

Acknowledgements

ARG, RA, and WAG were supported from NIH (R01-NS071112, R01-NS073115, and R01-AI040567) and other funds donated to the Materials and Process Simulation Center. The computers used in this research were funded by grants (to WAG) from DURIP (Defense University Research Instrument Program) and from NSF (equipment part of the NSF-MRSEC-CSEM). CJR, GM, and LCH-W were supported by the National Institutes of Health (grant R01 GM084724 to L.C.H-W).

References

1. Rogers, C. J., Miller, G., Hsieh-Wilson, L. C. (to be published)
2. Griffith, A. R., Rogers, C. J., Miller, G., Abrol, R., Hsieh-Wilson, L. C., Goddard, W. A., III. (2016, to be published). Predicting glycosaminoglycan-surface protein interactions: Implications for studying axonal growth.
3. Scott, C. E., Ahn, K. H., Graf, S. T., William A Goddard, I., Kendall, D. A., & Abrol, R. (2016). Computational Prediction and Biochemical Analyses of New Inverse Agonists for the CB1 Receptor. *Journal of Chemical Information and Modeling*, *56*(1), 201–212. http://doi.org/10.1021/acs.jcim.5b00581
4. Li, Q., Kim, S.-K., Goddard, W. A., III, Chen, G., & Tan, H. (2015). Predicted Structures for Kappa Opioid G-Protein Coupled Receptor Bound to Selective Agonists. *Journal of Chemical Information and Modeling*, *55*(3), 614–627. http://doi.org/10.1021/ci500523z
5. Abrol, R., Trzaskowski, B., Goddard, W. A., III, Nesterov, A., Olave, I., & Irons, C. (2014). Ligand- and mutation-induced conformational selection in the CCR5 chemokine G protein-coupled receptor. *Proceedings of the National Academy of Sciences*, *111*(36), 13040–13045. http://doi.org/10.1073/pnas.1413216111
6. Bray, J. K., Abrol, R., Goddard, W. A., III, Trzaskowski, B., & Scott, C. E. (2014). SuperBiHelix method for predicting the pleiotropic ensemble of G-protein–coupled receptor conformations. *Proceedings of the National Academy of Sciences*, *111*(1), E72–E78. http://doi.org/10.1073/pnas.1321233111
7. Kim, S.-K., & Goddard, W. A. (2014). Predicted 3D structures of olfactory receptors with details of odorant binding to OR1G1. *Journal of Computer-Aided Molecular Design*, *28*(12), 1175–1190. http://doi.org/10.1007/s10822-014-9793-4
8. Kim, S.-K., Goddard, W. A., Yi, K. Y., Lee, B. H., Lim, C. J., & Trzaskowski, B. (2014). Predicted Ligands for the Human Urotensin-II G Protein-Coupled Receptor

with Some Experimental Validation. *ChemMedChem*, *9*(8), 1732–1743. http://doi.org/10.1002/cmdc.201402087
9. Tan, J., Abrol, R., Trzaskowski, B., & William A Goddard, I. (2012). 3D Structure Prediction of TAS2R38 Bitter Receptors Bound to Agonists Phenylthiocarbamide (PTC) and 6-n-Propylthiouracil (PROP). *Journal of Chemical Information and Modeling*, *52*(7), 1875–1885. http://doi.org/10.1021/ci300133a
10. Kim, S.-K., Fristrup, P., Abrol, R., & William A Goddard, I. (2011a). Structure-Based Prediction of Subtype Selectivity of Histamine H3 Receptor Selective Antagonists in Clinical Trials. *Journal of Chemical Information and Modeling*, *51*(12), 3262–3274. http://doi.org/10.1021/ci200435b
11. Kim, S.-K., Riley, L., Abrol, R., Jacobson, K. A., & Goddard, W. A. (2011b). Predicted structures of agonist and antagonist bound complexes of adenosine A3 receptor. *Proteins: Structure, Function, and Bioinformatics*, *79*(6), 1878–1897. http://doi.org/10.1002/prot.23012
12. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences*, *98*(18), 10037–10041. http://doi.org/10.1073/pnas.181342398
13. Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., & Baker, N. A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, *35*(suppl 2), W522–W525. http://doi.org/10.1093/nar/gkm276
14. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, *32*(suppl 2), W665–W667. http://doi.org/10.1093/nar/gkh381
15. Ligand binding domain of human ephrin type-B receptor 3. (2011) http://dx.doi.org/10.2210/pdb3p1i/pdb
16. Chrencik, J. E., Brooun, A., Recht, M. I., Nicola, G., Davis, L. K., Abagyan, R., et al. (2007). Three-dimensional Structure of the EphB2 Receptor in Complex with an Antagonistic Peptide Reveals a Novel Mode of Inhibition. *Journal of Biological Chemistry*, *282*(50), 36505–36513. http://doi.org/10.1074/jbc.M706340200
17. Seiradake, E., Harlos, K., Sutton, G., Aricescu, A. R., & Jones, E. Y. (2010). An extracellular steric seeding mechanism for Eph-ephrin signaling platform assembly. *Nature Structural & Molecular Biology*, *17*(4), 398–402. http://doi.org/10.1038/nsmb.1782
18. Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, *22*(2), 195–201. http://doi.org/10.1093/bioinformatics/bti770
19. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, *42*(W1), gku340–W258. http://doi.org/10.1093/nar/gku340
20. Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical

perspective. *Electrophoresis*, *30*(S1), S162–S173. http://doi.org/10.1002/elps.200900140

21. Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(suppl 1), D387–D392. http://doi.org/10.1093/nar/gkn750
22. Mayo, S. L., Olafson, B. D., & Goddard, W. A. (2002). DREIDING: a generic force field for molecular simulations. *The Journal of Physical Chemistry*, *94*(26), 8897–8909. http://doi.org/10.1021/j100389a010
23. Lim, K. T., Brunett, S., Iotov, M., & McClurg, R. B. (1997). Molecular dynamics for very large systems on massively parallel computers: the MPSim program. *Journal of Computational Chemistry*, *18*(4), 501-521
24. Kam, V. W. T., & William A Goddard, I. (2008). Flat-Bottom Strategy for Improved Accuracy in Protein Side-Chain Placements. *Journal of Chemical Theory and Computation*, *4*(12), 2160–2169. http://doi.org/10.1021/ct800196k

Figures & Tables

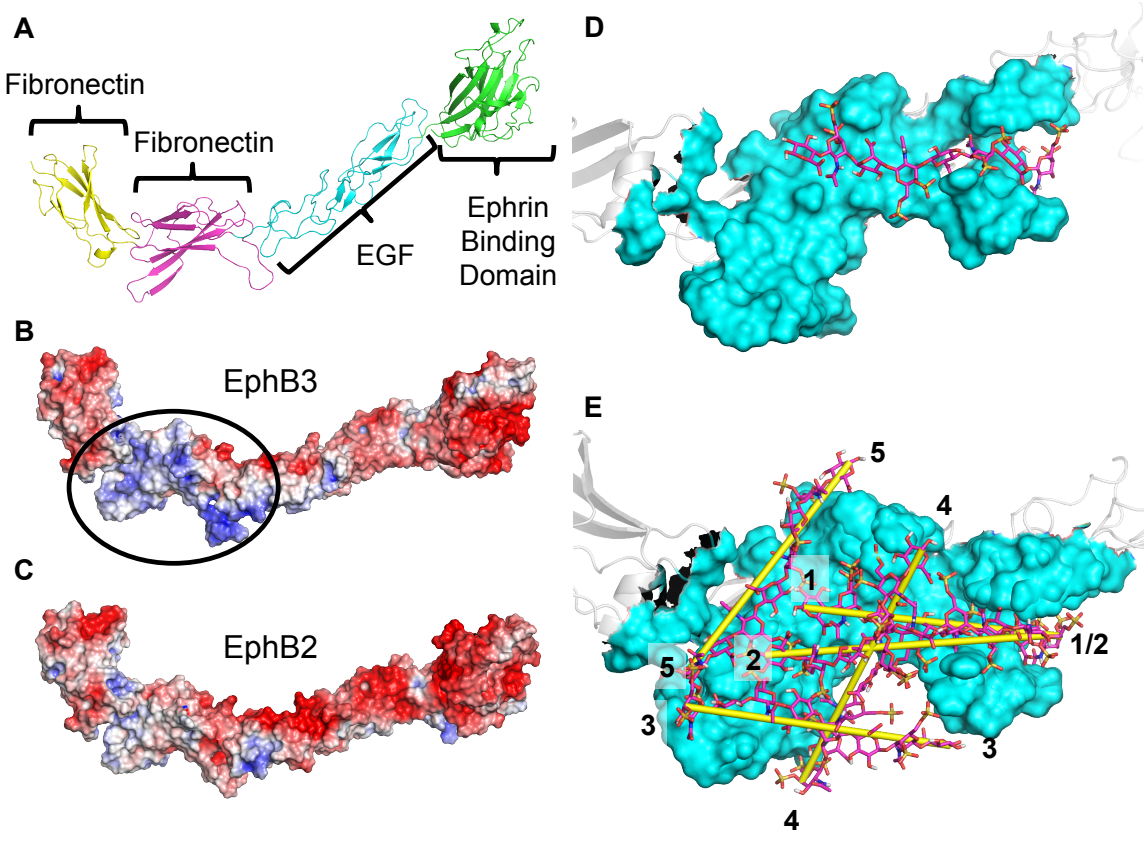

Figure 4-33 – (A) Model of EphB3. (B-C) Electrostatics mapped onto the surfaces of EphB3 and EphB2. Circled region denotes binding region for top five EphB3/CS-E binding modes (cyan region in D-E). (D) Predicted best EphB3/CS-E binding mode. (E) Overlay of predicted Top five EphB3/CS-E binding modes. The general orientation of binding modes shown in yellow.

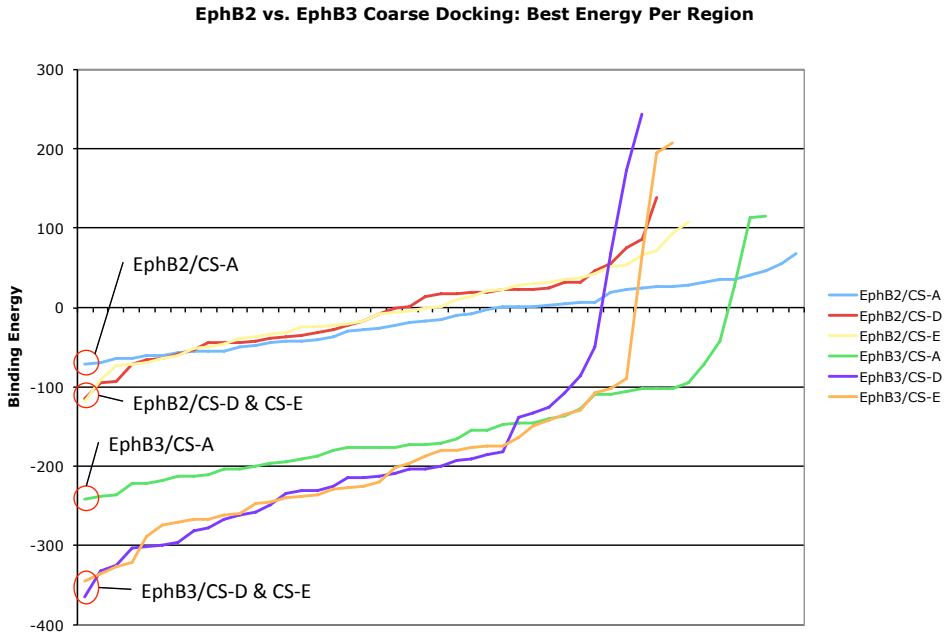

Figure 4-34 –Plot of the energy of the best pose in each region after coarse docking for CS-A, CS-D, and CS-E docked to EphB2 and EphB3. It is clear from the chart that the binding energies are much worse for EphB2 than EphB3. Additionally, CS-A has a much worse binding energy to EphB3 than CS-D and CS-E.

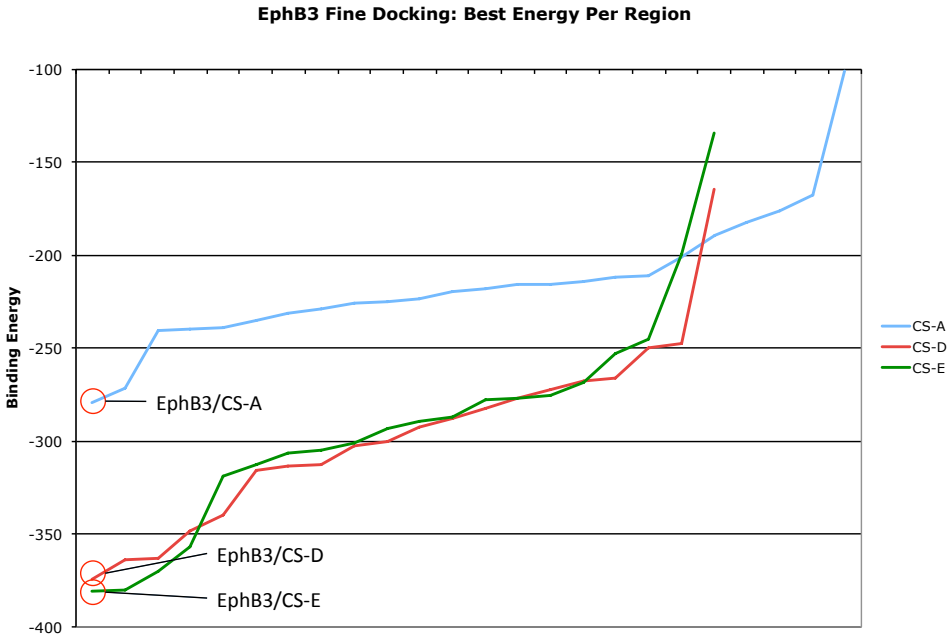

Figure 4-35 – Plot of the energy of the best pose in each region after fine docking for CS-A, CS-D, and CS-E docked to EphB3. After fine-level docking, CS-E binds slightly better than CS-D, and both bind significantly better than CS-A.

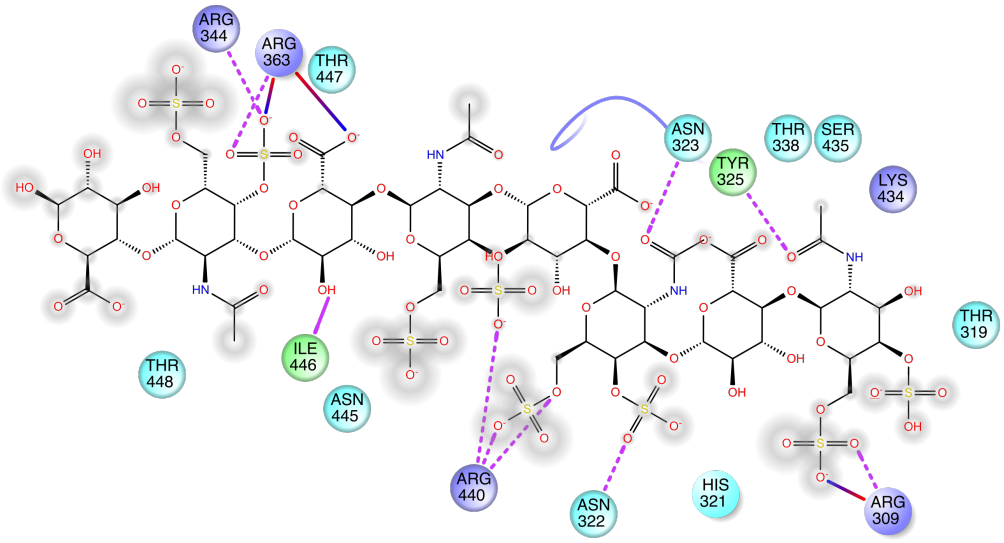

Figure 4-36 – Pharmacophore for best pose in EphB3/CS-E mode 1.

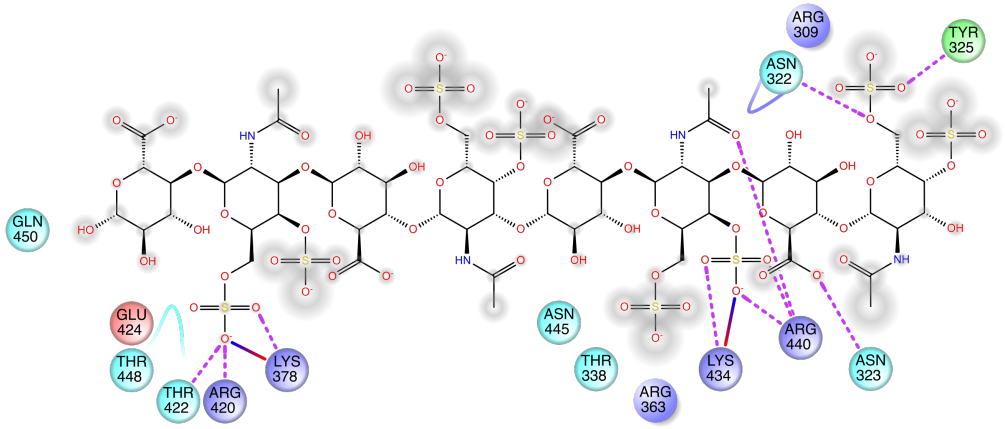

Figure 4-37 – Pharmacophore for best pose in EphB3/CS-E mode 2.

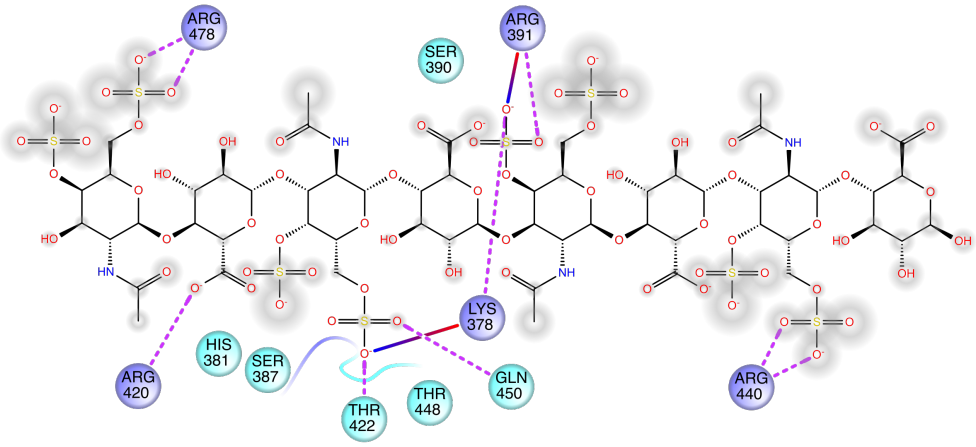

Figure 4-38 – Pharmacophore for best pose in EphB3/CS-E mode 3.

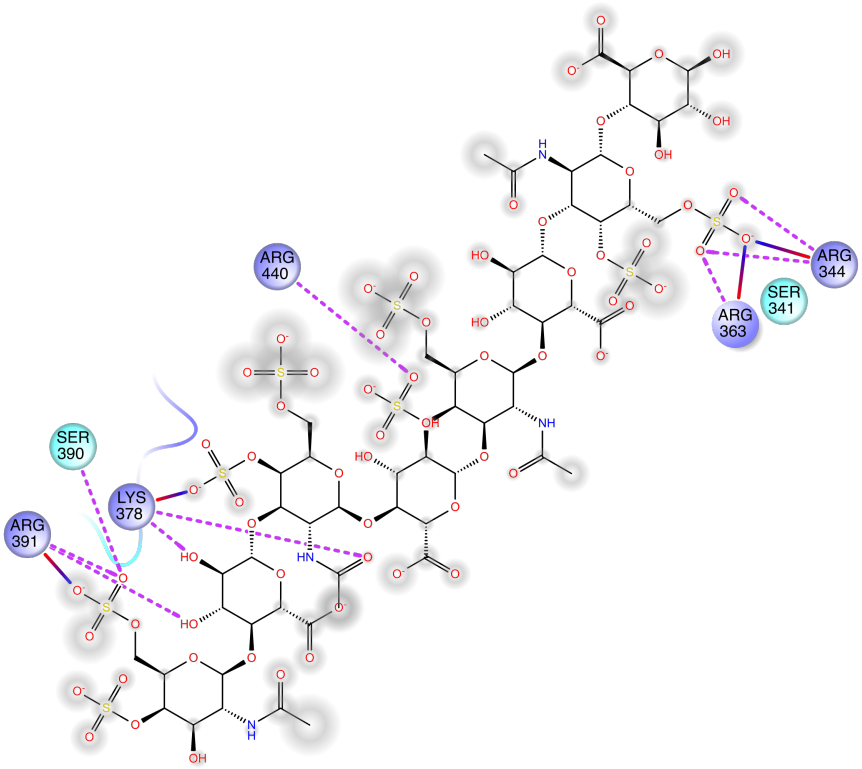

Figure 4-39 – Pharmacophore for best pose in EphB3/CS-E mode 4.

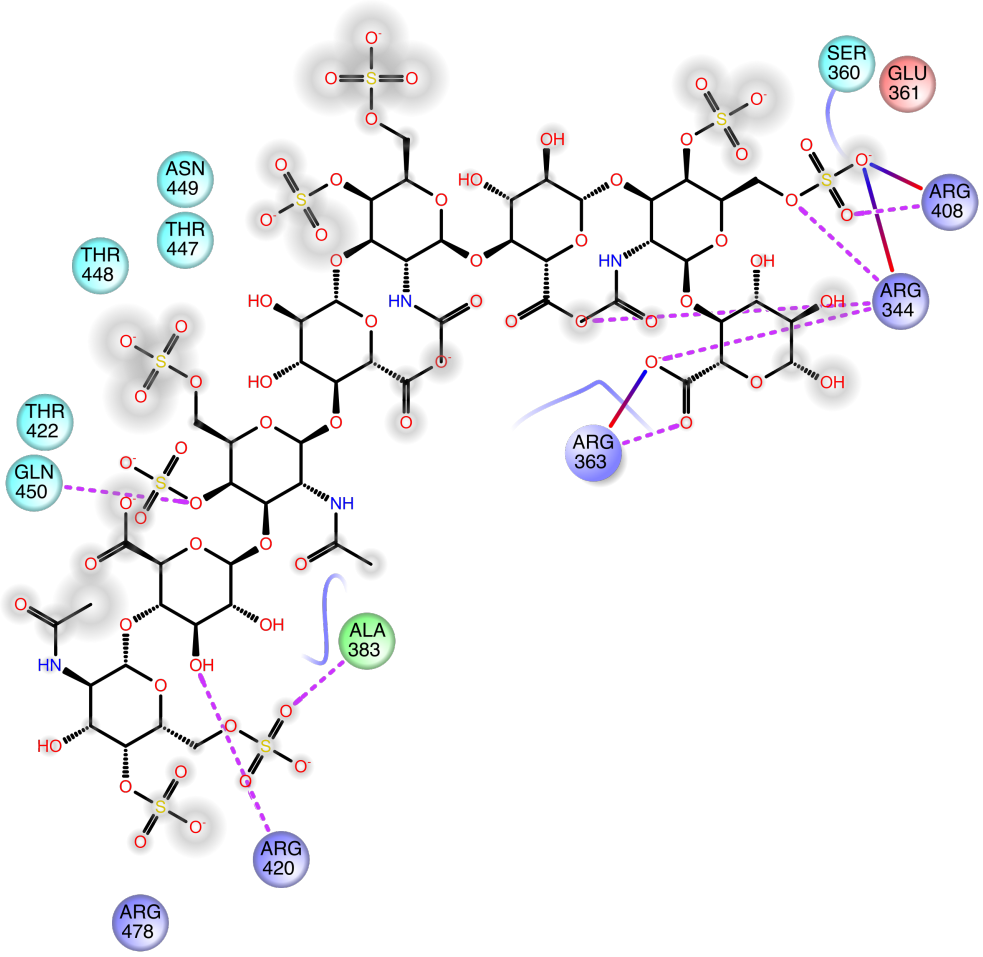

Figure 4-40 – Pharmacophore for best pose in EphB3/CS-E mode 5.

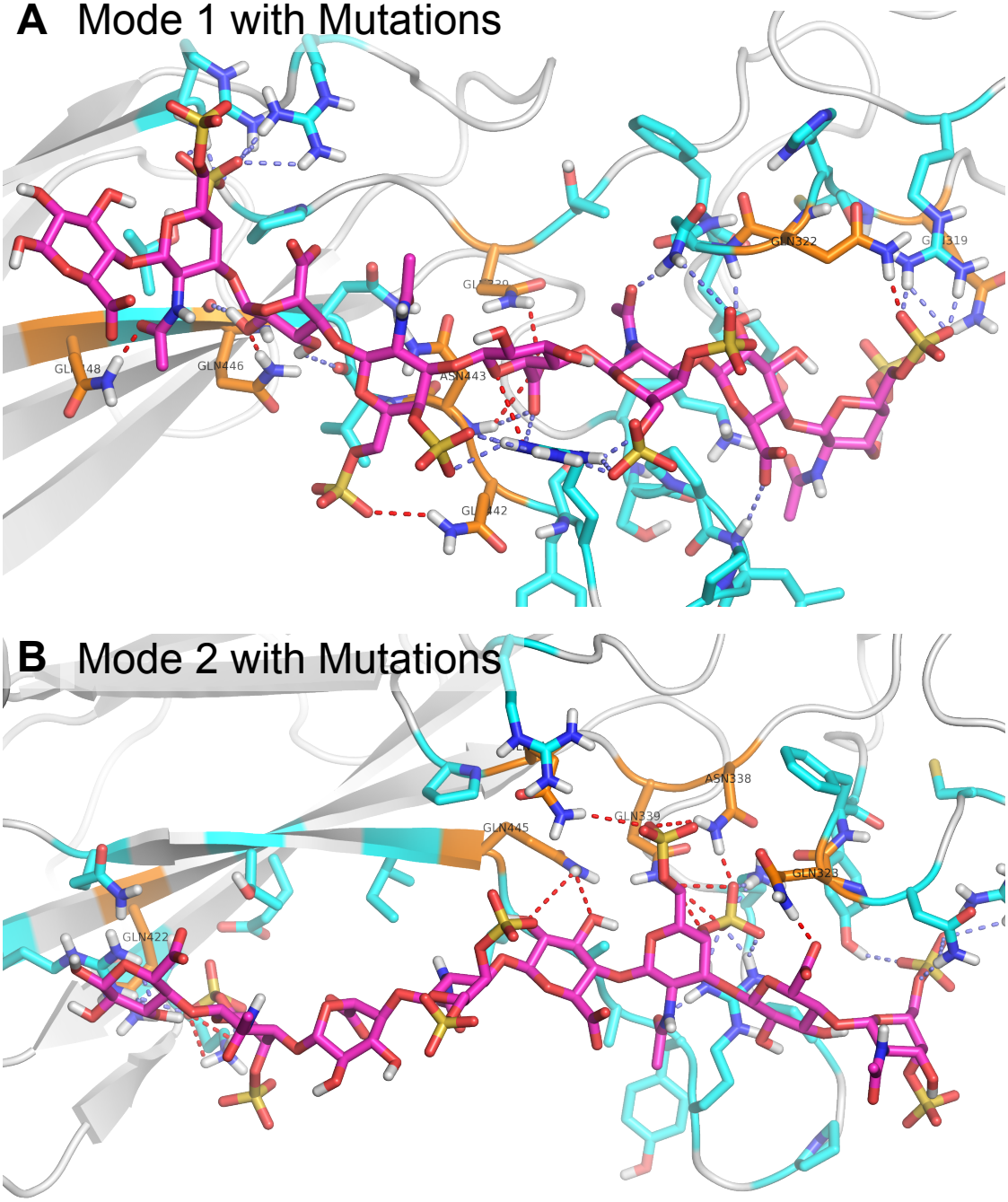

Figure 4-41 – Mutations to Gln predicted to increase EphB3/CS-E binding. Mutated residues are colored orange. Red hydrogen bond markers denote new hydrogen bonds with the ligand due to mutations and blue markers denote hydrogen bonds to the ligand that are common to both mutant and wild type. (A) Mutations for binding mode 1: T448Q, V339Q, I446Q, A442Q, T319Q, A443N, N322Q. Binding energy improved by 66.0 kcal/mol or 16.5% over wild type. (B) Mutations for binding mode 2: V339Q, T422Q, T338N, N445Q, N323Q, S341Q. Binding energy improved by 46.6 kcal/mol or 11.4% over wild type.

Supplemental Information

Mutation Methodology

In order to identify mutations that could validate our predicted CS-E/EphB3 binding modes we performed *in silico* mutations. Each residue – excluding proline and glycine – in the 5Å binding site was individually mutated to glutamine using SCREAM[24]. Simultaneously, the rest of the sidechains were also optimized to allow them to accommodate the mutated sidechain's position. The binding site and ligand were then minimized for 50 steps of conjugate gradient minimization using DREIDING[22] in MPSim[23]. At the end of this procedure mutations were identified that increased the binding energy of the ligand, summarized in Tables S4-3, S4-5, S4-7, S4-8, and S4-10. Based on these single mutants, sets of combined mutants that should increase binding were identified and tested. Again, SCREAM was used to perform the mutations as well as optimize the remaining sidechains in the binding site, followed by 50 steps of minimization. In some cases two mutant sidechains would clash, resulting in non-optimal interactions with the ligand. Thus additional sets that omitted some mutations were tested. Additionally, glutamine proved to be too large to make a good interaction with the ligand in some cases, thus asparagine was tried instead. In the end, one set of mutants was identified for each mode that maximized ligand binding and resulted in each mutated residue making a new hydrogen bond with the ligand. An additional set was generated each for Mode 2 and Mode 5. These modes have nearby glutamic acids (E424 and E361, respectively). We are wary of mutating these residues because they may have a special role in the structure or function of the EphB3 receptor or binding site. However, sets of mutations were generated for Mode 2 and Mode

5 that included the respective E424Q and E361Q mutations. The binding energies for the sets of mutations are summarized in Tables S4-4, S4-6, S4-7, S4-9, and S4-11.

Supplemental Figures & Tables

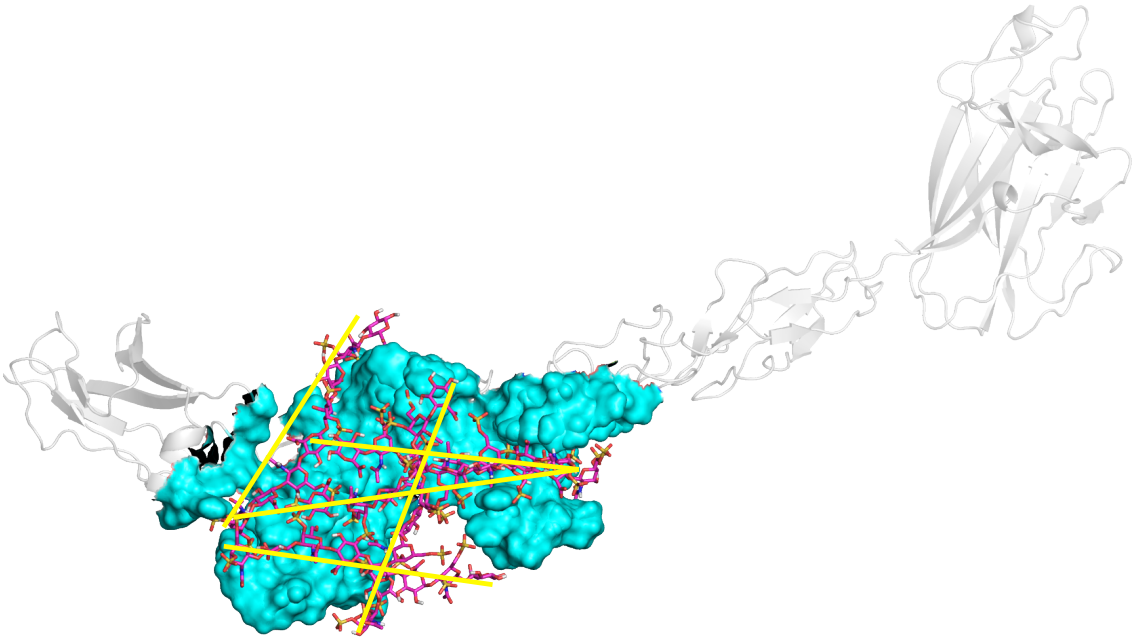

Figure S4-42 – Schematic showing placement of CS-E binding modes bound to EphB3.

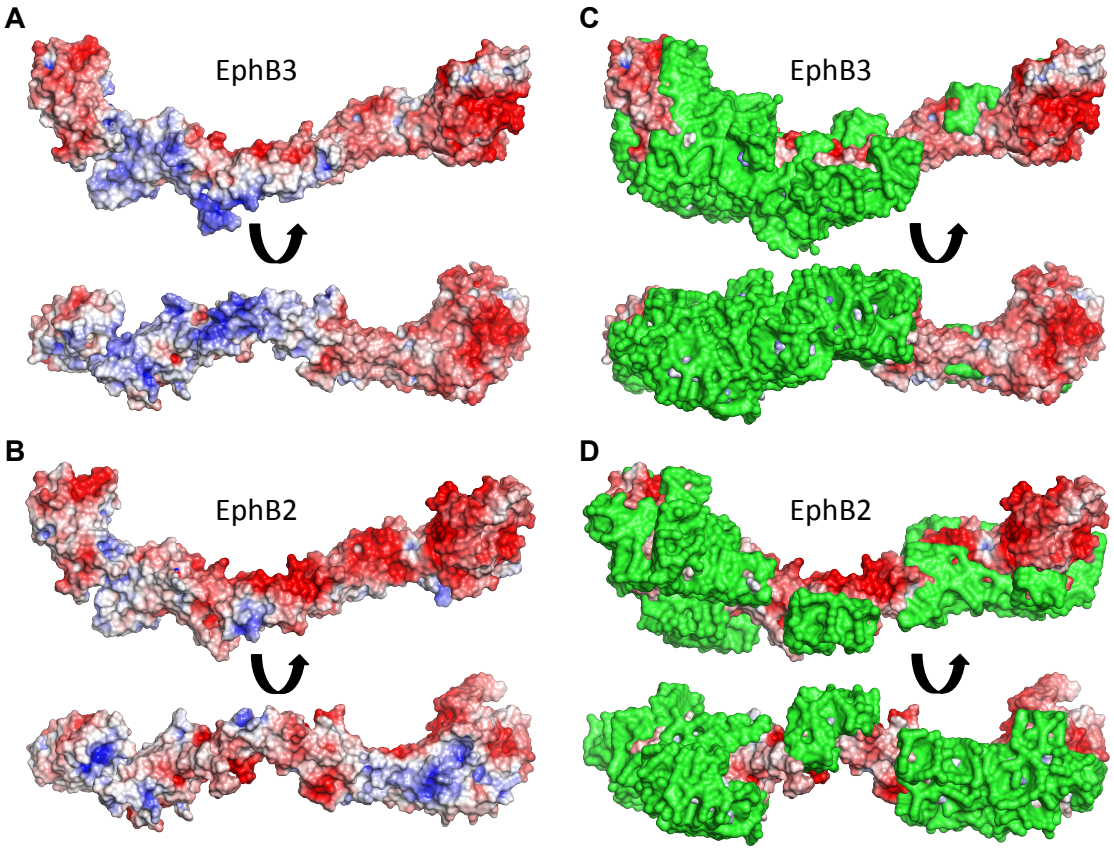

Figure S4-43 – Electrostatic surfaces of (A) EphB3 and (B) EphB2. Sphere regions used for coarse docking are shown in green for (C) EphB3 and (D) EphB2. Note that the regions sampled cover the positively charged regions of the proteins.

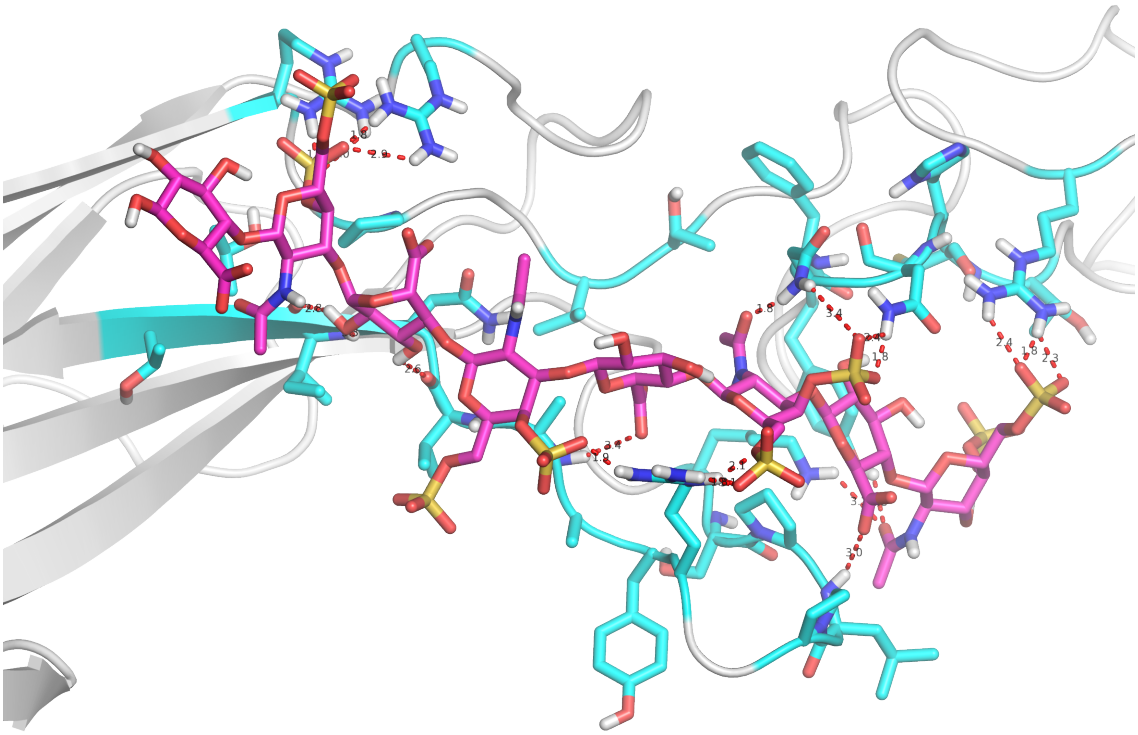

Figure S4-44 – Detailed view of the best CS-E/EphB3 Mode 1 binding pose.

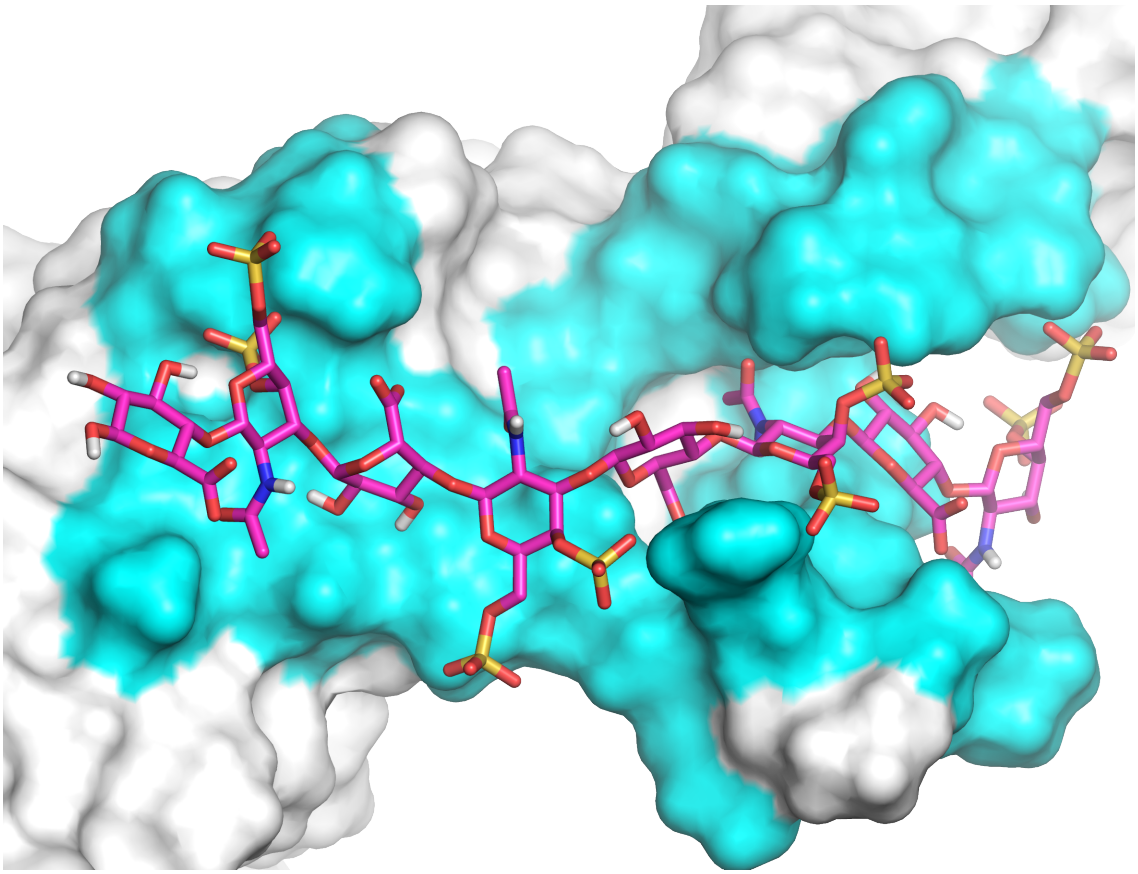

Figure S4-45 - The best CS-E/EphB3 Mode 1 binding pose, with the VDW surface of the protein shown to illustrate how well the ligand fits to the protein.

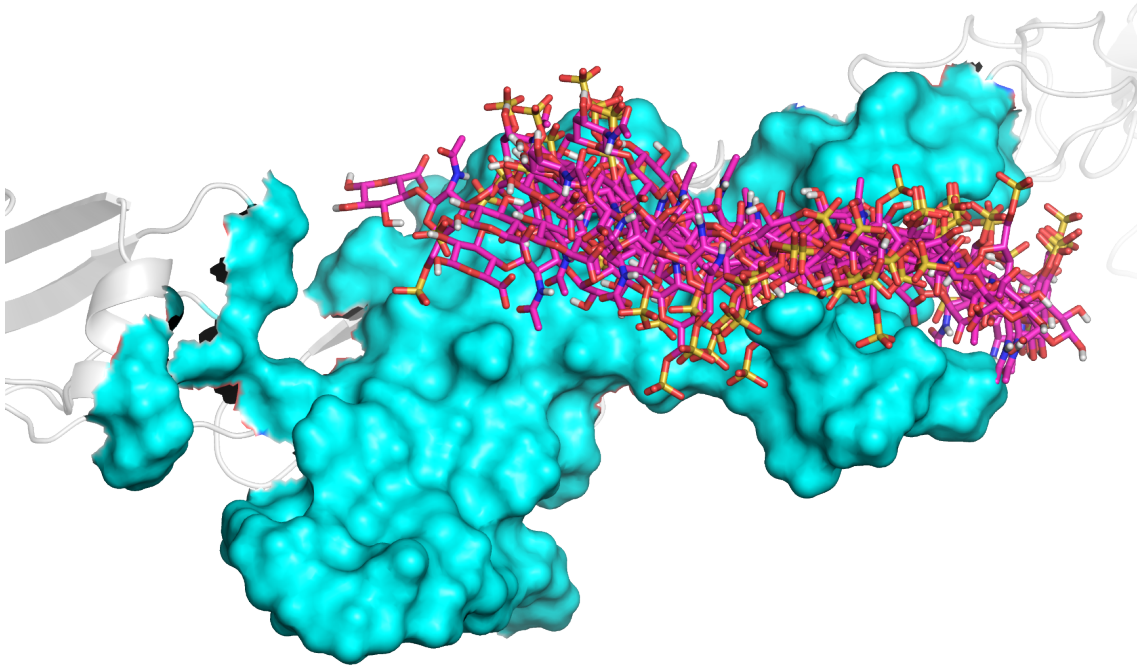

Figure S4-46 – The placement of all CS-E/EphB3 Mode 1 poses. The top pose in this mode is the #3 pose overall (-377.46 kcal/mol), but this mode shows the most consistency in placement.

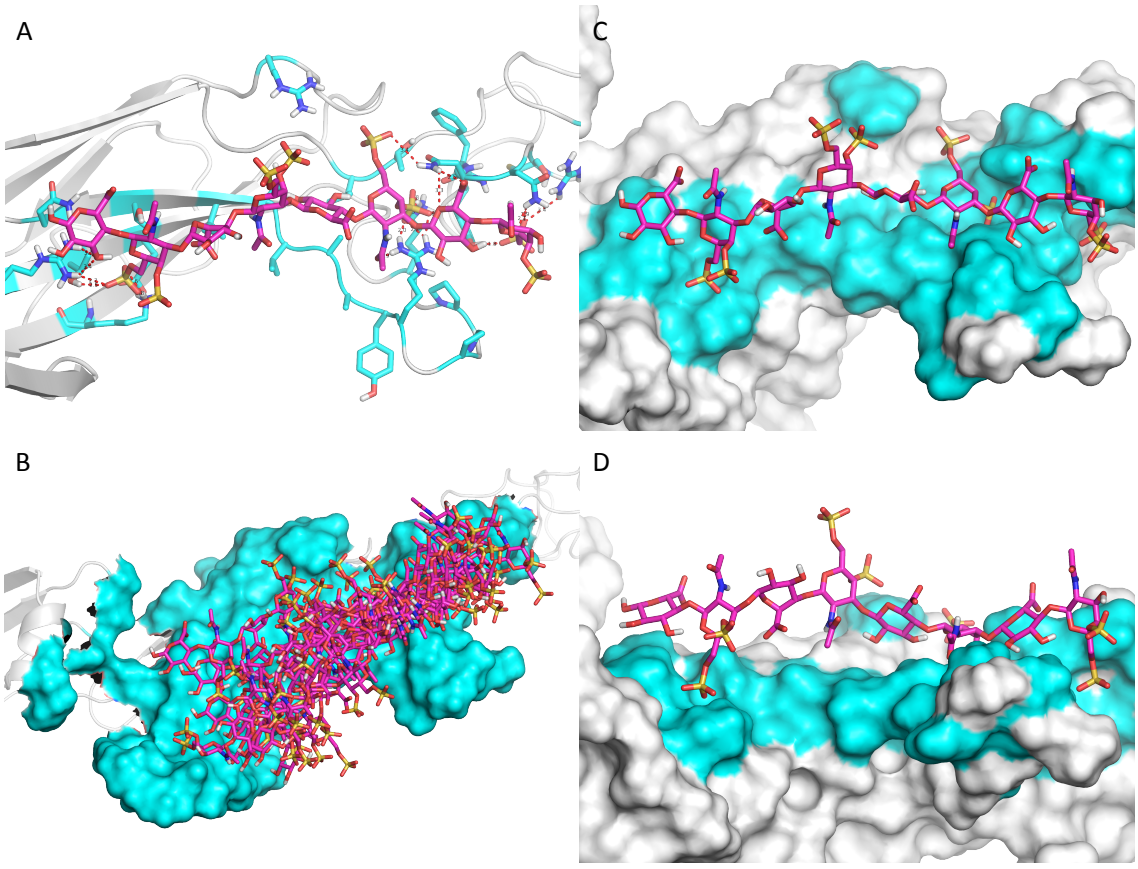

Figure S4-47 – (A) Detailed view of CS-E docked to EphB3 in top pose from binding mode 2. (B) Placement of all CS-E poses docked to EphB3 in binding mode 2. The top pose in this mode is the #1 pose overall (-380.80 kcal/mol), but shows less consistency in pose placement than Mode 1. (C) Top view of the best Mode 2 pose appears to fit closely to the protein surface, but the rotated view (D) shows that the middle section of the octasaccharide is separated from the surface.

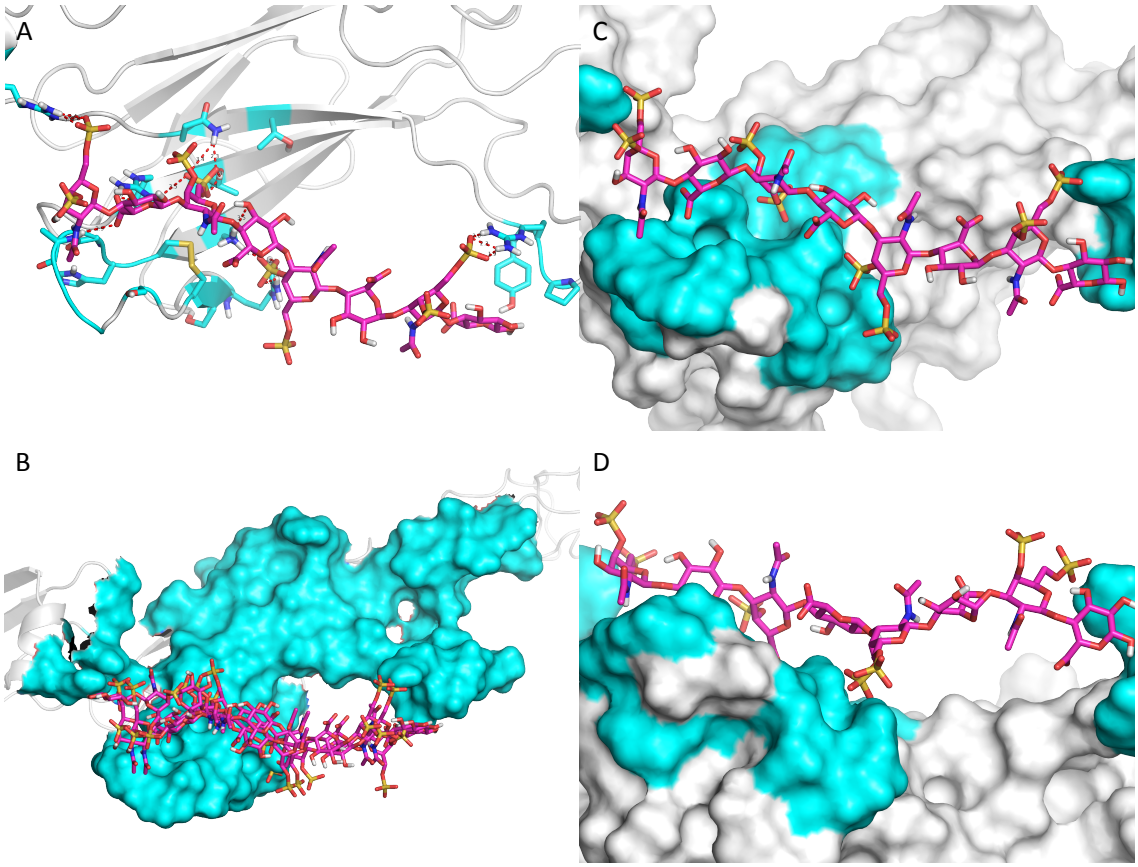

Figure S4-48 – (A) Detailed view of CS-E docked to EphB3 in top pose from binding mode 3. (B) Placement of all CS-E poses docked to EphB3 in binding mode 3. The top pose in this mode is the #2 pose overall (-380.53). This mode shows less contact with the surface of the protein.. (C) Top view of the best Mode 3 pose appears to fit closely to the protein surface, but the rotated view (D) shows that the much of the octasaccharide is separated from the surface.

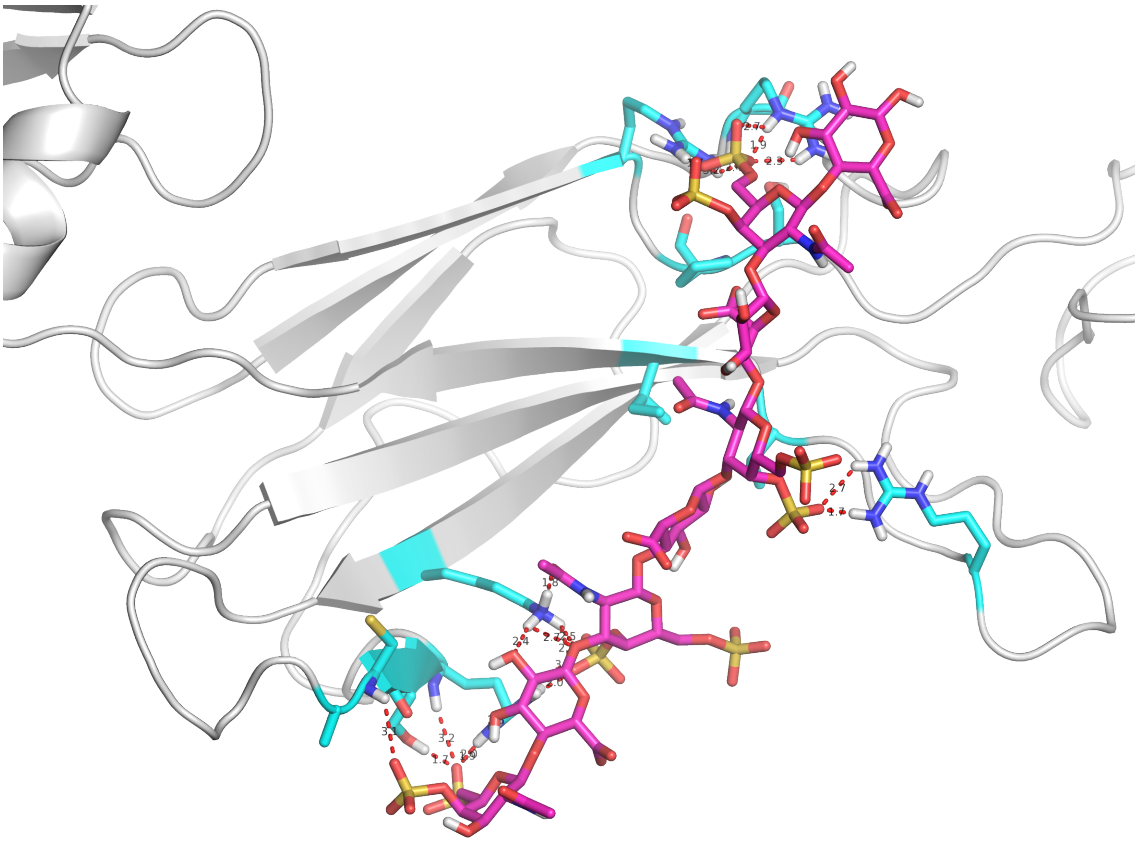

Figure S4-49 – Placement of only CS-E pose docked to EphB3 in binding mode 4. This mode contains only one pose in the top 25 poses. This pose is #6 overall (-351.91 kcal/mol).

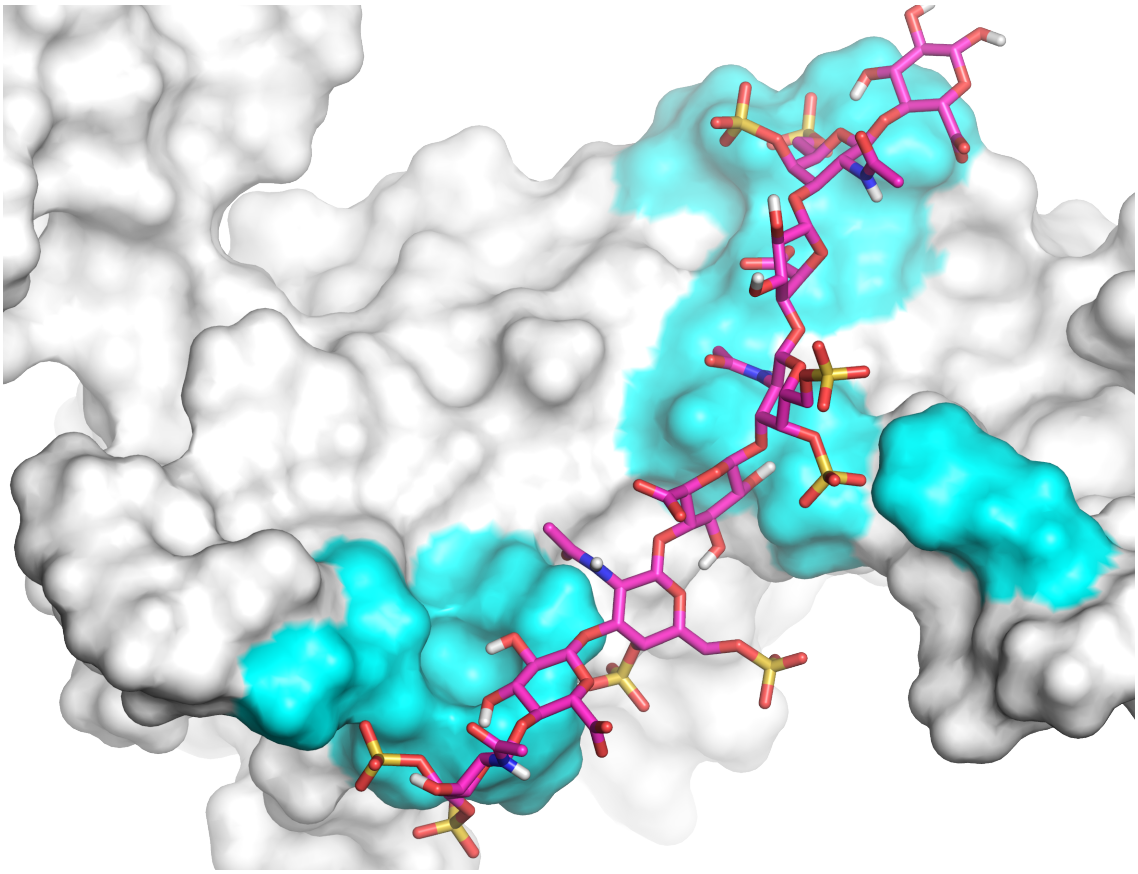

Figure S4-50 – Placement of the only CS-E pose docked to EphB3 in binding mode 4, with the protein surface shown.

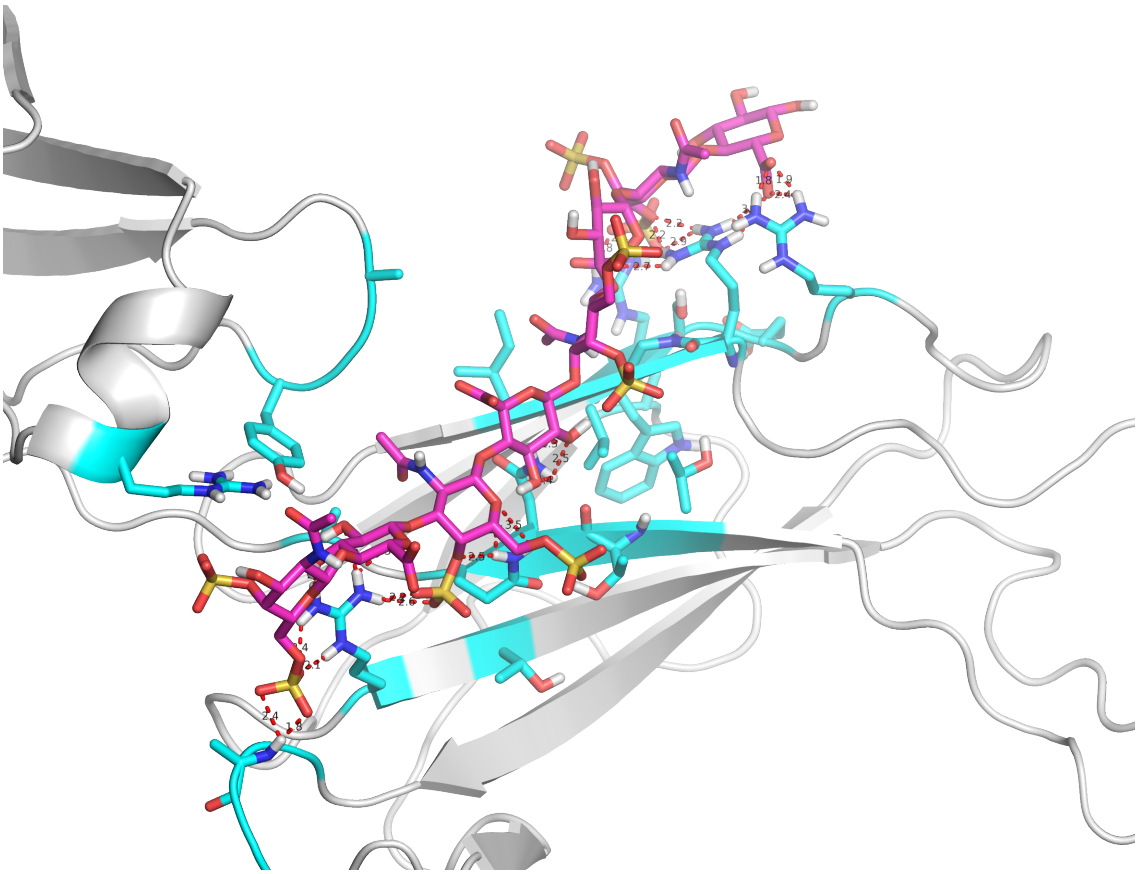

Figure S4-51 – Placement of the only CS-E pose docked to EphB3 in binding mode 5. This mode contains only one pose in the top 25 poses. This pose is #22 overall (-318.65 kcal/mol).

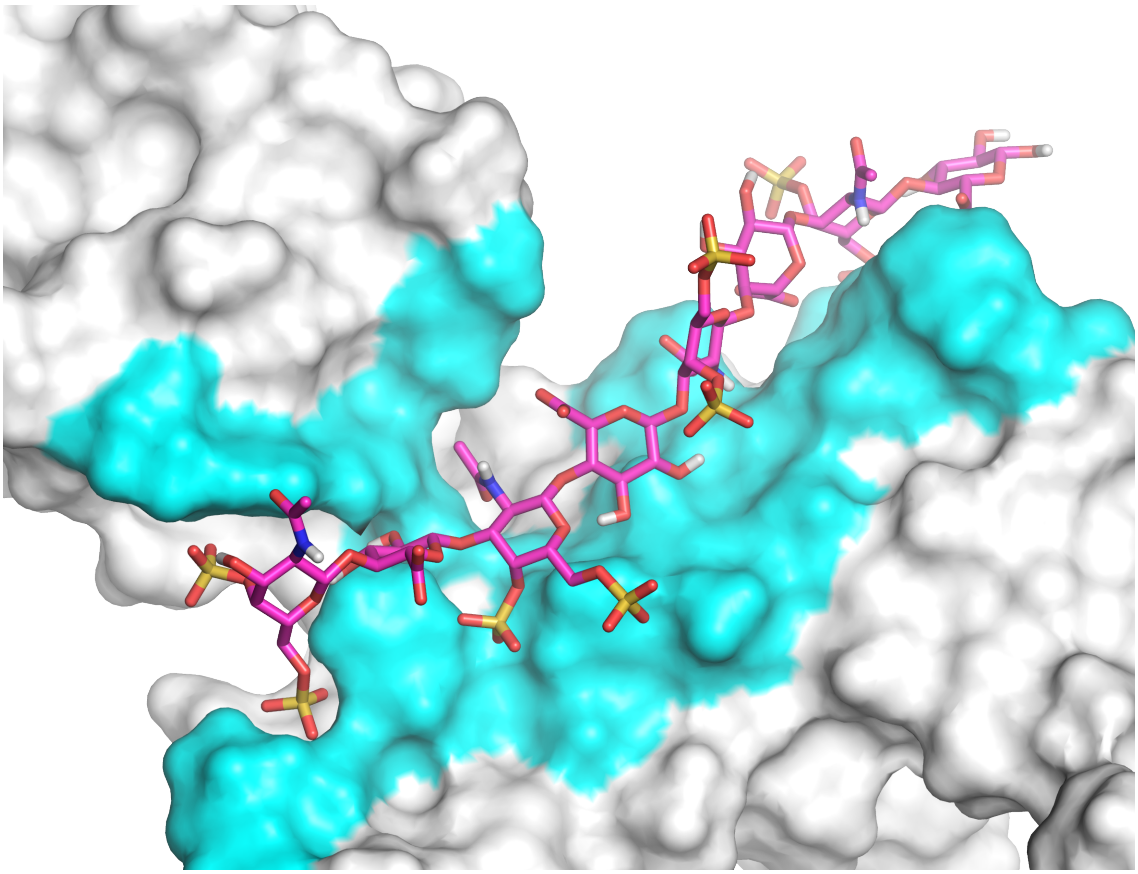

Figure S4-52 – Placement of the only CS-E pose docked to EphB3 in binding mode 5, with the protein surface shown.

Table S4-26 – Nonbond interactions by residue for the top pose in each of the five binding modes. As expected, favorable interactions are dominated by arginines and lysines (green), unfavorable interactions are dominated by glutamic acids (red). Ordered by nonbond energy for Mode 1.

Res	Num	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
ARG	440	-174.5	-167.1	-128.4	-137.4	-57.1
ARG	363	-137.7	-120.6	-57.8	-127.1	-138.0
ARG	309	-128.0	-98.4	-40.8	-45.4	-38.8
LYS	434	-125.7	-142.9	-48.7	-61.4	-54.9
ARG	344	-120.9	-71.5	-50.8	-115.5	-143.2
LYS	378	-85.9	-130.6	-167.5	-175.6	-92.9
ARG	420	-59.6	-115.5	-149.5	-72.9	-162.5
ARG	391	-57.7	-69.6	-139.1	-158.2	-56.5
ARG	408	-56.3	-55.9	-46.9	-55.7	-124.8
ARG	478	-40.6	-47.1	-120.7	-48.4	-121.1
ASN	323	-17.0	-6.2	1.8	2.3	0.5
ASN	322	-10.5	-14.5	-1.5	-1.4	0.2
ALA	443	-9.2	-8.0	-1.6	-3.8	-0.7
THR	319	-7.0	-2.9	-0.6	-0.4	0.2
TYR	325	-6.4	-5.0	0.1	0.3	0.8
THR	338	-6.4	-10.4	-0.5	0.1	0.1
SER	341	-6.3	-5.0	-0.2	-4.0	0.6
PRO	342	-3.8	-2.5	-2.0	-2.2	0.6
ASN	449	-3.3	-3.8	-4.2	0.4	-8.0
ALA	442	-3.1	-2.9	0.6	0.9	0.4
THR	422	-2.7	-9.5	-10.6	-2.3	-5.5
TRP	359	-2.2	-1.7	-0.8	-1.8	-0.7
SER	435	-2.1	-2.3	-0.7	-0.1	-0.4
THR	447	-1.7	-0.1	-0.3	0.6	-6.3
PRO	438	-1.6	-1.3	0.0	0.8	0.4
LEU	437	-1.6	-2.1	1.4	1.0	0.2
ASN	445	-1.5	-2.4	-1.0	-2.3	0.2
ALA	529	-1.0	-0.8	-0.1	-0.8	-2.8
HIS	321	-0.9	-1.1	0.8	-0.3	1.1

Res	Num	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
THR	448	-0.9	-5.6	-2.0	-1.4	-0.3
GLY	530	-0.9	-1.4	-2.3	-1.0	-4.1
GLY	382	-0.7	-1.3	-7.2	-0.2	-9.7
ALA	383	-0.7	-1.2	-6.3	-0.8	-6.9
CYX	389	-0.7	-1.8	-2.2	-0.3	-1.0
ALA	452	-0.2	-0.5	-1.0	-0.5	-2.2
GLN	450	-0.1	-6.8	-16.2	-2.8	-10.1
ALA	388	-0.1	0.1	-1.1	-6.0	2.3
SER	390	0.0	1.1	0.6	-8.3	1.9
VAL	339	0.0	1.4	-0.9	0.4	0.2
GLY	345	0.1	-0.1	-0.6	0.2	-5.4
ILE	347	0.1	-0.1	-0.6	0.2	-3.4
CYX	380	0.2	0.2	-2.3	-0.3	-0.4
SER	360	0.2	0.3	0.6	0.9	-7.1
TYR	531	0.3	0.2	-1.8	0.1	-5.7
PRO	439	0.7	0.8	1.1	2.7	0.8
GLY	385	0.7	1.1	1.8	2.5	0.5
GLY	384	0.8	1.3	2.9	1.3	2.4
CYX	320	0.8	-0.1	-0.2	-0.7	-0.6
VAL	346	0.8	0.8	0.7	0.5	1.4
HSE	381	1.0	1.9	-1.8	-0.8	-3.2
VAL	444	1.5	2.8	0.3	0.5	0.5
TYR	441	2.0	3.1	-5.0	-4.6	-0.1
SER	387	2.0	4.1	0.4	0.9	-0.7
ILE	446	2.2	1.3	-0.5	0.7	0.6
PRO	436	2.9	-5.5	1.2	0.6	0.0
PHE	324	3.2	2.4	0.2	-0.1	-0.7
GLU	361	65.7	64.1	48.1	62.9	98.6
GLU	424	79.6	100.8	103.3	108.3	92.6

Table S4-27 – Comparison of the arginine and lysine interactions with the five binding modes. Note that the best interactions differ for each binding mode. Ordered by nonbond energy for Mode 1. For instance, R344 contributes much more strongly to mode 1 than to mode 2. Similarly, R420 contributes much more strongly to mode 2 than to mode 1. Mutations targeting mode-specific residues could be used to identify which mode is correct.

Res	Num	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5
ARG	440	-174.5	-167.1	-128.4	-137.4	-57.1
ARG	363	-137.7	-120.6	-57.8	-127.1	-138.0
ARG	309	-128.0	-98.4	-40.8	-45.4	-38.8
LYS	434	-125.7	-142.9	-48.7	-61.4	-54.9
ARG	344	-120.9	-71.5	-50.8	-115.5	-143.2
LYS	378	-85.9	-130.6	-167.5	-175.6	-92.9
ARG	420	-59.6	-115.5	-149.5	-72.9	-162.5
ARG	391	-57.7	-69.6	-139.1	-158.2	-56.5
ARG	408	-56.3	-55.9	-46.9	-55.7	-124.8
ARG	478	-40.6	-47.1	-120.7	-48.4	-121.1

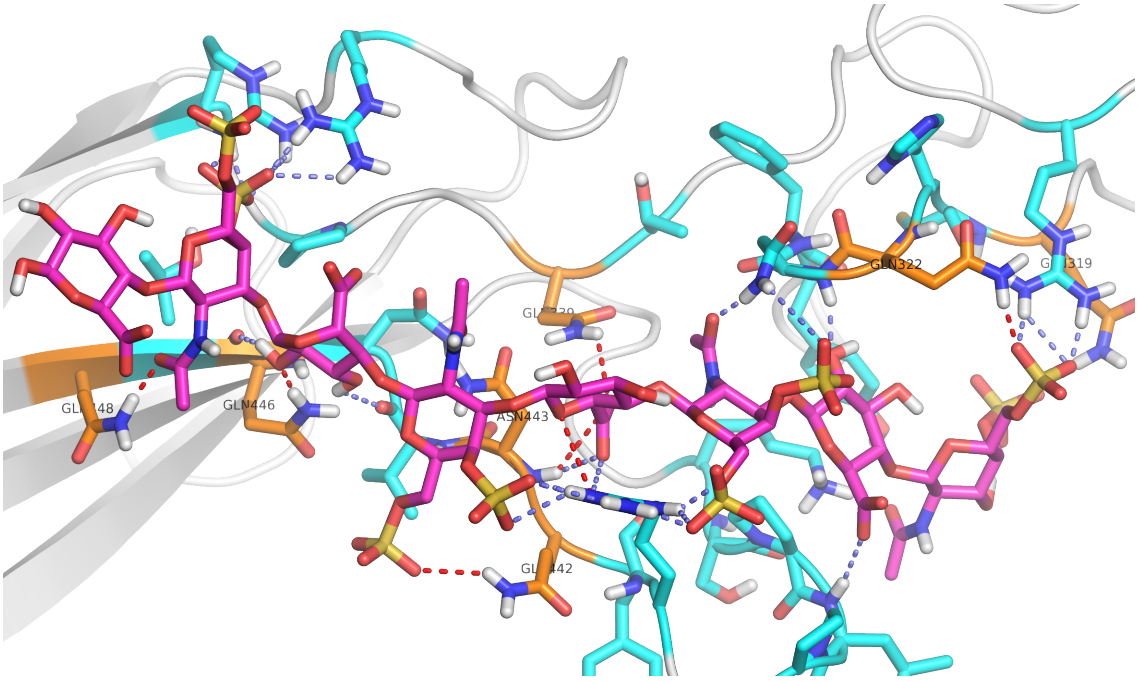

Figure S4-53 – Structure with proposed mutations for Mode 1: T319Q, N322Q, V339Q, A442Q, A443N, I446Q, and T448Q. We predict that this set of mutations for Mode 1 improves binding energy by 66.03 kcal/mol, or 16.5% better than binding to the wild-type.

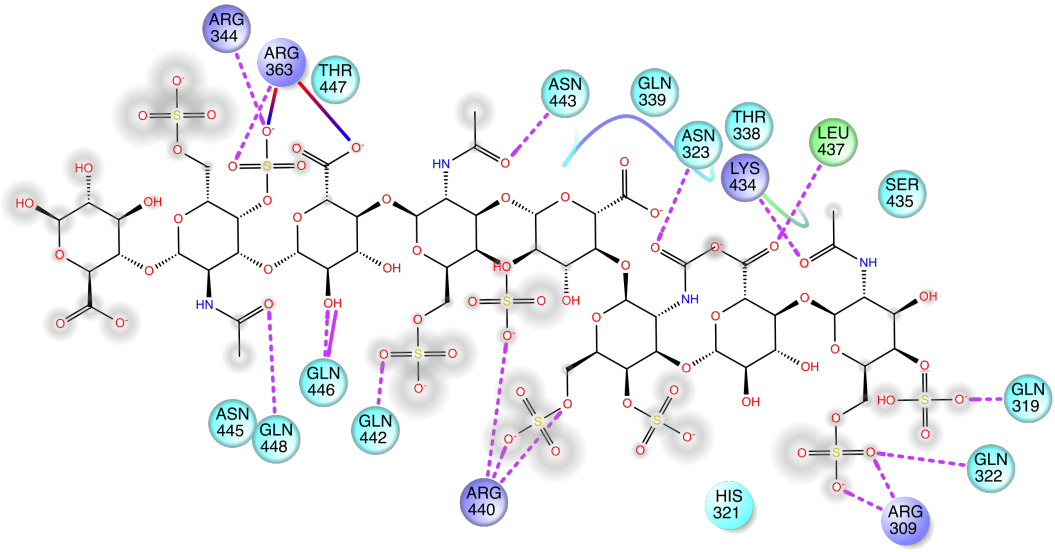

Figure S4-54 – Pharmacophore with proposed mutations for Mode 1: T319Q, N322Q, V339Q, A442Q, A443N, I446Q, and T448Q. We predict that this set of mutations for Mode 1 improves binding energy by 66.03 kcal/mol, or 16.5% better than binding to the wild-type.

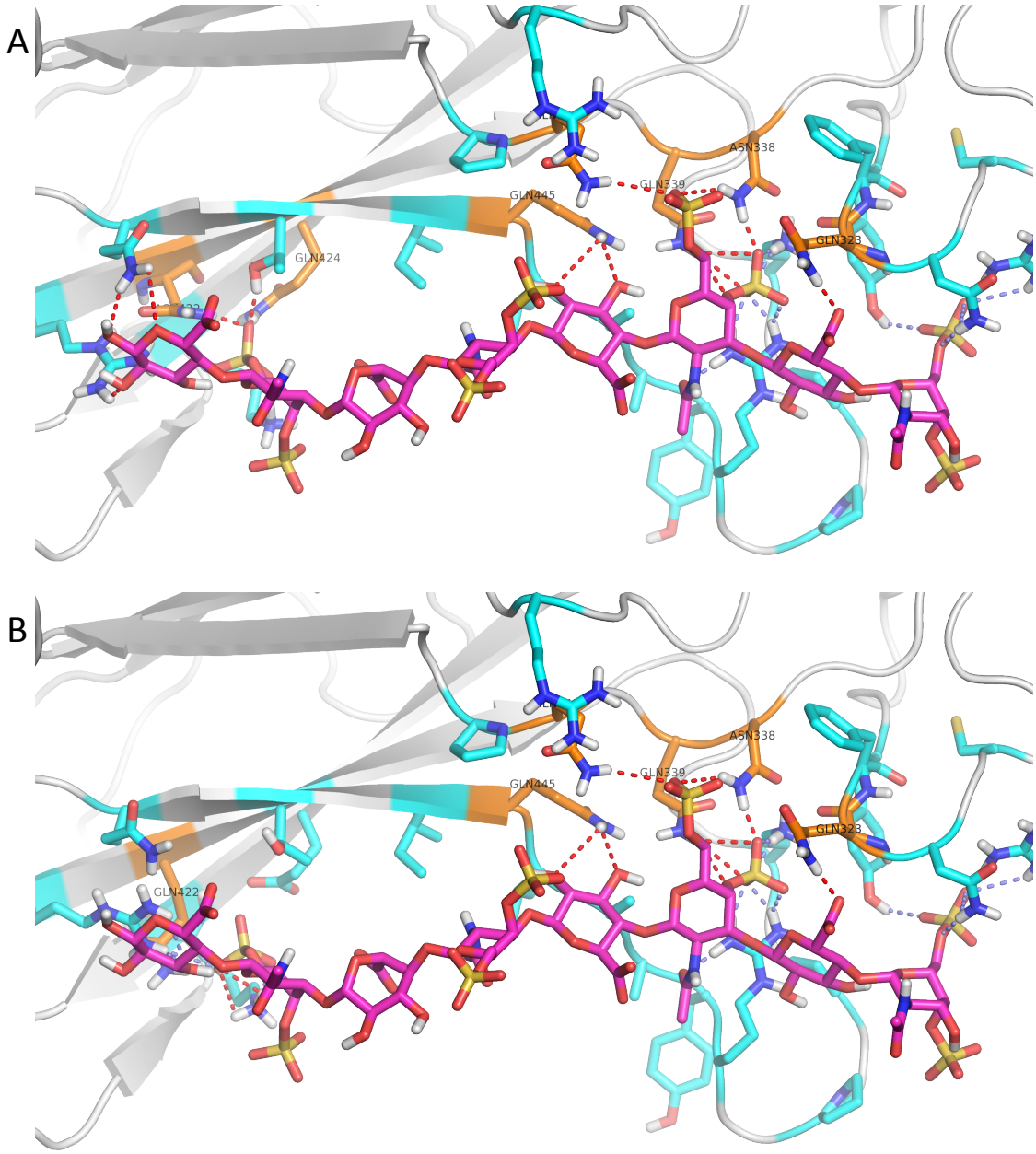

Figure S4-55 – Structures with proposed mutations for Mode 2. (A) Mutations: N323Q, T338N, V339Q, S341Q, T422Q, and N445Q. This improves binding energy by 46.64 kcal/mol or 11.4% better than the wild-type. (B) Adds the E424Q mutant, resulting in an improvement of 163.84 kcal/mol or 40%.

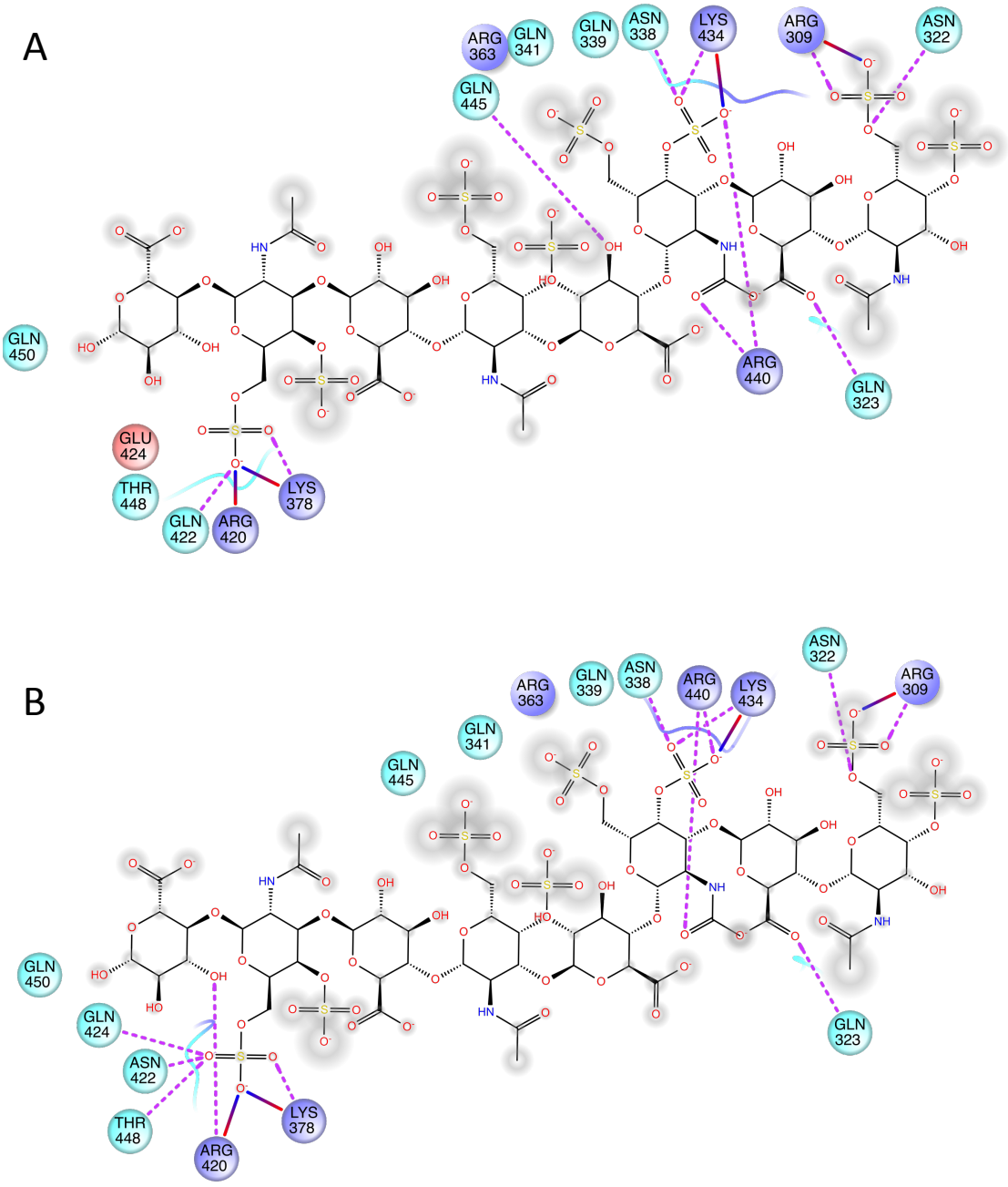

Figure S4-56 – Pharmacophores for proposed mutations for Mode 2. (A) Mutations: N323Q, T338N, V339Q, S341Q, T422Q, and N445Q. This improves binding energy by 46.64 kcal/mol or 11.4% better than the wild-type. (B) Adds the E424Q mutant, resulting in an improvement of 163.84 kcal/mol or 40%.

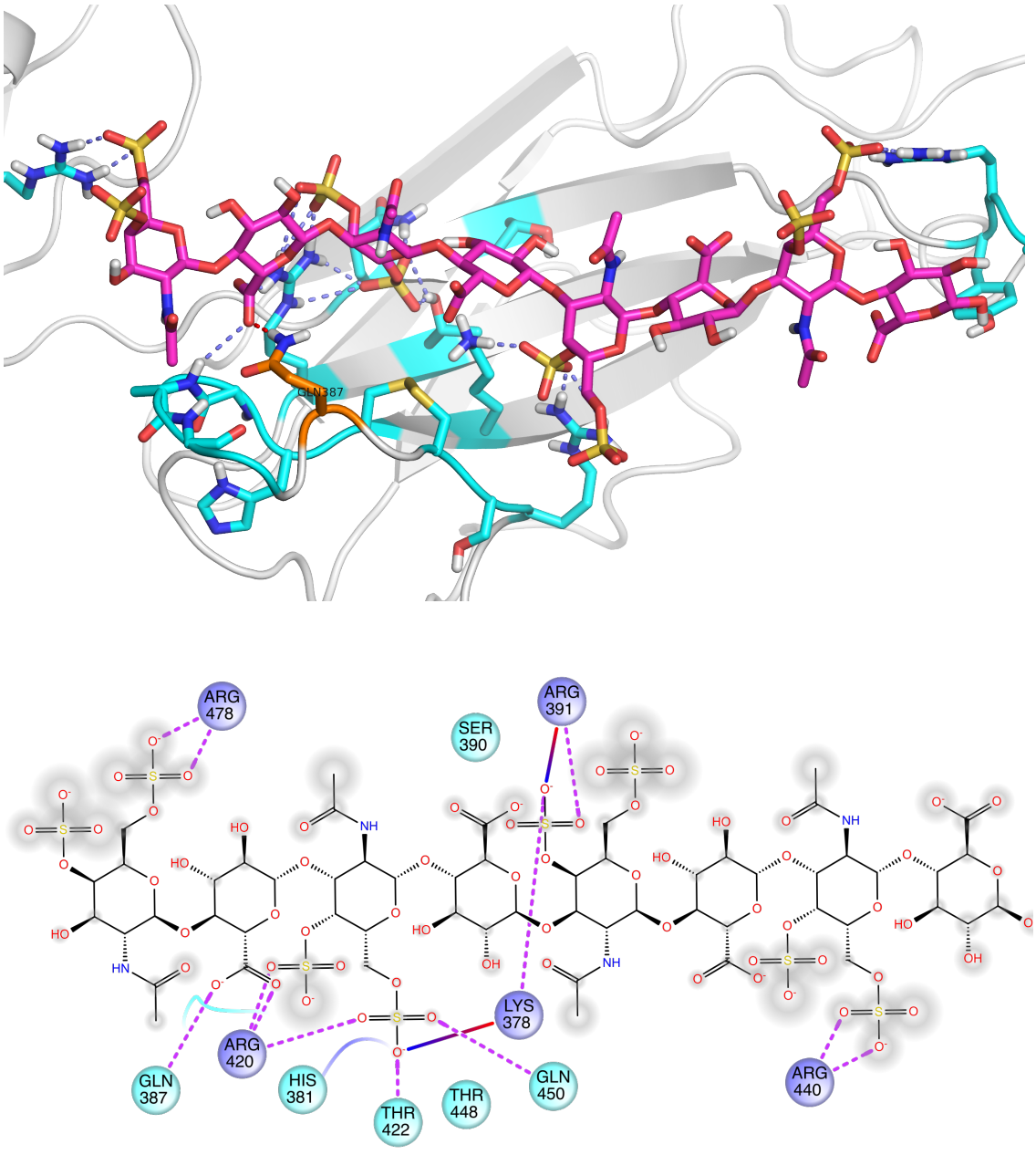

Figure S4-57 – Structure and pharmacophore for Mode 3 mutation: S387Q. This improves binding energy by 7.84 kcal/mol or 1.53% better than the wild-type.

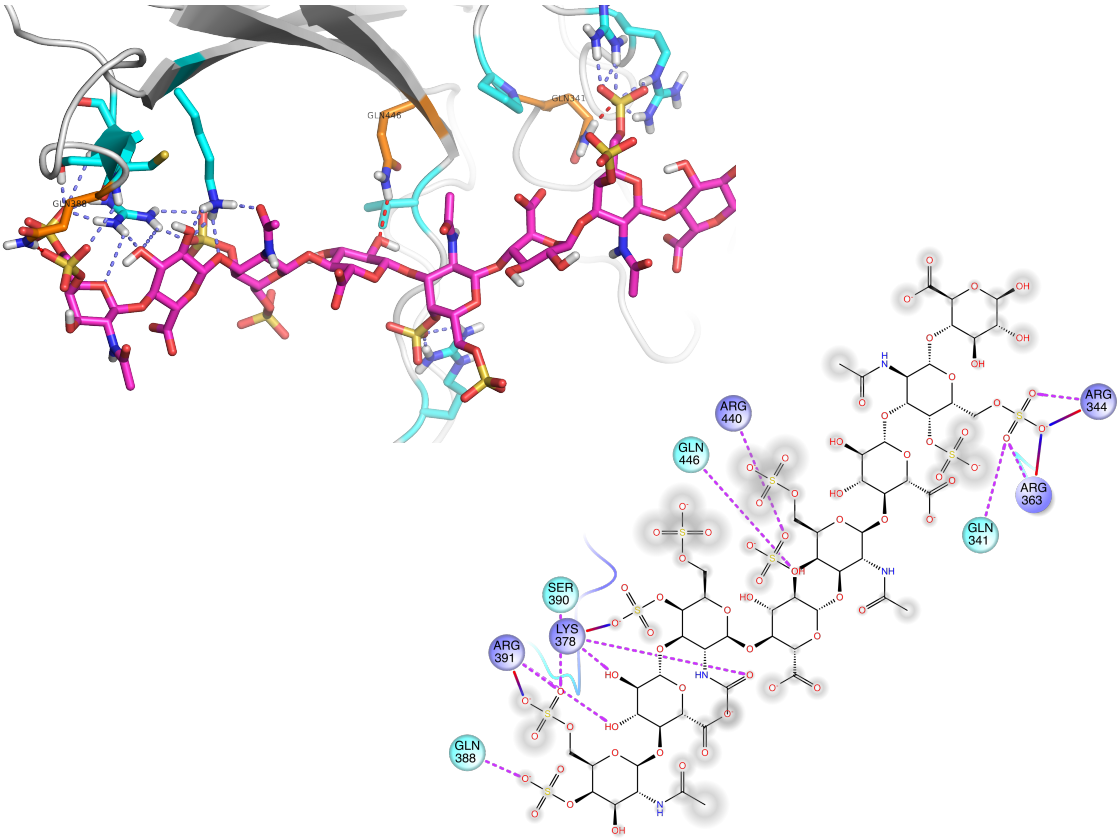

Figure S4-58 – Structure and pharmacophore for Mode 4 mutations: S341Q, A388Q, I446Q. This improves binding energy by 25.71 kcal/mol or 6.43% better than the wild-type.

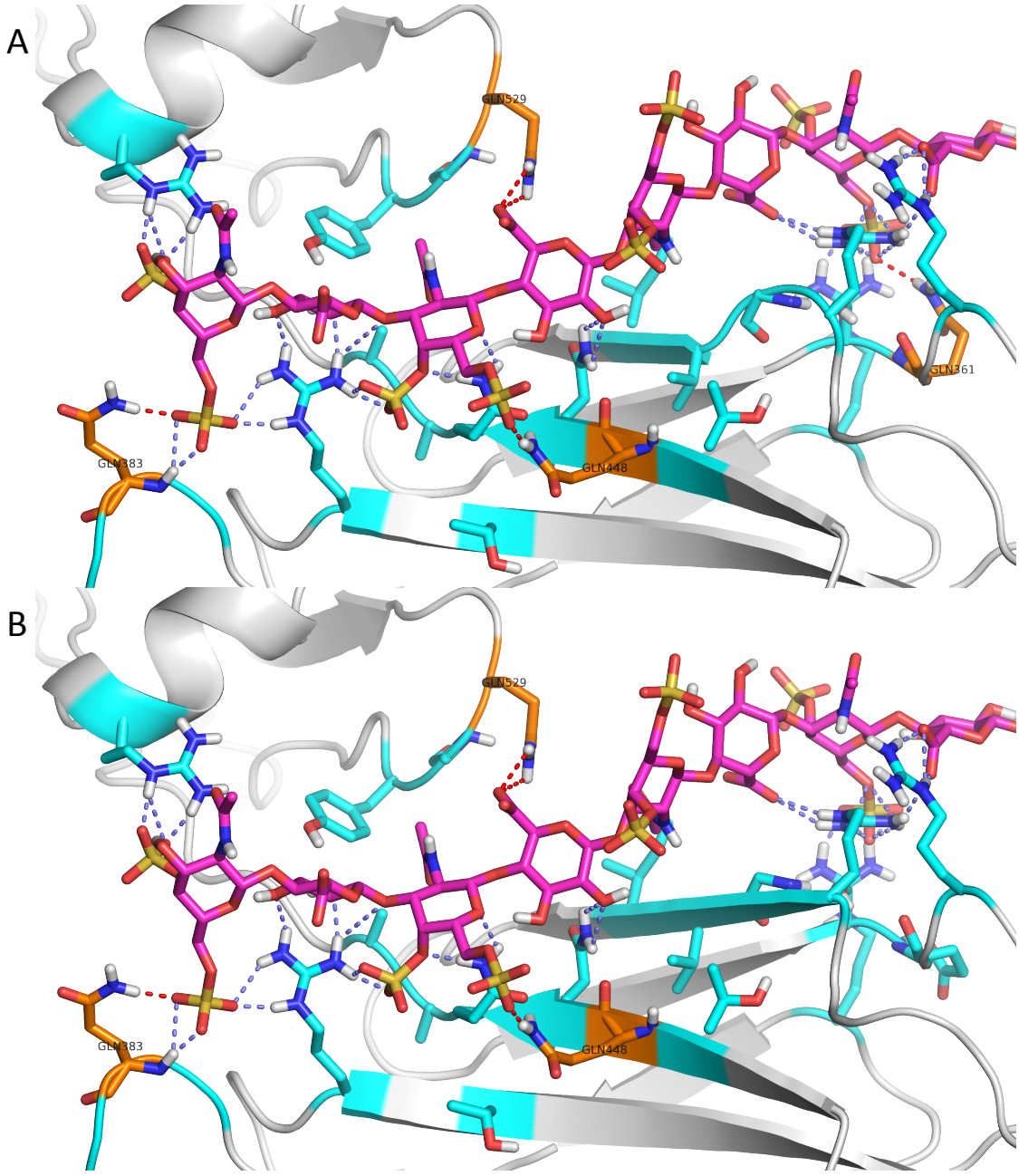

Figure S4-59 – Structures with proposed mutations for Mode 5. (A) Mutations: A383Q, T448Q, A529Q. This improves binding energy by 29.11 kcal/mol or 8% better than the wild-type. (B) Adds the E361Q mutant, resulting in an improvement of 133.82 kcal/mol or 37%.

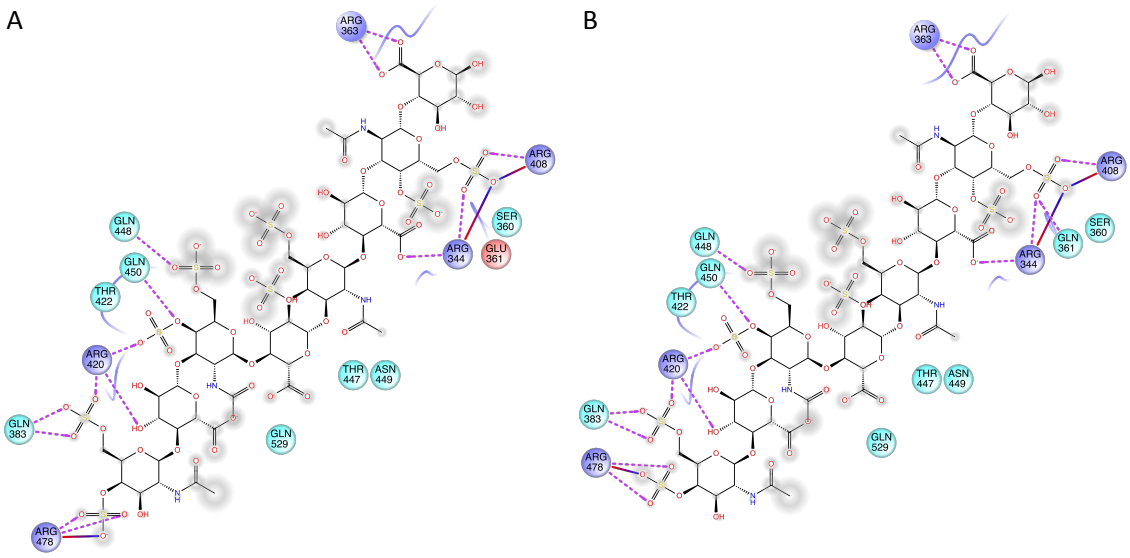

Figure S4-60 – Pharmacophores with proposed mutations for Mode 5. (A) Mutations: A383Q, T448Q, A529Q. This improves binding energy by 29.11 kcal/mol or 8% better than the wild-type. (B) Adds the E361Q mutant, resulting in an improvement of 133.82 kcal/mol or 37%.

Table S4-28 - Binding energies for all mutations to glutamine that improve binding energy for the best pose in mode 1. Note that most mutations do not make new hydrogen bonds (highlighted in red). The increase in binding energy for those mutations can be attributed to Coulomb energy. We only wish to use mutants that make new contacts with the ligand (highlighted in green).

residue	energy	rel. to wt	% incr.	new hbond?
E361	-477.71	-76.57	16.03	no
E424	-473.64	-72.49	15.18	no
T448	-419.91	-18.77	3.93	yes
V339	-414.79	-13.65	2.86	yes
I446	-413.62	-12.48	2.61	yes
A442	-413.11	-11.97	2.50	yes
N445	-412.56	-11.42	2.39	yes
T319	-412.34	-11.20	2.34	yes
L437	-408.68	-7.54	1.58	no
A529	-407.92	-6.78	1.42	no
A443	-406.67	-5.53	1.16	no
A452	-406.09	-4.95	1.04	no
N323	-405.37	-4.23	0.89	yes
A383	-405.12	-3.98	0.83	no
N449	-404.09	-2.95	0.62	no
T338	-403.90	-2.76	0.58	no
I347	-403.35	-2.20	0.46	no
S435	-403.08	-1.94	0.41	no
S341	-402.80	-1.66	0.35	no
T422	-402.79	-1.64	0.34	no
Y531	-402.41	-1.27	0.27	no
F324	-402.13	-0.99	0.21	no
H381	-401.95	-0.81	0.17	no
N322	-401.93	-0.78	0.16	yes
Y441	-401.87	-0.73	0.15	no
W359	-401.70	-0.56	0.12	no
S360	-401.61	-0.47	0.10	no
V444	-401.52	-0.38	0.08	no
wt	-401.14			

Table S4-29 – Binding energies in kcal/mol for different sets of mutations for best pose in binding mode 1. Sets are ranked by binding energy. Set 8.2 (66 kcal/mol or 16.5% improvement in binding energy) is the best set where all mutated residues make a new contact with the ligand.

set	round	energy	rel. to wt	% incr.	mutations								
set7	1	-476.91	-75.77	18.89	Q319	Q323	Q322	Q339	Q442	N443	Q445	Q446	Q448
set5	1	-472.46	-71.32	17.78	Q319	Q323	Q322	Q339	Q442		Q445	Q446	Q448
set8	1	-471.38	-70.24	17.51	Q319		Q322	Q339	Q442	N443	Q445	Q446	Q448
set3	1	-471.34	-70.20	17.50	Q319	Q323		Q339	Q442	N443	Q445	Q446	Q448
set8.2	2	-467.18	-66.03	16.46	Q319		Q322	Q339	Q442	N443		Q446	Q448
set2	1	-464.81	-63.66	15.87	Q319			Q339	Q442		Q445	Q446	Q448
set7.2	2	-463.32	-62.17	15.50	Q319	Q323	Q322	Q339	Q442	N443		Q446	Q448
set3.2	2	-462.70	-61.55	15.34	Q319	Q323		Q339	Q442	N443		Q446	Q448
set4	1	-462.00	-60.86	15.17	Q319			Q339	Q442	N443	Q445	Q446	Q448
set1	1	-460.38	-59.24	14.77	Q319	Q323		Q339	Q442		Q445	Q446	Q448
set4.2	2	-459.85	-58.70	14.63	Q319			Q339	Q442	N443		Q446	Q448
set6	1	-459.00	-57.86	14.42	Q319		Q322	Q339	Q442		Q445	Q446	Q448
set6.2	2	-448.17	-47.03	11.72	Q319		Q322		Q442			Q446	Q448
set5.2	2	-446.28	-45.14	11.25	Q319	Q323	Q322		Q442			Q446	Q448
set2.2	2	-444.95	-43.81	10.92	Q319				Q442			Q446	Q448
set1.2	2	-435.31	-34.17	8.52	Q319	Q323			Q442			Q446	Q448
wt		-401.14			T319	N323	N322	V339	A442	A443	N445	I446	T448

Table S4-30 - Binding energies for all mutations to glutamine that improve binding energy for the best pose in mode 2. Note that most mutations do not make new hydrogen bonds (highlighted in red). The increase in binding energy for those mutations can be attributed to Coulomb energy. We only wish to use mutants that make new contacts with the ligand (highlighted in green).

residue	energy	rel. to wt	% incr.	new hbond?
E424	-525.91	-117.08	22.26	yes
E361	-468.11	-59.28	11.27	no
V339	-424.97	-16.14	3.07	yes
T422	-424.27	-15.44	2.94	yes
T338	-422.96	-14.13	2.69	yes
A442	-417.47	-8.64	1.64	no
N445	-416.27	-7.44	1.42	yes
S435	-415.01	-6.18	1.17	no
Y325	-414.43	-5.60	1.06	no
I446	-412.76	-3.93	0.75	no
T319	-412.71	-3.88	0.74	no
N323	-411.80	-2.97	0.56	yes
Y441	-411.07	-2.24	0.43	no
I347	-410.22	-1.39	0.26	no
S341	-409.75	-0.92	0.17	yes
A388	-409.66	-0.83	0.16	no
Y531	-409.30	-0.47	0.09	no
S360	-408.96	-0.13	0.02	no
A452	-408.86	-0.03	0.00	no
wt	-408.83			

Table S4-31 - Binding energies in kcal/mol for different sets of mutations to best pose from binding mode 2. Sets are ranked by binding energy. Set 4 (163.5 kcal/mol or 40% improvement in binding energy) and Set 11 (46.6 kcal/mol or 11.4% improvement in binding energy) are both selected for mode 2 due to the presence of residue E424. It is interesting to find E424 in such close proximity to a negatively charged ligand like CS-E and we are unsure of what other role E424 may be playing in the protein. Thus, mutations to E424 may have unexpected consequences, even with a relatively close mutant such as glutamine.

set	energy	rel. to wt	%incr	mutations						
set4	-572.31	-163.48	39.99	Q323	N338	Q339	Q341	N422	Q424	Q445
set8	-571.53	-162.70	39.80		N338	Q339	Q341	N422	Q424	Q445
set2	-571.27	-162.44	39.73	Q323	Q338	Q339	Q341	N422	Q424	Q445
set6	-568.59	-159.76	39.08		Q338	Q339	Q341	N422	Q424	Q445
set3	-546.79	-137.96	33.74	Q323	N338	Q339	Q341	Q422	Q424	Q445
set1	-544.24	-135.41	33.12	Q323	Q338	Q339	Q341	Q422	Q424	Q445
set7	-542.07	-133.24	32.59		N338	Q339	Q341	Q422	Q424	Q445
set5	-540.38	-131.54	32.18		Q338	Q339	Q341	Q422	Q424	Q445
set11	-455.47	-46.64	11.41	Q323	N338	Q339	Q341	Q422		Q445
set9	-455.03	-46.20	11.30	Q323	Q338	Q339	Q341	Q422		Q445
set10	-451.95	-43.12	10.55	Q323	Q338	Q339	Q341	N422		Q445
set13	-448.74	-39.91	9.76		Q338	Q339	Q341	Q422		Q445
set15	-448.39	-39.56	9.68		N338	Q339	Q341	Q422		Q445
set12	-447.20	-38.37	9.38	Q323	N338	Q339	Q341	N422		Q445
set14	-443.97	-35.14	8.59		Q338	Q339	Q341	N422		Q445
set16	-441.47	-32.64	7.98		N338	Q339	Q341	N422		Q445
wt	-408.83			N323	T338	V339	S341	T422	E424	N445

Table S4-32 - Binding energies for all mutations to glutamine that improve binding energy for the best pose in mode 3. Note that only *one* residue was able to make a new contact with the ligand. This is a strong indicator that binding mode 3 is not a reliable result. The increase in binding energy for those mutations can be attributed to Coulomb energy.

residue	energy	rel. to wt	% incr.	new hbond?
E424	-511.04	-104.87	20.52	no
E361	-471.32	-65.16	12.75	no
I446	-419.60	-13.44	2.63	no
T338	-415.66	-9.50	1.86	no
A388	-415.12	-8.96	1.75	no
V346	-415.10	-8.94	1.75	no
S435	-414.11	-7.95	1.56	no
S387	-414.00	-7.84	1.53	yes
T447	-413.63	-7.47	1.46	no
S390	-413.47	-7.31	1.43	no
I347	-413.13	-6.97	1.36	no
Y325	-412.65	-6.48	1.27	no
N445	-412.47	-6.31	1.23	no
Y531	-411.85	-5.69	1.11	no
A529	-411.72	-5.56	1.09	no
S360	-411.18	-5.02	0.98	no
F324	-410.95	-4.78	0.94	no
L437	-410.48	-4.32	0.85	no
S341	-408.40	-2.24	0.44	no
T319	-408.29	-2.12	0.42	no
A443	-408.18	-2.02	0.39	no
N323	-408.08	-1.91	0.37	no
V339	-407.77	-1.61	0.31	no
V444	-407.03	-0.87	0.17	no
W359	-406.77	-0.61	0.12	no
N322	-406.33	-0.16	0.03	no
Y441	-406.20	-0.04	0.01	no
wt	-406.16			

Table S4-33 - Binding energies for all mutations to glutamine that improve binding energy for the only pose in mode 4. Note that most mutations do not make new hydrogen bonds (highlighted in red). The increase in binding energy for those mutations can be attributed to Coulomb energy. We only wish to use mutants that make new contacts with the ligand (highlighted in green).

residue	energy	rel. to wt	% incr.	new hbond?
E424	-495.76	-95.77	19.32	no
E361	-465.02	-65.03	13.12	no
A443	-419.46	-19.46	3.93	no
A388	-412.03	-12.04	2.43	yes
S341	-411.02	-11.03	2.22	yes
N449	-410.33	-10.33	2.08	no
Y531	-410.20	-10.21	2.06	no
I446	-407.81	-7.82	1.58	yes
T338	-407.64	-7.64	1.54	no
I347	-405.08	-5.08	1.03	no
W359	-402.47	-2.48	0.50	no
T422	-402.32	-2.33	0.47	no
Y441	-402.25	-2.25	0.45	no
S435	-401.01	-1.01	0.20	no
wt	-399.99			

Table S4-34 - Binding energy in kcal/mol for the only set of mutations to the only pose from binding mode 4. Mutation results in 25.7 kcal/mol or 6.4% improvement in binding energy.

set	energy	rel. to wt	% incr.	mutations		
set1	-425.70	-25.71	6.43	Q341	Q388	Q446
wt	-399.99			S341	A388	I446

Table S4-35 - Binding energies for all mutations to glutamine that improve binding energy for the only pose in mode 5. Note that most mutations do not make new hydrogen bonds (highlighted in red). The increase in binding energy for those mutations can be attributed to Coulomb energy. We only wish to use mutants that make new contacts with the ligand (highlighted in green).

residue	energy	rel. to wt	% incr.	new hbond?
E361	-464.85	-103.36	22.23	yes
E424	-462.29	-100.79	21.68	no
A383	-378.48	-16.98	3.65	yes
T448	-369.95	-8.46	1.82	yes
A388	-369.66	-8.17	1.76	no
A529	-369.62	-8.13	1.75	yes
S390	-369.57	-8.08	1.74	no
I347	-368.36	-6.87	1.48	no
H381	-366.63	-5.13	1.10	no
V444	-365.92	-4.42	0.95	no
S435	-365.33	-3.83	0.82	no
T422	-364.86	-3.37	0.72	no
Y325	-364.47	-2.98	0.64	no
T338	-362.17	-0.67	0.14	no
N323	-361.95	-0.45	0.10	no
I446	-361.70	-0.20	0.04	no
wt	-361.50			

Table S4-36 - Binding energy in kcal/mol for mutations to the only pose from binding mode 5. Set 1, which includes a mutation to E361, improves binding energy by 133.8 kcal/mol or 37%. Set 2, which skips the mutation to E361, improves binding by 29.1 kcal/mol or 8.1%. Similarly to binding mode 2, it is interesting to find E361 (a different glutamic acid) in such close proximity to a negatively charged ligand like CS-E and we are unsure of what other role E361 may be playing in the protein. Thus, mutations to E361 may have unexpected consequences, even with a relatively close mutant such as glutamine.

set	energy	rel. to wt	% incr.	mutations			
set1	-495.32	-133.82	37.02	Q361	Q383	Q448	Q529
set2	-390.61	-29.11	8.05		Q383	Q448	Q529
wt	-361.50			E361	A383	T448	A529