Molecular simulations of carbohydrate–protein complexes

I Generation and validation of a free-energy model for carbohydrate binding
II Simulating the binding of Lewis-type ligands to DC-SIGN
III Developing a molecular modeling toolbox for medicinal chemists

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät der

Universität Basel

von

Sameh Mansour Abbas Eid

aus Gharbia, Ägypten

Basel, 2013

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
Auf Antrag von:

Prof. Dr. A. Vedani, Departement Pharmazeutische Wissenschaften, Universität Basel
Prof. Dr. F. M. Böckler, Pharmazeutisches Institut, Eberhard Karls Universität Tübingen

Basel, den 23. April 2013

Prof. Dr. Jörg Schibler
Dekan

The greatest challenge to any thinker is stating the problem in a way that will allow a solution
Bertrand Russell

My definition of an expert in any field is a person who knows enough about what's really going on to be scared
Phillip James Plauger

If the facts don't fit the theory, change the facts
Albert Einstein

There are no such things as applied sciences, only applications of science
Louis Pasteur

If mankind minus one were of one opinion, then mankind is no more justified in silencing the one than the one - if he had the power - would be justified in silencing mankind
John Stuart Mill

Data is not information, information is not knowledge, knowledge is not understanding, understanding is not wisdom
Clifford Stoll

How much easier it is to be critical than to be correct
Benjamin Disraeli

You've achieved success in your field when you don't know whether what you're doing is work or play
Warren Beatty

Acknowledgements

First and foremost I would like to express my deepest gratitude to Prof. Angelo Vedani for giving me the opportunity to join his research group and for his continuous assistance, understanding, and guidance throughout the entire study. In his team, and under his supervision, I learnt how to be a better researcher, a questioning and methodical scientist, and most importantly a great teacher. In my culture, we were brought up believing that the most valuable gift a man could present to a fellow man is *knowledge*. Therefore, I am, and will always be, in his debt for the priceless knowledge he passed unto me, in scientific as well as life experiences.

I am also very grateful to my co-supervisor Prof. Beat Ernst for giving me the chance to collaborate with his vibrant research team in such a wonderful scientific setting, for his invaluable insight and motivation, and for proofreading manuscripts. I shall never forget his advice to always strive for the best quality and never settle for less.

I owe special thanks to my colleague and dear friend in the group of molecular modeling, Adam Zalewski. He was always by my side in the highs and lows, the successful job completions and the system crashes, always having something smart, funny, and mood-lifting to say. My Ph.D. experience would have been very different if it were not for him.

My special thanks extend to all members of the molecular modeling group, Dr. Martin Smieško, Dr. Gianluca Rossato, Christoph Sager, and Zhenquan Hu for the amazing work environment, and for the most helpful answers and advice they gave me in innumerable situations. Assistance provided by Nour Saleh in collecting and reviewing crystallographic and biochemical data during his three-month internship is greatly appreciated. The contribution of Biographics Laboratory 3R in the development of the *VirtualDesignLab* is gratefully acknowledged.

I am particularly grateful to Dr. Mathias Wittwer and Dr. Morena Spreafico for their warm and friendly help on countless occasions during my early stay in Basel.

I would like to thank the following members of the Department of Pharmaceutical Sciences, University of Basel (formerly IMP) for the fruitful collaborations and the friendly working atmosphere: Katharina Mayer, Meike Scharenberg, Said Rabbani, Arjan Odedra, Brian Cutting, Daniela Abgottspon, Roland Preston, Katrin Lemme, and Jacqueline Bezençon. I would like to express my appreciation to Gabi Lichtenhan and Claudia Huber for their most valued assistance with all the administrative issues and tedious paper work.

There are hardly words that can express my love and appreciation for my wife and kids, for they are the source of all my motivation and happiness. The efforts my wife had to spend to take care of the kids and convince them, somehow, to jump around a little less, and above

all to be patient with me staring at the screen and banging the keys of my laptop for hours are just miraculous. Frankly, I cannot think of a way to repay this debt.

I would not have been the man I am, nor have I achieved what I have achieved today if it were not for the unconditional love, nurture, support, and prayers of my family; my father, mother, brother, sister, and my parents-in-law. May Allah bless you all!

Thanks must go to the graceful city of Basel, Switzerland. It has indeed become my second home with many unforgettable memories and wonderful festivals and parades.

Finally, I would like to gratefully acknowledge the financial support from the Swiss National Science Foundation and the Department of Pharmaceutical Chemistry, University of Basel.

Abstracts

Generation and validation of a free-energy model for carbohydrate binding. Carbohydrates play a key role in a variety of physiological and pathological processes and, hence, represent a rich source for the development of novel therapeutic agents. Being able to predict binding mode and binding affinity is an essential, yet lacking, aspect of the structure-based design of carbohydrate-based ligands. To this end, we assembled a diverse data set of 316 carbohydrate–protein crystal structures with known binding affinity. We evaluated the prediction accuracy of a large collection of well-established scoring and free-energy functions, as well as empirical combinations thereof. Unfortunately, the tested functions were not capable of reproducing carbohydrates binding affinities in our complexes. To simplify the complex free-energy surface of carbohydrate–protein systems, we classified the studied complexes according to the topology and solvent exposure of the carbohydrate-binding site into five distinct categories. A free-energy model based on the proposed classification scheme reproduced binding affinities in the carbohydrate data set with an r^2 of 0.69 and root-mean-squared-error of 1.36 kcal/mol. The improvement in model performance underlines the significance of the differences in the local micro-environments of carbohydrate-binding sites and demonstrates the usefulness of calibrating free-energy functions individually according to binding-site topology and surface exposure.

Simulating the binding of Lewis-type ligands to DC-SIGN. Dendritic cells (DCs) have the function of presenting antigens to other processing cells of the immune system, particularly T-cells. DC-SIGN (DC-specific intercellular adhesion molecule-3-grabbing non-integrin) is one of the major receptors on DCs involved in the uptake of pathogens and has gained increasing interest over the last decade as it is crucially involved in infections caused by HIV-1, Ebola virus, *Mycobacterium tuberculosis*, and various other pathogens. High-mannosylated *N*-glycans or L-Fuc-containing trisaccharide motifs such as the Lewis (Le) blood group antigens Le^a and Le^x, which are surface components of these microorganisms, mediate binding to DC-SIGN. Crystallographic data for DC-SIGN in complex with a Le^x-containing pentasaccharide suggest that the terminal sugar residues, L-Fuc and D-Gal, are predominantly involved in binding. We elucidated the interaction of DC-SIGN with Le^a and Le^x bearing two different aglycones. Binding assays together with STD NMR analysis, molecular modeling and mutagenesis studies revealed distinct binding modes dependent on the nature of the aglycone. Introduction of phenyl aglycones at the Le trisaccharides offers the establishment of an additional hydrophobic contact with Phe313 in the binding site of DC-SIGN, which entails a switch of the binding mode. Based on this information a new series of DC-SIGN antagonists can be designed.

Developing a molecular modeling toolbox for medicinal chemists. In the current era of high-throughput drug discovery and development, molecular modeling has become an indispensable tool for identifying, optimizing and prioritizing small-molecule drug candidates. The required background in computational chemistry and the knowledge of how to handle the complex underlying protocols, however, might keep medicinal chemists from routinely using *in silico* technologies. Our objective is to encourage those researchers to exploit existing modeling technologies more frequently through easy-to-use graphical user interfaces. In this account, we present two innovative tools (which we are prepared to share with academic institutions) facilitating computational tasks commonly utilized in drug discovery and development: (1) the *VirtualDesignLab* estimates the binding affinity of small molecules by simulating and quantifying their binding to the three-dimensional structure of a target protein; and (2) the *MD Client* launches molecular dynamics simulations aimed at exploring the time-dependent stability of ligand–protein complexes and provides residue-based interaction energies. This allows medicinal chemists to identify sites of potential improvement in their candidate molecule. As a case study, we present the application of our tools towards the design of novel antagonists for the FimH adhesin.

The Author’s Contribution

Generation and validation of a free-energy model for carbohydrate binding

This project in term of planning, simulations, development, and analysis was completely carried out by me, except for the assistance of Nour Saleh (during 3-month internship) in procuring the carbohydrate–protein affinity database.

Simulating the binding of Lewis-type ligands to DC-SIGN

I performed all the molecular simulations of this study, including docking and molecular dynamics simulations.

Developing a molecular modeling toolbox for medicinal chemists

The development, testing, and application of the *MD Client* to the study of selected FimH ligands were completely carried out by me. The development of the *VirtualDesignLab* was done by Adam Zalewski, Martin Smieško, and Biographics Laboratory 3R.

Table of Contents

Abbreviations

CDR	Complex descriptors (NeoScore file format)
DoF	Degrees-of-Freedom
EDS	Electron-density map
FEP	Free energy perturbation
GA	Genetic algorithm
HET	PDB residue name used to designate non-standard residues
ITC	Isothermal Titration Calorimetry
LIE	Linear Interaction Energy
MD	Molecular Dynamics
MMFF	Merck Molecular Force Field
MM/GBSA	Molecular Mechanics – Generalized Born/Surface Area model
MM/PBSA	Molecular Mechanics – Poisson Boltzmann/Surface Area model
MOAD	Mother Of All Databases
NMR	Nuclear magnetic resonance
OPLS	Optimized Potentials for Liquid Simulations
PDB	Protein Data Bank
QSAR	Quantitative structure activity relationship
RMSD	Root Mean Squared Deviation
RMSE	Root Mean Squared Error
RRHO	Rigid-rotor harmonic oscillator approximation
SASA	Solvent-accessible surface area
SBD	Structure-based design
TI	Thermodynamic integration

Aim of the thesis

This thesis explores the use of computational methodologies for studying carbohydrate–protein binding and its potential applications in drug design. The increasing numbers of structurally and functionally characterized carbohydrate-binding proteins provide the basis for structure-based design tools, e.g. docking and scoring, virtual screening, and *de novo* design, and could thereby accelerate rational design and optimization of carbohydrate leads. Moreover, carbohydrate–protein complexes are of highly dynamic nature. Therefore, proper sampling of their conformational space, e.g. by molecular dynamics simulations, could provide valuable clues for medicinal chemists and guide the lead optimization process. Employment of molecular modeling tools in synergy with experimental techniques is the key to successful design and development of novel carbohydrate-based therapeutics.

This thesis is organized in three separate parts:

Generation and validation of a free-energy model for carbohydrate binding. Despite the great advances in molecular modeling methodologies, prediction of carbohydrate binding affinity from structural information remains largely unsolved. Here, we assembled and verified the experimental binding affinities of a large data set of carbohydrate–protein complexes with known crystal structures. We performed a thorough analysis of empirical free-energy functions to uncover the potential difficulties in deriving reliable structure–affinity relationships for carbohydrate ligands. We aimed to develop an improved treatment for predicting binding free energies in carbohydrate–protein systems, which can be used in structure-based design applications.

Simulating the binding of Lewis-type ligands to DC-SIGN. DC-SIGN (Dendritic Cell-specific intercellular adhesion molecule-3-grabbing non-integrin) is an interesting target for anti-infective treatments as it is involved in infections caused by HIV-1, Ebola virus, *Mycobacterium tuberculosis*, and various other pathogens. Binding assays together with STD NMR analysis, molecular modeling and mutagenesis studies were used to study the interaction of DC-SIGN with Lewis^a and Lewis^x bearing two different aglycones. The improved understanding of structure-dependent binding modes could guide the design of a new series of DC-SIGN antagonists.

Developing a molecular modeling toolbox for medicinal chemists. Employment of molecular modeling techniques to solve drug design problems can be facilitated by medicinal chemist-oriented interfaces to powerful computational tools. In this regard, we developed two innovative utilities targeting commonly required tasks: (1) the *VirtualDesignLab* for simulating and quantifying binding of small molecules to the three-dimensional structure of a target protein; and (2) the *MD Client* for launching molecular dynamics simulations to explore the time-dependent behavior of ligand–protein complexes and calculating residue-based interaction energies.

I. Generation and validation of a free-energy model for carbohydrate binding

I.1. Abstract

Carbohydrates play a key role in a variety of physiological and pathological processes and, hence, represent a rich source for the development of novel therapeutic agents. Being able to predict binding mode and binding affinity is an essential, yet lacking, aspect of the structure-based design of carbohydrate-based ligands. To this end, we assembled a diverse data set of 316 carbohydrate–protein crystal structures with known binding affinity. We evaluated the prediction accuracy of a large collection of well-established scoring and free-energy functions, as well as empirical combinations thereof. Unfortunately, the tested functions were not capable of reproducing carbohydrates binding affinities in our complexes. To simplify the complex free-energy surface of carbohydrate–protein systems, we classified the studied complexes according to the topology and solvent exposure of the carbohydrate-binding site into five distinct categories. A free-energy model based on the proposed classification scheme reproduced binding affinities in the carbohydrate data set with an r^2 of 0.69 and root-mean-squared-error of 1.36 kcal/mol. The improvement in model performance underlines the significance of the differences in the local micro-environments of carbohydrate-binding sites and demonstrates the usefulness of calibrating free-energy functions individually according to binding-site topology and surface exposure.

I.2. Introduction

The design of small-molecule ligands that can bind firmly to their biological target is the heart of rational drug design. The term structure-based design (SBD) is used to describe drug-design studies where the three-dimensional structure of the macromolecular target (typically a protein) is available at high resolution. Structures of macromolecules are either solved by experiments such as X-ray crystallography and NMR spectroscopy, or generated by modeling techniques such as homology modeling. Our knowledge of the structures of functional proteins and other biomolecules is expanding at an unprecedented rate with thousands of new structures being deposited every year into the Protein Data Bank (Bernstein *et al.*, 1977) (Figure I-1). Structure-based design tools such as virtual screening (Wildman, 2012) and *de novo* design (Hartenfeller and Schneider, 2011) play a key role in the initial stages of drug design, particularly lead identification and lead optimization. Many research efforts seek to boost the reliability and efficiency of these methodologies by means of innovative high-speed algorithms, improved theoretical treatments for molecular interactions, as well as exploiting the superior computational capability of high- performance computers. A central issue in this regard is the ability to infer binding affinities from three-dimensional configurations of ligand–protein complexes with acceptable speed and accuracy.

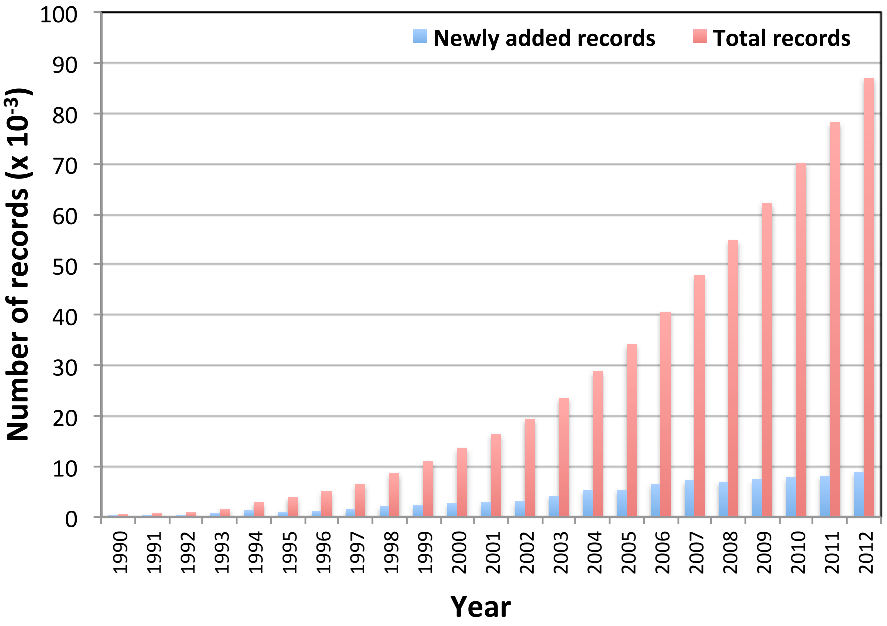


Figure I-1: The total number of records stored in the Protein Data Bank and the number of new records deposited every year (updated Dec 2012).

This thesis is focused on the structure-based design of carbohydrate and glycomimetic ligands, along with the quantification of their binding to macromolecular targets. We will begin our introduction by elucidating the potential of carbohydrates as sources for the design of novel therapeutics. Afterwards, the challenges responsible for slowing the pace of carbohydrate-based drug discovery will be highlighted, particularly those in the area of structure-based design. This is followed by an overview of the thermodynamics principles governing bimolecular associations and the computational methods employed to simulate and quantify ligand–protein recognition and binding. Then, the reported attempts of development of carbohydrate-specific functions for prediction of binding free energy are briefly reviewed. In closing, we outline the strategy employed in this study for construction and validation of predictive models for carbohydrate–protein binding free energy.

I.2.1 Carbohydrates in drug design and development

Carbohydrates are ubiquitous in nature as they represent one of the three main categories of biomolecules, besides proteins and nucleic acids. They are involved in a broad spectrum of pathophysiological processes ranging from protein folding, bacterial adhesion, viral infection, cancer metastasis, inflammatory reactions, cell proliferation, and cell–cell communication (Cummings, 2009). Carbohydrate research has gained considerable momentum in the past decade due to its potentially rewarding applications in therapeutics, diagnosis and vaccine development. The following paragraphs outline some examples for employment of carbohydrates in drug targeting and drug delivery and as novel therapeutic agents.

Drug targeting. Numerous glycoprotein and glycopolymer-based systems have been developed to deliver drugs selectively to their intended site of action, thereby reducing the unwanted side effects and employing smaller dose of the active principle (Davis and Robinson, 2002). Such systems take advantage of the specific nature of carbohydrate–protein interactions and the wide variety of cellular receptors that can be potentially targeted by such systems. This could be particularly useful, for instance, in reducing cytotoxic effects of anti-cancer drugs on non-cancerous cells (Singh *et al.*, 2008). The lectin-directed enzyme activated prodrug therapy (LEAPT) is another example, where sugar patterns were manipulated to control the site of release of sugar-capped prodrugs via a cell- or tissue-specific synthetic enzyme (Garnier *et al.*, 2010). An M-cell targeted oral mucosal immunization against Hepatitis B was successfully achieved via α-L-Fucose-specific lectin as homing device for drug nano-carriers (Mishra *et al.*, 2011). Successful delivery of HIV DNA vaccine particles has been reported using mannose as a DC (Dendritic Cells)-directed ligand (van den Berg *et al.*, 2010). Extensive investigation of mannose-conjugated drug delivery vehicles showed that in many cases such conjugates exhibit high macrophage uptake, good activity, and fewer side effects (Kumar *et al.*, 2006; Verma *et al.*, 2008). As and application, mannose-conjugated solid lipid nanoparticles were used for effective and targeted delivery of antituberculosis drug, rifabutin, to alveolar macrophages (Nimje *et al.*, 2009).

Drug delivery. Conjugation of custom-made glycan epitopes to proteins or biocompatible non-immunogenic polymeric scaffolds produces neoglycoconjugates with purpose-adaptable properties (Yamazaki *et al.*, 2000). Lectins were employed as vehicles for mucosal drug delivery (Clark *et al.*, 2000) and were shown to be useful as natural bioadhesives useful in mucosal vaccine delivery (Baudner and O'Hagan, 2010). Lectin (wheat-germ agglutinin) conjugated micro-particles were used for enhancing oral delivery of insulin due its superior mucoadhesive properties (Kim *et al.*, 2005). Similarly, Gal/GalNAc decorated LPD (liposomes/protamine/DNA) particles demonstrated improved efficiencies and lower toxicity as hepatocellular gene transfer vectors (Lu *et al.*, 2010). Lectin-conjugated nanoparticles improved efficiency of triple therapy (amoxicillin, clarithromycin and omeprazole) in eradicating *H. pylori* infections from gut due to selective release of triple therapy for a longer period of time (Ramteke *et al.*, 2008). Use of galactosylated liposomes as delivery vehicles for azidothymidine significantly reduced its hematopoietic toxicity and enhanced cellular uptake by lectin-bearing macrophages (Garg and Jain, 2006).

Therapeutic applications. Discovery and functional characterization of an increasing number of carbohydrate-related targets and biological pathways have paved the way for the development of several therapeutic agents (Figure I-2). Ernst and Magnani reviewed marketed carbohydrate-derived products in a handful of disease areas (Ernst and Magnani, 2009). Vancomycin, first isolated in1953, is probably the first carbohydrate-containing therapeutic agent. It was originally indicated for the treatment of penicillin-resistant *Staphylococcus aureus* infections (Levine, 2006). Kaplan *et al.* showed that the sugar residues in vancomycin enhance its binding affinity by restricting conformational flexibility of the aglycon part (Kaplan *et al.*, 2001). Another early therapeutic application of carbohydrate derivatives is the use of acarbose, an α-glucosidase inhibitor, for treatment of type 2 diabetes (Hoffmann and Spengler, 1997; Scheen, 1998). Two more products belong to the same pharmacological class of acarbose; miglitol (Scott and Spencer, 2000) and voglibose (Chen *et al.*, 2006).

Furthermore, zanamivir (von Itzstein *et al.*, 1993) and oseltamivir (Kim *et al.*, 1997) are examples of carbohydrate-based drugs with potent anti-influenza activity. The discovery and clinical development of low-molecular-weight heparins, which are glycosaminoglycans, marks a major breakthrough in the field of antithrombotic treatment (Weitz, 1997). Other examples of carbohydrate-based drugs include miglustat for type 1 Gaucher disease (Weinreb *et al.*, 2005), topiramate for epilepsy (Maryanoff *et al.*, 1987), and hyaluronan for osteoarthritis (Puhl and Scharf, 1997). In addition, some glycomimetic drugs are currently in the clinical development stages, e.g. GMI-1070 for treatment of vaso-occlusive crisis of sickle cell disease. Examples presented above barely scratch the surface of the great pharmaceutical potential of carbohydrate-based therapeutics. Research is being actively conducted in several disease areas where carbohydrates play pivotal roles, e.g. inflammatory diseases, neuronal regeneration (Yang and Schnaar, 2008), antitumor vaccination (Liu and Ye, 2012; Ouerfelli *et al.*, 2005; Ragupathi, 1996), and cancer therapy (Salatino *et al.*, 2008).

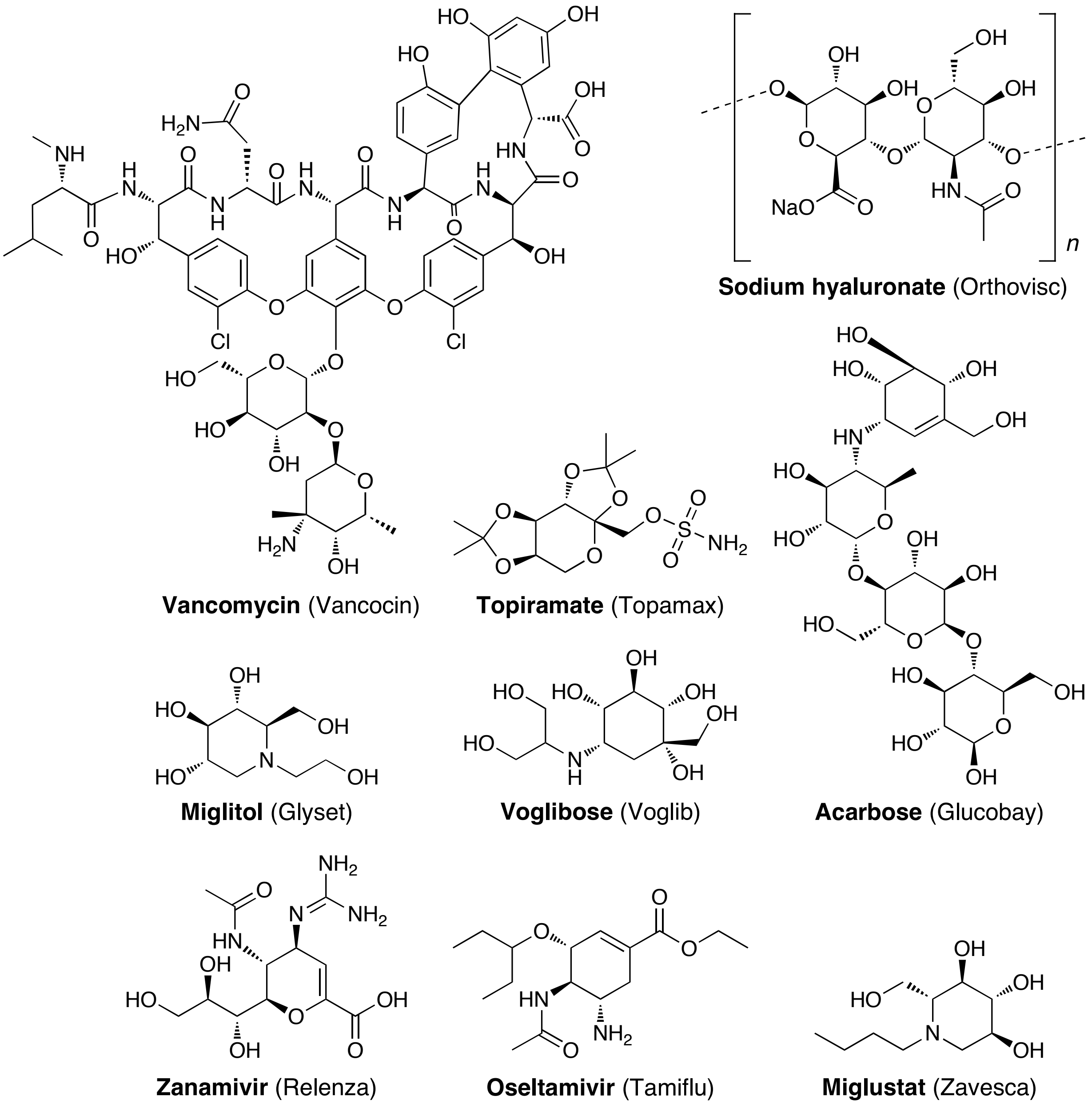


Figure I-2: Examples of carbohydrate-based drugs in the market (trade name in brackets).

Challenges of carbohydrate drug research. Despite the tremendous pharmaceutical potential, only a limited number of carbohydrate-based drugs have reached the market up to date. Carbohydrates are, thus, still considered to be relatively untapped as a promising source for new therapeutic agents (Ernst and Magnani, 2009). Development of carbohydrate-targeting therapeutics is more challenging compared to other chemical classes due to a multitude of factors. Firstly, despite the development of more efficient synthetic routes and separation and analysis technologies (Boltje *et al.*, 2009; Muthana *et al.*, 2009; Zhu and Schmidt, 2009), synthesis of complex carbohydrate structures still requires considerable effort and careful planning, and it might take up to weeks or even months (Galan *et al.*, 2011). Moreover, naturally occurring carbohydrates are rich in polar functional groups and very water soluble, which hampers their absorption by passive diffusion through lipid membranes (Magnani and Ernst, 2009). In addition to such undesirable pharmacokinetic characteristics, bioactive carbohydrates typically have low binding affinities (in the milli- to micro-molar range) towards their targets. However, multivalent binding of natural carbohydrates leads to higher *avidity*, which enables them to perform their biological functions. Thus, the success of rational design of drug-like carbohydrate derivatives requires correct identification and subsequent removal of the unnecessary polar groups from lead structures without compromising any of the pharmacophoric elements essential for recognition (Cipolla *et al.*, 2010; Ernst and Magnani, 2009). Furthermore, carbohydrates present a distinctive set of challenges to contemporary molecular-modeling methodologies due to their unique structural features. The following section investigates the causes in more detail.

I.2.2 Challenges in carbohydrate modeling

In early biomolecular investigations, carbohydrates and carbohydrate–protein interactions were overlooked in favor of peptides, proteins, and nucleic acids (Neumann *et al.*, 2004). In comparison to other classes of biologically relevant molecules, carbohydrates present a unique group of structural and energetic features that makes accurate modeling of their properties a daunting task. When simulating carbohydrates, some intra- or intermolecular interactions might require special treatments in the employed potential-energy functions to generate meaningful results. Over the past two decades, the increased awareness of the tremendous biological significance of carbohydrates motivated the development of computational tools specifically tuned for carbohydrate simulations. Below is a brief outline of some important carbohydrate-related modeling challenges (Figure I-3), as well as reported approaches for dealing with them.

Chemistry and stereochemistry. Carbohydrates are densely packed with polar functional groups relative to the small size of their monomeric building blocks. In aqueous solution, individual sugar units consist of small ring systems holding together a set of highly polar bonds separated by a bond or a single atom. As such, sugars are hardly distinguishable from clusters of water molecules (Kubik, 2012); hence proteins or other macromolecules would have no apparent reason to prefer binding sugar molecules over bulk solvent. However, the small hydrophobic patches on the two faces of sugar rings are sufficient to set them apart from water clusters, presumably by engaging in specific types of intermolecular contacts such as C–H···π interactions (Kubik, 2012). Moreover, certain phenomena, e.g. the anomeric and exoanomeric effects, are more frequently observed in carbohydrates because they normally possess the necessary structural motifs. The anomeric effect, for instance, is a stereoelectronic effect observed when an electronegative atom one bond away from the ring oxygen of the sugar shifts the conformational preference towards an otherwise sterically disfavored conformer (Tvaroska and Carver, 1998). Accounting for such effects requires the addition of specialized terms to the potential energy function (Kirschner *et al.*, 2008; Lii *et al.*, 2003). An accurate description of the electronic effects in carbohydrates is further complicated by their stereochemistry. Subtle variations of spatial charge distributions between stereoisomers, and even within different conformers of a single monosaccharide, might have substantial consequences on intra- and intermolecular interactions (Foley *et al.*, 2012). Fortunately, however, force fields specifically tuned to handle the peculiar geometrical and electrostatic features of carbohydrates are increasing in number and quality. Such carbohydrate force fields are being adopted more frequently in biomolecular simulations involving carbohydrate–macromolecule interactions (Foley, *et al.*, 2012).

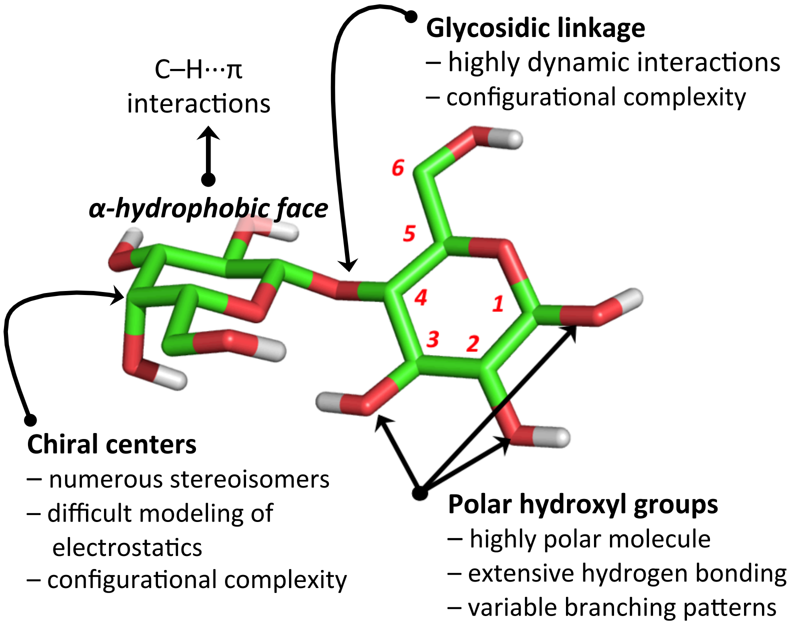


Figure I-3: Structure of a disaccharide (lactose) highlighting computationally challenging features of carbohydrate ligands and illustrating the common numbering scheme for pyranosyl sugar units.

Configurational complexity. Typical saccharide monomers have five sites of attachment for additional biomolecular units, two of which generate linear chains while the remaining three result in branched structures. Each glycosidic linkage between two sugar monomers has two torsional degrees of freedom, or three if one unit is linked via its 6'-hydroxyl group (Figure I-3). In some cases, however, the rotameric states of glycosidic bonds are not well defined (Bohne *et al.*, 1998). Combined with the fact that, in most monosaccharaides, all but one carbon center are chiral, the outcome is an enormous number of conformational and configurational possibilities (Foley, *et al.*, 2012). It is, thus, excessively difficult to simulate the complicated conformational space of carbohydrate molecules in reasonable time.

Electrostatic interactions. Most carbohydrates are very flexible and have a high number of hydroxyl groups, which enable them to make extensive hydrogen-bond networks and strong electrostatic interactions (Agostino *et al.*, 2009). The binding sites of carbohydrate-binding proteins, in turn, commonly contain charged residues and/or ions (Boraston *et al.*, 2004; Fadda and Woods, 2010; Quiocho, 1986). Moreover, numerous biologically relevant glycans contain charged moieties, e.g. N-acetylneuraminic acid and sulfated glycosaminoglycans. Consequently, charge-dependent interactions are commonly cited as major contributors to binding enthalpy in carbohydrate–protein systems (Fadda and Woods, 2010). Accurate handling of these interactions by docking and force field methods requires a proper treatment of charges for both ligand and protein atoms, which represents a significant bottleneck in development of force fields for biomolecular simulations (Fadda and Woods, 2010).

Hydrogen bonds. Hydrogen-bonds are crucial in defining carbohydrate-binding specificity. In general, lectins bind specifically to monosaccharide possessing the proper configuration of hydrogen-bonding partners (Rini, 1995), and loss or alteration of even a single group can lead to a significant drop in affinity (Drickamer, 1992; Lemieux, 1989; Sigurskjold and Bundle, 1992). It is difficult, however, to determine the actual reason for the importance of hydrogen bonds in carbohydrate–protein binding. Despite the obvious substantial enthalpic gain expected from hydrogen bonding between the ligand and protein, it is typically balanced, or sometimes even exceeded, by the corresponding cost of desolvating the polar H-bonding partners. In principle, however, the release of bound water molecules upon carbohydrate binding is associated with gain in entropy (Shimokhina *et al.*, 2006; Williams *et al.*, 1992), which could contribute for the overall free energy gain from H-bonding. Entropy has indeed been shown to be a major contributor to carbohydrate–protein binding (Lammerts van Bueren and Boraston, 2004).

C–H···π interactions. Contacts between C–H bonds on the hydrophobic face of carbohydrates and aromatic side chains of the protein residues are distinctive features of carbohydrate–protein interactions (Laughrey *et al.*, 2008). Based on their geometric parameters, C–H···π interactions are considered as weak hydrogen bonds (Brandl *et al.*, 2001), and experimental studies verified their importance in stabilizing lectin-sugar complexes (Muraki *et al.*, 2002). Lack of adequate treatment for C–H···π interactions in commonly used docking tools could negatively impact the quality of simulations of carbohydrate–protein systems (Agostino, *et al.*, 2009; Kerzmann *et al.*, 2006). Therefore, special treatment of C–H···π interactions has been incorporated into some force fields (Macias and Mackerell, 2005) and scoring functions (Kerzmann *et al.*, 2008; Kerzmann, *et al.*, 2006).

Dynamic and weak nature of interactions. Carbohydrates bind to their protein targets with relatively low affinity (Laederach and Reilly, 2005), most often in the millimolar to micromolar range (Ramkumar and Podder, 2000; Ramkumar *et al.*, 1995; Schwarz *et al.*, 1993). A significant fraction of carbohydrate-binding proteins (e.g. lectins) have shallow and un-structured binding sites in comparison to other binding pockets (Taroni *et al.*, 2000). As a consequence, carbohydrate–protein interactions are intrinsically more dynamic than many other ligand–protein systems (Fadda and Woods, 2010). Water molecules that bridge carbohydrate–protein interactions atoms show a high degree of dynamic exchange (Caffarena *et al.*, 2002; Tempel *et al.*, 2002). This inherently high mobility of carbohydrates makes approximating their binding to proteins by a single static configuration insufficient and calls for more computationally demanding methods, e.g. molecular-dynamics (MD) simulations. The timescale of MD simulations attainable on modern computers is typically sufficient for covering internal motions in glycans (Gonzalez-Outeiriño *et al.*, 2006; Kirschner and Woods, 2001). MD simulations can be employed to increase accuracy of docking results and provide ensemble averaging to aid in generating robust affinity and specificity predictions (Alonso *et al.*, 2006; Jorgensen, 2004; Taft *et al.*, 2008). MD simulations usually cost more time, though, and are typically reserved for cases where higher accuracy is necessary, e.g. the fine-tuning stage of lead optimization.

I.2.3 Thermodynamics of ligand–protein binding

Most, if not all, biological processes are tightly coupled to some binding event, where a macromolecular target, typically a protein, recognizes and selectively binds to its ligand. The non-covalent association of two (macro-) molecules is a reversible process governed by the laws of equilibrium thermodynamics. Like any other spontaneous process, this association happens if, and only if, it is accompanied by a negative change in Gibb's free energy of the system, which, in turn, has enthalpic and entropic components (equation I-1). Favorable (negative) enthalpic contributions in bimolecular associations are results of interactions between the binding partners, such as halogen and hydrogen bonds, electrostatic, van der Waals, and ionic interactions. Some events in the binding process, however, incur an enthalpic penalty, such as breaking of established hydrogen bond networks in the unbound partners or stripping molecules from their solvation shells. Entropic changes are related to alterations in overall dynamics of the system. Freezing of degrees of freedom (translational, rotational, and configurationally) of the binding partners or the solvent result in unfavorable (negative) entropic changes. On the other hand, entropic cost of binding could be partially compensated by releasing tightly-bound water molecules (De Lucca *et al.*, 1997), or increased protein mobility in response to ligand binding (MacRaild *et al.*, 2007). The currently prevalent thinking is that van der Waals interactions contribute the most to affinity, while hydrogen bonds and electrostatic interactions are more important to selectivity (Gilson and Zhou, 2007; Gohlke and Klebe, 2002; Smith *et al.*, 2012; Xu *et al.*, 1999). The smaller impact of the latter interactions on affinity stems from the fact that, when unbound, both ligand and protein can establish these interactions with water molecules and counterions (Gilson and Zhou, 2007; Gohlke and Klebe, 2002).

$$\Delta G = \Delta H - T\Delta S \qquad \textbf{I-1}$$

$\boldsymbol{\Delta G}$ *Gibb's free energy of ligand–protein association*
$\boldsymbol{\Delta H}$ *Enthalpic change upon ligand–protein association*
$\boldsymbol{\Delta S}$ *Entropic change upon ligand–protein association*
$\boldsymbol{T}$ *Absolute temperature*

Ligand–protein binding results from a delicate balance between favorable and unfavorable contributions from the different free-energy components. Finding an appropriate scheme to decompose the binding free energy into quantifiable components, as well as the precise calculation of these components, have given rise to significant obstacles in the development of fast and accurate structure-based design techniques (Smith, *et al.*, 2012). Despite the abundance of methodologies for predicting binding affinities from atomic configurations, the problem is so far considered unsolved. Table I-1 provides a brief overview the different types of methods for predicting binding free energy and highlights their primary strengths and weaknesses. These methods vary in the level of approximation used from rigorous statistical mechanics-based methods employing MD simulations in explicit solvent for sampling, e.g. FEP and TI, to the comparatively simple and computationally efficient ligand-based methods such as pharmacophore modeling and QSAR. The mixed-model approach lies in the middle between rigorous but time consuming methods (e.g. FEP) and fast but less accurate methods (e.g. classical QSAR), which do not account for important phenomena such as induced-fit and solvation. The mixed model approach combines automated flexible docking and multidimensional-QSAR by taking into consideration multiple ligand orientations (4D), induced-fit scenarios (5D) and solvation models (6D) (Vedani and Dobler, 2002; Vedani *et al.*, 2005; Vedani and Zbinden, 1998). Since docking and scoring are of highest relevance to the work presented in this thesis, they are discussed in more details in the following section. However, several review articles discussing the current status and future directions of the different approaches for binding affinity prediction can be found in the literature (Audie and Swanson, 2013; Bissantz *et al.*, 2010; DeMarco and Woods, 2008; Ferrara *et al.*, 2004; Gilson and Zhou, 2007; Gohlke and Klebe, 2002; Grosdidier *et al.*, 2009; Parenti and Rastelli, 2012).

Table I-1: Comparison of computational approaches for predicting the binding affinity of ligand–protein complexes.

Comparison	Free energy pathway	Direct ΔG calculation	Docking and scoring	Mixed-model approaches	Ligand-based approaches
Examples	FEP, TI	MM/GBSA, MM/PBSA, LIE	Glide, AutoDock	Quasar (mQSAR)	QSAR, CoMFA
Computational efficiency	Low	Moderate	High	Moderate–High	High
Protein structure	Required	Required	Required	Optional	Not required
Fitting to training set	Not required	Required	Required	Required	Required
Ligands to predict affinity	Close structural analogues only	Can be diverse	Can be diverse	Can be diverse	Can be diverse
Water model	Explicit	Implicit	Implicit	Explicit	Implicit
Accuracy mainly dependent on	Sampling, force field	Water model, force field, sampling	Compounds used for calibration	Compounds used for calibration	Compounds used for calibration

FEP: free-energy perturbation (Jorgensen and Thomas, 2008; Zwanzig, 1954), TI: thermodynamic integration (Rodinger *et al.*, 2005), LIE: linear interaction energy (Nicolotti *et al.*, 2012), MM: molecular mechanics, GBSA: generalized Born/surface area, PBSA: Poisson-Boltzmann/surface area, QSAR: quantitative structure-activity relationship (Verma *et al.*, 2010), mQSAR: multidimensional-QSAR(Lill *et al.*, 2004; Vedani and Dobler, 2002; Vedani, *et al.*, 2005; Vedani and Zbinden, 1998), CoMFA: comparative molecular field analysis (Cross and Cruciani, 2010).

Experimentally determined binding affinities of ligand–protein complexes are of central importance in the development and validation of computational methodologies. Equation I-2 delineates the relationship between the experimentally determined binding affinity and the association free energy of ligand–protein complexes. The binding process is an equilibrium reaction between the ligand–protein complex and the unbound partners:

$$[L] + [P] \rightleftharpoons [L{:}P]$$

Typically, experimental binding affinities are reported as K_a, K_d, K_i, IC_{50}, or EC_{50} (equation I-3). The equilibrium association constant (K_a, sometimes written as K_b for binding constant) and dissociation constant (K_d) are determined *directly* in a binding assay, e.g. by isothermal titration calorimetry (ITC). The equilibrium inhibition constant (K_i) is usually measured in an inhibition assay, where the compound of interest displaces a radiolabelled reference compound. The three constants (K_a, K_d, and K_i) are true thermodynamic equilibrium constants. On the other hand, IC_{50} and EC_{50} are measures of potency classically defined as 'the concentration of the ligand that produces 50% of the maximal response or reduces that response to 50% of its maximal value'. The interpretation of IC_{50} depends on the experimental setup and the employed inhibition model, but in most cases it has the same meaning as K_i.

$$\Delta G = -RTlnK_a \tag{I-2}$$

$$K_a = K_d^{-1} = K_i^{-1} = \frac{[L\colon P]}{[P][L]} \tag{I-3}$$

$\boldsymbol{\Delta G}$	*Gibb's free energy of ligand–protein association*
$\boldsymbol{T}$	*Absolute temperature*
$\boldsymbol{R}$	*Universal gas constant*
$\boldsymbol{K_a}$	*Association constant*
$\boldsymbol{K_d}$	*Dissociation constant*
$\boldsymbol{K_i}$	*Inhibition constant*
$[\boldsymbol{L}\colon \boldsymbol{P}]$	*Equilibrium concentration of the ligand–protein complex*
$[\boldsymbol{P}]$	*Equilibrium concentration of the unbound protein*
$[\boldsymbol{L}]$	*Equilibrium concentration of the unbound ligand*

I.2.4 Scoring functions and prediction of binding affinity

Molecular docking has become an integral component in structure-based drug design projects. Accurate and fast prediction of the binding modes of hypothetical molecules to macromolecular targets helps modelers and medicinal chemists understand the structural determinants of strong binding and hence guide and accelerate the hit finding and hit-to-lead optimization processes. The primary role of docking programs is to identify native binding modes of the ligand, typically a small molecule, within the binding site of the macromolecular target, typically a protein. Docking workflows consist primarily of two consecutive stages, docking and scoring. In the first stage, the *docking algorithm* generates a large number of *poses*, which should adequately cover potential solutions to the problem. Poses are defined by the conformation of the modeled ligand and its orientation in the binding site. Subsequently, the generated poses pass through one or more stages of scoring where unrealistic and unfavorable poses are excluded, and the remaining 'reasonable' poses are rank ordered. The prioritization step is carried out by the *scoring function*, which should ideally rank the poses more likely to occur in nature higher than it ranks decoy poses.

Although a fairly large number of scoring functions are available, none so far provides optimal performance in terms of both prediction accuracy and general applicability (Huang *et al.*, 2010). Development of rigorous and accurate scoring functions is still, thus, an active area of research with a sizeable room for improvement. Comprehensive coverage of available docking and scoring approaches is, however, beyond the scope of this introduction. A number of excellent reviews covering scoring and free-energy functions, their application in structure-based design, as well as critical assessment of their performance can be found in the literature (Cheng *et al.*, 2009; Ferrara, *et al.*, 2004; Gohlke and Klebe, 2001; Gohlke and Klebe, 2002; Grosdidier and Fernández-Recio, 2009; Guimaraes, 2011; Hou *et al.*, 2011b;

Huang and Zou, 2010a; Huang, *et al.*, 2010; Jain, 2006; Kitchen *et al.*, 2004; Mooij and Verdonk, 2005; Núñez *et al.*, 2010; Perola *et al.*, 2004; Rajamani and Good, 2007; Rastelli *et al.*, 2010; Schulz-Gasch and Stahl, 2004; Warren *et al.*, 2006). In this section we will shed some light on the three fundamental aspects of scoring functions: types, applications, and criteria used for performance assessment. In the following section carbohydrate-specific scoring approaches will be discussed in more detail.

I.2.4.1 Types of scoring functions

According to the theoretical foundations and mathematical form of the energy function, scoring functions are roughly classified into three categories: force field scoring functions, empirical scoring functions, and knowledge-based scoring functions.

Force field scoring functions. In molecular mechanics, the potential energy of molecular systems is decomposed into bonded (stretching, bending and torsional) and non-bonded (van der Waals and Columbic) terms. The use of the non-bonded energy components from force fields in docking and scoring dates back to the first docking program DOCK (Kuntz *et al.*, 1982) whose energy parameters were taken from the AMBER force field (Weiner *et al.*, 1984). The scoring function in DOCK comprises two energy terms; Lennard-Jones van der Waals term and an electrostatic term:

$$E = \sum_{i \neq j}^{non-bonded} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \left(\frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right)$$

where A_{ij} and B_{ij} are the vdW parameters, r_{ij} is the distance between the non-bonded atoms i and j, q_i and q_j are the atomic charges. The Coulombic term employs a distance-dependent dielectric constant $\varepsilon(r_{ij})$ to account for the shielding effect of the solvent. Force field parameters are typically optimized to reproduce experimental or theoretical *ab initio* data. The use of simple physical model for non-bonded interactions makes interpretation of force field-derived scores straightforward. Moreover, the high computational efficiency of force field scoring functions makes them suitable for fast pose ranking as well as database screening. Examples of docking programs using force field terms in their scoring functions include YETI (Vedani, 1988), DOCK 4.0 (Ewing *et al.*, 2001), GOLD (Jones *et al.*, 1997), SYBYL/D-Score (Meng *et al.*, 1992), SYBYL/G-Score (Jones, *et al.*, 1997), and AutoDock (Huey *et al.*, 2007; Rosenfeld *et al.*, 2003).

The major limitation of the force field model is the lack of proper treatment for solvent effects. In addition to the inaccuracy of modeling shielding effect via simple a distance-dependent dielectric constant, the approach employed in DOCK (discussed above) does not take desolvation effects into account. A direct undesirable consequence of ignoring desolvation effects is overestimation of electrostatic interactions and biasing the scoring function towards highly charged molecules (Huang, *et al.*, 2010). Accurate treatment of solvation effects is indeed a persistent challenge in modeling free-energy changes in molecular systems. Rigorous methods such as free energy perturbation (FEP) (Jorgensen and Thomas, 2008; Zwanzig, 1954) or thermodynamic integration (TI) (Rodinger, *et al.*, 2005) simulate the solvent molecules explicitly, but they are usually associated with a substantial increase in computation time. A reasonable compromise, however, is offered by implicit solvent models which treat the solvent as a continuum electrostatic, e.g. Poisson–Boltzmann/surface area (PB/SA) model (Baker *et al.*, 2001; Grant *et al.*, 2001; Rocchia *et al.*, 2002; Tjong and Zhou, 2008) and the generalized-Born/surface area (GB/SA) model (Hawkins *et al.*, 1995; Qiu *et al.*, 1997; Still *et al.*, 1990). Several studies have successfully employed the PB/SA (Huang and Caflisch, 2004; Kuhn *et al.*, 2005; Kuhn and Kollman, 2000; Pearlman, 2005; Sims *et al.*, 2003; Thompson *et al.*, 2008; Wang *et al.*, 2001a) and GB/SA (Cecchini *et al.*, 2004; Cho *et al.*, 2005; Guimaraes, 2011; Guimarães, 2012; Guimarães and Cardozo, 2008; Liu *et al.*, 2004; Liu *et al.*, 2009; Liu and Zou, 2006; Lyne *et al.*, 2006; Zou *et al.*, 1999) approaches for relative potency predictions and virtual screening.

Despite the well-documented successes, the implicit solvent approaches suffer from a couple of limitations. For example, Zou and colleagues pointed out that in spite of the accurate reproduction of solvation energies of both ligand and protein; the implicit solvent models are not necessarily suitable for binding affinity calculation (Liu, *et al.*, 2009; Liu and Zou, 2006). The authors attributed the inaccuracy in affinity prediction to improper treatment of phenomena such as partial desolvation of ligand and/or protein and proposed an improved GB multiscale approach optimized for virtual screening. In this approach a subset of atoms, which are more critical to binding electrostatics, are calculated using accurate GB model at the cost of increased computation time (Liu, *et al.*, 2009). Another limitation of the GB/SA approach stems from the inaccurate modeling of protein electrostatics and protein–solvent interactions by the approximations used in the GB model (Guimarães and Mathiowetz, 2010). Moreover, MM/GBSA scores tend to have a wide dynamic range probably due the application of a protein dielectric constant of 1 in a model where protein motions and polarization are not taken into account (Guimarães and Mathiowetz, 2010). These limitations were addressed in some improved GB/SA models such as the VSGB 2.0. The VSGB 2.0 model employs the Surface Generalized Born (SGB) model (Ghosh *et al.*, 1998; Yu *et al.*, 2006) in conjunction with a variable dielectric (VD) treatment to account for polarization effects from protein side chains by varying the internal dielectric constants from 1.0 to 4.0 (Zhu *et al.*, 2007).

Empirical scoring functions. The molecular mechanics model employed by in force field scoring functions lacks specific terms accounting for important energy components in molecular associations such as entropic changes and ligand/protein strain penalties. The second type of scoring functions, empirical scoring functions, employs a more adaptable functional form to estimate binding free energy using a weighted sum of energy terms:

$$\Delta G = \sum_i w_i \Delta G_i$$

Empirical scoring functions employ a suitable decomposition scheme to break down the free energy of binding, ΔG, into a set of additive energy components, ΔG_i, such as vdW energy, electrostatics, solvation, entropic penalty, ligand strain penalty, etc. The empirical weighting coefficients are typically derived by fitting experimental binding affinity data for a set of ligand–protein complexes with known three-dimensional structures. As a result of employing simple energy terms, empirical scoring functions are generally faster than force field scoring functions (Huang, *et al.*, 2010). Examples of empirical scoring functions include FlexX (Rarey *et al.*, 1996), Glide (Friesner *et al.*, 2006; Friesner *et al.*, 2004; Halgren *et al.*, 2004), ICM (Abagyan *et al.*, 1994), LUDI (Böhm, 1994; Böhm, 1998), PLP (Gehlhaar *et al.*, 1995), ChemScore (Eldridge *et al.*, 1997), Surflex (Jain, 2003), MedusaScore (Yin *et al.*, 2008), AIScore (Raub *et al.*, 2008), and SFCscore (Sotriffer *et al.*, 2008).

Development of generally applicable empirical scoring functions is faced by three major challenges (also common to force field scoring functions):

- *Accurate calibration of weighting coefficients.* Weighting coefficients are necessary because different energy components might come from unrelated methods and, hence, have very different scales (Huey, *et al.*, 2007; Liu, *et al.*, 2004; Morris *et al.*, 1998; Zou, *et al.*, 1999). Although it is relatively easy to obtain appropriate empirical coefficients for a specific protein or protein family, it is rather difficult to obtain a universal training set of diverse ligand–protein complexes (Huang, *et al.*, 2010).
- *Non-additivity of energy components.* Empirical free-energy models are based on the additivity assumption, i.e. the notion that the total free energy can be expressed as a sum of independent free-energy components. This concept, albeit it tremendous utility, does not always hold to careful scrutiny. For a set of free energy components to be truly additive they need to be fully independent of each other, which is not always the case in energy components commonly used in empirical free-energy functions (Dill, 1997). Several studies have shown that non-covalent interactions are often mutually reinforcing (positively cooperative) or mutually weakening (negatively cooperative) rather than additive (Baum *et al.*, 2010; Dill, 1997; Williams *et al.*, 1993; Williams *et al.*, 2004).
- *Holes in the free-energy landscape.* Hill and Reilly used the term 'holes' to describe the apparent 'lack of [experimental] energetics data for atomic configurations that deviate from optimality' (Hill and Reilly, 2008). The absence of non-optimal geometries in the training data used to fit empirical functions could reduce the sensitivity of the resultant models towards penalizing bad structural motifs. Unfortunately, this problem cannot be solved by increasing the number or quality of experimentally determined structures, as their atomic configurations are found only in optimal states (Hill and Reilly, 2008).

Knowledge-based scoring functions. Knowledge-based scoring functions (also known as statistical potential-based scoring functions) model intermolecular interactions as a sum of atom-pair potentials derived from the occurrence frequencies of atom pairs in databases of experimentally determined structures (Miyazawa and Jernigan, 1985; Sippl, 1990; Tanaka and Scheraga, 1976). The pairwise potentials are calculated using the inverse Boltzmann relation (Koppensteiner and Sippl, 1998; Thomas and Dill, 1996a):

$$w(r_{ij}) = -k_B T ln \left[\frac{\rho(r_{ij})}{\rho^*(r_{ij})} \right]$$

where $w(r_{ij})$ is the pairwise potential associated with finding atoms i and j at some distance r, k_B is the Boltzmann constant and T is the absolute temperature of the system, and $\rho(r)$ and $\rho^*(r)$ are the occurrence density functions for the ligand–protein atom pair in crystal structures of the training set and in a reference state where the interatomic interactions are zero. Examples of knowledge-based scoring functions include ITScore (Huang and Zou, 2006a; Huang and Zou, 2006b; Huang and Zou, 2010b), PMF (Muegge, 2006; Muegge and Martin, 1999), DrugScore (Gohlke *et al.*, 2000; Velec *et al.*, 2005), DFIRE (Zhang *et al.*, 2005), BLEEP (Mitchell *et al.*, 1999a; Mitchell *et al.*, 1999b), MScore (Yang *et al.*, 2006), and KScore (Zhao *et al.*, 2008).

Since knowledge-based scoring functions are derived from large and diverse data sets of structural data rather than by reproducing a limited set of binding affinities, they are typically more robust and less sensitive to the training set (Huang and Zou, 2006a; Huang and Zou, 2006b; Muegge, 2006; Muegge and Martin, 1999). Due to their pairwise characteristic, the scoring process could be as fast as empirical scoring functions. The primary challenge in deriving knowledge-based scoring functions, however, is the proper definition of the reference state. Knowledge-based scoring functions commonly approximate the reference state with an atom-randomized state. Such approximations, however, ignore some important structural features of ligand–protein systems such as excluded volume and interatomic connectivity (Thomas and Dill, 1996b). Moreover, the fact that knowledge-based scoring functions are trained using structural data only and no binding affinity data could compromise their prediction accuracies. Interestingly, however, a knowledge-based quantitative structure-activity relationship approach has been introduced to fill in this gap by using the atom-pair occurrences as descriptors and fitting them to binding affinities in the training set (Ballester and Mitchell, 2010; Deng *et al.*, 2004). In general, knowledge-based scoring functions do not explicitly account for solvation or entropy effects; a deficiency yet to be fully addressed in new knowledge-based approaches (Huang and Zou, 2010b).

Consensus scoring. In consensus scoring, a combination of the scores from several scoring functions is employed to increase the chances of finding the correct answer to docking and binding affinity problems (Charifson *et al.*, 1999). Several consensus scoring strategies can be used to combine the different scores, e.g. vote-by-number, rank-by-number, linear combination, etc.; and the choice of the appropriate strategy is essential to obtain accurate predictions and for computational tractability (Oda *et al.*, 2006).

I.2.4.2 Applications and performance assessment of scoring functions

Scoring functions are employed to achieve three related goals: identifying native binding mode, predicting binding affinity and finding active hits in database screening. At the most basic level, a reliable scoring function should be able to distinguish native docking modes from decoys by assigning better scores to the former. Success of docking/scoring applications is classically defined by the RMSD values between top-ranked ligand pose(s) and the experimentally determined configuration. Typically, if one of the highly ranked poses (ideally the top-ranked pose) exhibits RMSD value ≤ 2.0 Å, the docking prediction is deemed accurate. However, the use of RMSD for judging prediction accuracy might give misleading results in small and symmetrical ligands as well as in large ligands where an irrelevant (e.g. solvent-exposed) part of the molecule is incorrectly predicted (Huang, *et al.*, 2010). Thus, alternative metrics have been proposed to overcome this limitation, e.g. relative displacement error (Abagyan and Totrov, 1997), interaction-based accuracy classification (Kroemer *et al.*, 2004), and Generally Applicable Replacement for rmsD (Baber *et al.*, 2009).

Although a multitude of docking/scoring programs have achieved considerable success in reproducing crystal poses, accurate prediction of binding affinity from these poses is still largely elusive (Huang and Zou, 2010b; Warren, *et al.*, 2006). Scores calculated by scoring and free-energy functions are generally of different scales than experimental binding affinity and a scaling factor is typically required to reproduce the absolute experimental values. Therefore, instead of comparing scores to the exact values of experimental data, it is common to employ a correlation metric to evaluate the prediction accuracy of scoring functions. For example, the Pearson product-moment correlation coefficient *r*, is used to assess the linear correlation between calculated (y_i) and experimental (x_i) values:

$$r = \frac{\sum_{i=1}^{N}(x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}\sqrt{\sum_{i=1}^{N}(y_i - \langle y \rangle)^2}}$$

The Spearman's rank correlation coefficient r_s is typically used in cases where accurate rank ordering of ligands is more important, e.g. in database screening applications:

$$r_s = 1 - \frac{6\sum_{i=1}^{N} D_i^2}{N(N^2 - 1)}, \quad D_i = x_{i,rank} - y_{i,rank}$$

The third application of scoring function is to mine compound chemical databases for potentially active hits against a specific protein target. Lead identification in drug discovery studies is probably the most important application of docking/scoring functions as virtual screening tools (Wildman, 2012). Conventionally, an enrichment factor is employed in the assessment of scoring functions performance in this virtual screening. Enrichment factors estimate the accumulated rate of correctly identified actives as opposed to the number of compounds hypothetically found if compounds were screened randomly (Bender and Glen, 2005). A higher enrichment factor for certain scoring methods is an indication that this method can correctly rank the molecules in the screened database distinguishing active from decoy molecules. The area under the receiver operating characteristics (ROC) curve is another commonly employed measure for virtual database screening efficiency (Jain, 2000; Warren, *et al.*, 2006).

Examples of critical assessments and comparison of the performance of commonly used scoring functions, along with some relevant review articles, were given in the beginning of this section. A detailed discussion of the results of these studies is certainly beyond the scope of this introduction. It is necessary, however, to emphasize a couple of common attributes of scoring functions that are relevant to this study:

- **Lack of universal validation set.** Collections of ligand–protein complexes used to construct training and/or test sets for scoring functions are typically extracted from databases such as and Binding MOAD (http://www.bindingmoad.org/) (Benson *et al.*, 2008; Hu *et al.*, 2005), BindingDB (http://www.bindingdb.org/) (Chen *et al.*, 2001; Liu *et al.*, 2007), AffinDB ((http://www.agklebe.de/affinity) (Block *et al.*, 2006) and PDBbind (Wang *et al.*, 2004; Wang *et al.*, 2005). However, every scoring function and every performance comparison done on several scoring functions employ different collections of complexes. In absence of a standardized and unified set of ligand–protein complexes across docking scoring studies, it is rather difficult to use the results of one study in another or to draw general conclusions about certain scoring function or families of scoring functions.
- **Target-dependent performance.** Another commonly observed feature of scoring and free-energy functions is the fluctuation of their prediction accuracies among different protein families (Ferrara, *et al.*, 2004; Huang, *et al.*, 2010; Mooij and Verdonk, 2005; Perola, *et al.*, 2004; Warren, *et al.*, 2006). A scoring function that accurately predicts affinities of ligands to a certain target does not necessarily perform equally well on other targets. Marsden *et al.*, for instance, demonstrated that scoring functions generally perform better on the training sets used in their own publications than on other, apparently similar, sets (Marsden *et al.*, 2004). They also pointed out that the correlation between scores and binding affinities are better within family-specific subsets in comparison to the whole data set.
- **Application-dependent performance.** Most of the existing scoring functions perform well in only one or two of the three applications discussed above and fail in others (Huang and Zou, 2010a; Huang, *et al.*, 2010). This is somewhat counter-intuitive, since all three applications (pose selection, binding affinity, and database ranking) are governed by the same physical laws. Warren *et al.* performed an extensive assessment of 37 different scoring functions for their performance in each of the three applications (Warren, *et al.*, 2006). The authors reported that although docking algorithms could essentially perform virtual crystallography (i.e. generate experimental small molecule conformations), scoring functions could not reliably identify the best-docked pose in all cases. Moreover, they observed a surprising

discrepancy between the ability of some scoring functions to rank poses and/or predict binding affinity in some target and their ability to correctly identify actives in database screening. This result led them to a rather radical statement: ‘under certain circumstances scoring functions are not ranking compounds based on structural information’. Nevertheless, the study was concluded with a recommendation that the quality and reliability of docking/scoring programs could be improved by the intervention of a ‘skilled computational chemist’.

I.2.5 Carbohydrate-specific scoring functions

To the best of our knowledge, three methods are reported that deal specifically with quantification of carbohydrate–protein interactions. These studies have two main characteristics in common. First, all of them report the development and assessment of two functions; a computationally efficient *scoring function* for ranking of putative binding modes produced by docking software and a *free-energy function* for prediction of binding affinity from docking poses. The latter function typically features an improved treatment of solvation and entropy that makes it less suited for fast prioritization of docking poses, but more suitable for accurate binding free-energy prediction. The second common trait in these studies is that the free-energy functions therein comprised a linear combination of terms adopted from another previously reported scoring function or force field, with linear weighting coefficients derived by fitting to a training set of carbohydrate–protein complexes with known structure and affinity. The proposed free-energy functions differ, however, in the way they decompose the binding free energy and in the included special treatment for some of the free-energy components.

In the first study, Laederach and Reilly (2003) employed a set of 30 carbohydrate–protein complexes to optimize a docking protocol for prediction of structure and affinity of carbohydrate–protein complexes. The authors employed an empirical formulation based on AutoDock scoring function (Morris, *et al.*, 1998):

$$\Delta G = f_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + f_{hbond}\left[\sum_{i,j} E(t)\left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}}\right) + E_{hbond}\right] +$$
$$f_{elec}\sum_{i,j}\left(\frac{q_i q_j}{\varepsilon(r_{ij})r_{ij}}\right) + \Delta G_{tor}N_{tor} + f_{solv}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{\left(-r_{ij}/2\sigma^2\right)}$$

The first three terms (Lennard-Jones 6-12 potential, hydrogen bonding, and electrostatic interactions) represent intermolecular interaction energy. The Lennard-Jones parameters are based on the AMBER force field (Pearlman *et al.*, 1995) for protein and metals and on GLYCAM_93 force field (Woods *et al.*, 1995) for the carbohydrate. *E(t)* is a directional attenuation factor used to enforce optimum hydrogen bonds geometries (Goodford, 1985). The Coulomb electrostatic term is evaluated using a distance-dependent dielectric constant

$\varepsilon(r_{ij})$ (Mehler and Solmajer, 1991). The last term accounts for the solvation contribution from the change in solvent accessible surface area of non-polar ligand atoms as defined in (Morris, *et al.*, 1998). This term does not model polar atoms, however, as they should be covered in the explicit hydrogen bonding term.

The E_{hbond} energy was added to the second (hydrogen-bond) term to compensate for the penalty associated with disruption of ligand–water hydrogen bond network when the ligand leaves the solvent and binds to the protein. The authors assessed six different approaches to model this energy using three base values for the average H-bond energy *(E_{hb}=5.0, 2.5, and 1.0 kcal/mol)* and two ways of counting ligand groups that should receive this penalty:

$$E_{hbond} = (n_A + n_B)E_{hb} \qquad \text{or} \qquad E_{hbond} = (n_A + n_B - n_{OH})E_{hb}$$

where n_A, n_D, and n_{OH} are the numbers of H-bond acceptors, donors and hydroxyl groups in the ligand. The latter formula assumes that a hydroxyl group cannot simultaneously function as hydrogen bond acceptor and donor in the solvent. Additionally, the authors examined three methods to count the number of torsions *(N_{tor})* used to estimate entropy: using the number of heavy atom torsions, the total number of torsions in the molecule, and the number of glycosidic bond torsions. In total, 18 (6 × 3) different models were tested for the ability to reproduce affinities of the training set by fitting the equations to the binding affinities in the training set using linear regression.

The best performing model, with E_{hb} = 1.0 kcal/mol and counting heavy-atom torsions only, exhibited a residual standard error of 1.4 kcal/mol in the training set, clearly out-performing the standard AutoDock function, which showed an standard error of 2.2 kcal/mol (Morris, *et al.*, 1998). The authors applied the suggested free-energy function to predict affinities of a test set of 17 *Aspergillus niger* glucoamylase inhibitors for which binding energies had been determined experimentally. The model predicted the free energies of the test set with 1.1 kcal/mol residual error, which further confirmed its validity. The authors noticed that the coefficient for the electrostatic term is significantly larger compared to AutoDock scoring function, which they suggested as an indication of its higher relative importance in determining carbohydrate binding affinities.

Later, Hill and Reilly expanded this study to a much larger data set (Hill and Reilly, 2008). They followed the same approach starting from AutoDock scoring function and examined more alternative methodologies to calculate free-energy components. In their extended analysis, the Lennard-Jones and hydrogen-bonding parameters from AMBER99 (Pearlman, *et al.*, 1995), CHARMM22 (Brooks *et al.*, 1983), MM3PRO (Ponder and Case, 2003), and AutoDock (Morris, *et al.*, 1998) force fields were compared. For the solvation term, two sets of parameters were also tested using either heavy atoms or carbons only. Four rational values for E_{hb} were used: 0.0, 1.0, 2.5, and 5.0 kcal/mol. Finally, they introduced a novel entropic term that accounts for ligand's translational and rotational degrees-of-freedom via an empirical coefficient, ξ, to couple them with torsional DoF according to the formula

$$\Delta S_{bind} = -k\left[ln(6 + \xi N_{tors}) - \frac{\xi N_{tors}}{6 + \xi N_{tors}} ln\xi\right]$$

where N_{tors} are the number of freely rotatable bonds in the ligand and k is the Botlzmann constant. Four values for this coupling coefficient were examined: ξ=0.1, 0.33, 0.67, and 1.0, in addition to the three methods of counting rotatable bonds described in the previous study (Laederach and Reilly, 2003). In total, 288 different models were evaluated.

The final training set employed for computing the linear regression comprised 115 unique carbohydrate–protein pairs for which AutoDock found a docking solution within 1.0 Å of the crystal structure pose. The best model (JA model) achieved a root-mean-squared-error of 2.0 kcal/mol. The JA model utilizes the AutoDock force field and uses heavy-atoms to compute the solvation parameter and to count the torsional angles. Surprisingly, the entropy term in JA model did not use the novel treatment proposed by the authors in this study. The authors also noted that the models they assessed were most sensitive to the choice of force field, AutoDock being the best, and less sensitive to the choice of entropy or solvation treatments. The authors used jackknife analysis (leave-one-out cross-validation) to confirm the robustness of the suggested model, however, they did not report the use of a test set for validation.

The third approach in this regard was the SLICK (Sugar–Lectin Interactions and DoCKing) scoring functions introduced by Kerzmann *et al.* (2006). SLICK has two variants, SLICK/score for rescoring structures generated by docking programs and the more extended SLICK/energy for estimating binding free energy. SLICK/energy estimates ΔG as a sum of five weighted components:

$$\Delta G = c_0 + c_{CH\pi}\Delta G_{CH\pi} + c_{hb}\Delta G_{hb} + c_{vdW}\Delta G_{vdW} + c_{np}\Delta G_{solv}^{np} + c_{es}(\Delta G_{solv}^{es} + \Delta G_{int}^{es})$$

The $\Delta G_{CH\pi}$ term accounts for C–H···π interactions, which are important mediators of carbohydrate–protein interactions (Fernández-Alonso *et al.*, 2005). The authors employ the model described by Brandl *et al.* (2001) to compute C–H···π interactions, which uses three simple geometric parameters analogous to those used for H-bonds: d_{CX}, the distance from the carbon atom to the center of the ring X, α_{CHX}, the angle between the C—H bond and the line connecting the hydrogen to the center of ring X, and d_{H_pX}, the distance of the H-atom to the ring center projected into the ring plane (Kerzmann, *et al.*, 2006). The authors combine the parameters suggested by Brandl *et al.* for locating C–H···π interactions with a sigmoidal switching function to reduce sensitivity to small experimental errors in structure.

Hydrogen-bonds *(*ΔG_{hb}*)* were treated by a modified version of the model described by Böhm (Böhm, 1994) with the parameterization of Eldridge and co-workers (Eldridge, *et al.*, 1997). The van der Waals contribution was computed using a modified version of the AMBER force field (Cornell *et al.*, 1995). The authors used a softened Lennard-Jones potential in the van der Waals calculation to improve tolerance for minor structural reorganization (Ferrari *et al.*, 2004). Lennard-Jones 6-12 parameters were taken from the Glycam2000a force field

(Basma *et al.*, 2001). Electrostatic interactions were computed using Coulomb's law. Finally, the solvation term was computed using the solvation model of Jackson and Sternberg (Jackson and Sternberg, 1995).

The authors employed a data set of 20 lectin–sugar complexes with affinities measured by Isothermal Titration Calorimetry (ITC) to calibrate and validate the empirical free-energy function. The set was divided equally into a training set (10 complexes) used to derive the linear weighting coefficients and a test set (10 complexes) used to assess the quality of binding affinity prediction. SLICK/energy predicted binding affinities in with a maximum absolute error of 2.8 kJ/mol (0.7 kcal/mol), and a mean absolute error of 3.3 kJ/mol (0.8 kcal/mol) in a randomized five-fold cross-validation run. According to the authors, the hydrogen-bonding and C–H···π terms are most important for the identification of binding conformations. The authors noted, however, that prediction quality was size-dependent, with the smaller ligands predicted more accurately. They attributed this to the conformational complexity associated with numerous glycosidic linkages in oligomers compared to monomers, and to the need for a more robust, possibly explicit, treatment of water molecules in the binding site.

Notably, the authors warned against the danger of over-fitting due to the relatively small size of the calibration set employed in the study. In a later study, the authors employed a larger data set (18 training and 22 test complexes) to improve the performance of SLICK in both docking and binding affinity predictions (Kerzmann, *et al.*, 2008). The new approach has been implemented in the freely available docking program BALLDock (Kohlbacher and Lenhof, 2000). The results of the docking study confirmed the superior quality the new approach in comparison to another docking program, FlexX (Rarey, *et al.*, 1996). Out of 18 training complexes, 17 could be successfully redocked with an average rmsd of 0.85 Å and an average absolute error of 3.6 kJ/mol in the binding free-energy estimate. However, the study did not comment on the quality of free-energy predictions in the test set, probably because only five complexes in the test set had known experimental binding affinities.

I.2.6 Aim and strategy

The foundation of structure-based design is the notion that all properties of a given ligand–protein complex, such as the binding free energy, could be inferred from the atomic configuration. However, the development of computational methods to accurately account for the enthalpic and entropic components of the binding process remains a central challenge in molecular modeling. Although numerous successful examples are reported for quantifying one or more of the ΔG components, these approaches are not always transferrable. Thus, the relationship between the structure of the ligand–protein complex and free-energy components has essentially remained elusive so far.

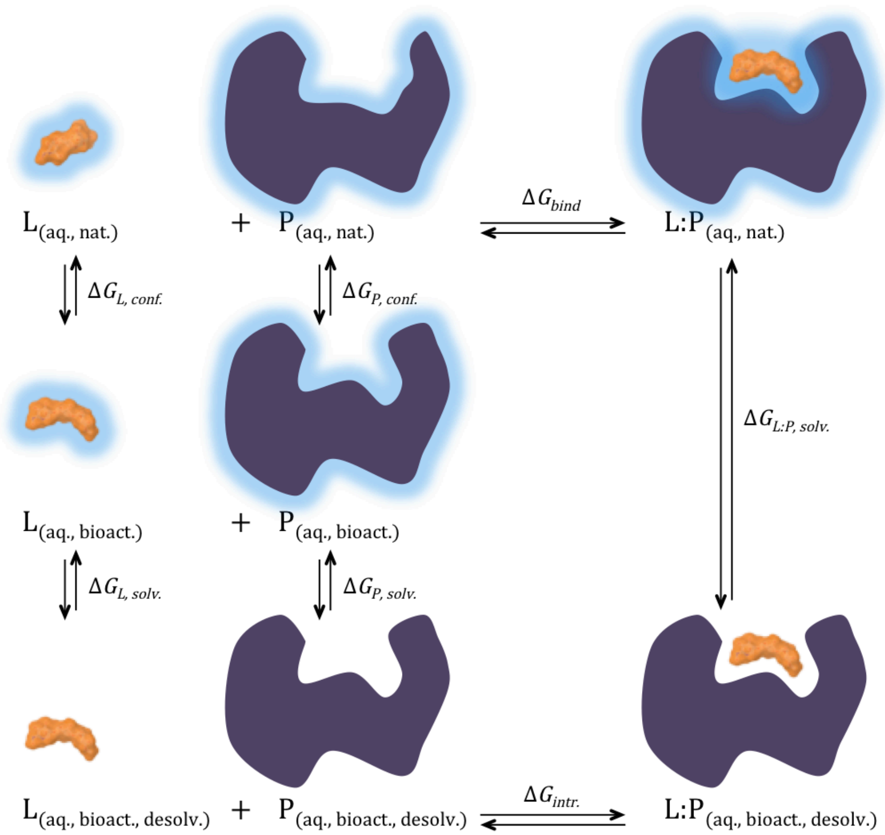


Figure I-4: Thermodynamic cycle of ligand–protein binding in water. Abbreviations: L=Ligand, P=Protein, L:P=Ligand–Protein complex, aq.=aqueous, nat.=native, conf.=conformational, bioact.=bioactive, solv.=solvation, desolv.=desolvated, intr.=intrinsic.

The ultimate goal of this study is to formulate and validate an empirical scoring function to quantify binding of small carbohydrate ligands to their macromolecular targets. Each term in this scoring function should represent a component of the free energy of binding. *(Note: throughout this thesis, the terms equation, model, and scoring function will be used interchangeably)*. The free energy of ligand–protein binding was decomposed into calculable components according to the thermodynamic cycle shown in Figure I-4. Since Gibb's free energy (G) is a state function, it is dependent only on the specific state of the system not the path taken to reach this particular state (equation I-4). Therefore, the observable free energy of binding (ΔG_{bind}) is path-independent, allowing it to be represented as a sum of free-energy changes in any closed thermodynamic cycle such as the one presented in Figure I-4.

$$\Delta G_{bind} = G(L{:}P)_{aq.,native} - \left[G(L)_{aq.,native} + G(P)_{aq.,native}\right] \qquad \textbf{I-4}$$

$\boldsymbol{\Delta G_{bind}}$ *Gibb's free energy change upon ligand–protein association*

$\boldsymbol{G(L{:}P), G(L), G(P)}$

Standard Gibb's free energies of the ligand–protein complex, free ligand, and free protein in their native aqueous states at equilibrium, all of which are path-independent state functions

Since the sum of ΔG changes in any closed thermodynamic cycle equals zero, we can calculate binding free energy as a sum of ΔG components according to equation I-5 by rearranging the terms from Figure I-4. Same arguments apply to the enthalpy and entropy components of Gibb's free energy, since both of them are state functions as well. Consequently, equation I-5 remains valid if we replace ΔG*'s* by ΔH*'s* or ΔS*'s*.

$$\Delta G_{bind} = \Delta G_{intr.} + \Delta G_{L,conf.} + \Delta G_{P,conf.} + \Delta G_{L,solv.} + \Delta G_{P,solv.} - \Delta G_{L:P,solv.} \tag{I-5}$$

$\boldsymbol{\Delta G_{bind}}$	*Gibb's free energy change upon ligand–protein association*
$\boldsymbol{\Delta G_{intr.}}$	*Intrinsic binding free energy of ligand–protein complex in vacuum (direct intermolecular interactions)*
$\boldsymbol{\Delta G_{L,conf.}}$	*Conformational free energy of ligand induced-fit*
$\boldsymbol{\Delta G_{P,conf.}}$	*Conformational free energy of protein induced-fit*
$\boldsymbol{\Delta G_{L,solv.}}$	*Desolvation free energy of the ligand*
$\boldsymbol{\Delta G_{P,solv.}}$	*Desolvation free energy of the protein*
$\boldsymbol{\Delta G_{L:P,solv.}}$	*Solvation free energy of the ligand–protein complex*

Equation I-5 separates solvation effects (the last three terms) from the potential energy of interaction ($\Delta G_{intr.}$), which represents direct ligand–protein intermolecular interactions *in vacuo*. These interactions are mediated via familiar contacts, e.g. van der Waals and electrostatic interactions, H-bonds, π-π stacking, etc., which are described in traditional force fields and scoring functions. Moreover, empirical free-energy functions commonly employ yet another separation of terms; where enthalpy and entropy contributions are split up. This approximation is done primarily for pragmatic reasons, since formal treatment of entropy is very computationally demanding. Entropy is typically accounted for by a single empirical term, or sometimes even neglected. In this study, we adopted a similar scheme to construct and assess empirical functions for binding free-energy prediction. The following generic formula (the Master Equation) was employed as the basis of our investigations.

Master Equation

$$\Delta G_{bind} = c_1 \Delta G_{inter} + c_2 \Delta G_{solv} + c_3 \Delta G_{strain} + c_4 \Delta S_{lig} + c_5 \Delta G_{reward/penalty}$$

, where ΔG_{inter} is the ligand–protein interaction energy, ΔG_{solv} is the desolvation penalty associated with binding, ΔG_{strain} is the conformational strain penalty, ΔS_{lig} is the entropy lost by the ligand upon binding, and $\Delta G_{reward/penalty}$ represent special rewards and penalties, e.g. $SASA_{ligand}^{buried-on-binding}$. Each one of these terms can be computed using a number of methods found in the literature. For example, ligand–protein interaction and ligand conformational strain can be calculated using several force fields and solvation treatments. The empirical weighting coefficients, c_i's, are determined by fitting to a training set of carbohydrate–protein complexes with known structure and binding affinity. The use of linear regression models, or linear response models, is a recurring theme with several successful examples in the development of free-energy functions (Aqvist and Marelius, 2001; Hansson *et al.*, 1998; Hill and Reilly, 2008; Kerzmann, *et al.*, 2006; Laederach and Reilly, 2003; Lamb *et al.*, 1999; Marelius *et al.*, 2001; Morris, *et al.*, 1998; Rizzo *et al.*, 2002; Smith *et al.*, 1998; Wesolowski and Jorgensen, 2002).

Scanning literature reveals a staggering number of methods for computing each term in the proposed Master Equation. These methods vary in their theoretical derivation, degree of sophistication, and associated computational cost. They could vary from a simple integer representing the number of freely rotatable bonds in the ligand up to a fully-blown free-energy function employing advanced implicit solvent model such as MM/GBSA. Factoring in the number of possible treatments for each term in the Master Equation, the result is a huge number of combinations comprising free-energy components computed using different methods, referred to as *complex descriptors* within this thesis. The descriptors employed in this study to estimate the various free-energy components and details of their computation are described in the Methods section (page 51).

From this point two routes can be followed: (1) mine the existing pool of descriptors to find a valid free-energy function, or (2) develop new descriptors offering alternative methods to calculate one or more of the components of the binding free energy. We decided to initially follow the first route; i.e. start by assessing models of increasing levels of complexity, yet composed of established energy terms, before resorting to the development of more rigorous descriptors or case-specific penalties and/or rewards and incorporate them into completely new models. Moreover, the investigated systems were sampled via molecular dynamics (MD) simulations in explicit solvent to account for conformational flexibility of both ligand and protein and for solvent effects, in a manner analogous to multidimensional (6D-) QSAR concept developed in our group. Customization of free energy terms to compensate for discrepancies in one or more carbohydrate–protein complexes could be viewed as introducing bias to the free energy model, which in turn carries the risk of building an artificial and non-generalizable model. Moreover, there is a huge collection of techniques and algorithms that can simulate and quantify several constituents of molecular interactions, from force fields based on simple ball-and-spring models to the more robust thermodynamic integration methods. Therefore, it seems less likely that, in the search for a free-energy function for carbohydrates, we are more lacking in methodologies for computation of free-energy components. What seems more lacking, in our opinion, is a better understanding of why the traditional free-energy functions do not produce good correlation with experimental results.

Therefore, this study started by investigating the exhaustively enumerated combinations of free-energy terms in the Master Equation and evaluating the ability of the resultant models to predict experimental binding affinities in a data set of carbohydrate–protein complexes, which was carefully compiled and refined beforehand. The primary intention is to search for the right blend(s) of well-established computational tools that could serve as an objective free-energy function for carbohydrate–protein complexes, if such a blend exists. In case of failure, the reason(s) behind this failure were to be thoroughly investigated: is it the improper or complete lack of treatment of one or more free-energy components, or some other reason? We devoted more time trying to find the underlying reasons for the failures and limitations, rather than to come up with a neat arrangement of points in a binding affinity prediction plot.

This study also addressed three relevant issues: (1) target-dependence of scoring functions (Ferrara, *et al.*, 2004; Mooij and Verdonk, 2005; Perola, *et al.*, 2004; Warren, *et al.*, 2006); why is it that certain scoring functions could predict binding affinities accurately in some protein families and fail in others, (2) dynamic nature of carbohydrate–protein interactions, and (3) external validation of the proposed empirical scoring functions. In the end, however, it is necessary to reassert that formulating models to describe and quantify molecular interactions is only possible within certain boundaries, largely due to the highly non-additive nature of the components of interaction energy (Bissantz, *et al.*, 2010). Recognition of the boundaries and limitations of a given binding-affinity model is a fundamental prerequisite for its successful employment in structure-based drug design.

I.3. Methods

I.3.1 Compiling the carbohydrate–protein data set

The assembled collection of carbohydrate–protein complexes, along with their experimental affinities and important metadata (e.g. PDB resolution, molecular weights, corresponding publications, etc.) was stored in a relational database to facilitate searching and filtering later on. The initial goal was to collect *as many carbohydrate–protein complexes as possible* for which experimental binding affinities are available. From this *master database* we could select a suitable set of entries fulfilling certain criteria for use in our development and validation of scoring functions for predicting binding affinities.

In the start, a large pool of ligand–protein complexes was gathered by mining three databases: the Protein Data Bank (http://www.pdb.org) for structural information, and Binding MOAD (http://www.bindingmoad.org/) (Benson, *et al.*, 2008; Hu, *et al.*, 2005) and BindingDB (http://www.bindingdb.org/) (Chen, *et al.*, 2001; Liu, *et al.*, 2007) for binding affinities (Figure I-5). The starting query included general substructure search for ligands having a pyran or furan moieties. The results were combined with those of a keyword search for terms such as “carbohydrate”, “sugar”, and “lectin”. Complexes used previously in similar studies were also included (Hill and Reilly, 2008; Kerzmann, *et al.*, 2006; Laederach and Reilly, 2003). This initial stage resulted in a large collection of over 8’000 entries, which ended up in a *crude* collection of 6’398 candidate entries after removing redundancies.

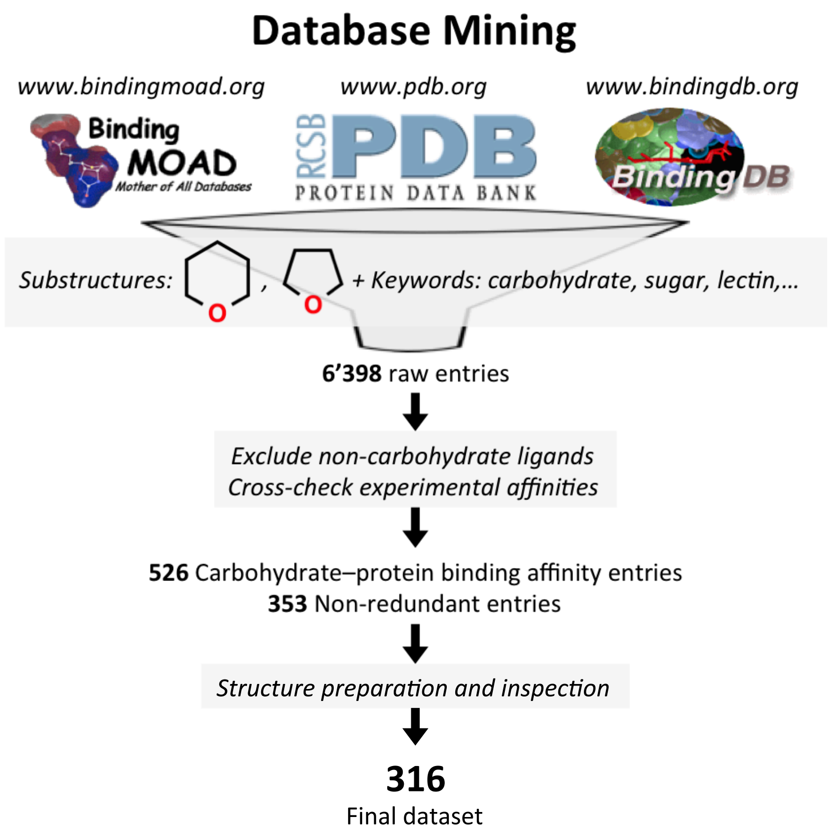

Database Mining
www.bindingmoad.org
www.pdb.org
www.bindingdb.org
Binding MOAD
Mother of All Databases
RCSB PDB
PROTEIN DATA BANK
BindingDB
Substructures:
O
O
+ Keywords: carbohydrate, sugar, lectin,…
6'398 raw entries
Exclude non-carbohydrate ligands
Cross-check experimental affinities
526 Carbohydrate–protein binding affinity entries
353 Non-redundant entries
Structure preparation and inspection
316
Final dataset

Figure I-5: Construction of the carbohydrate–protein affinity database

All the entries were inspected to confirm the existence of a *carbohydrate ligand* (i.e. excluding co-factors and other co-crystallized compounds) correctly linked to experimental binding affinity measurement(s) in Binding MOAD and/or BindingDB. For each ligand–protein pair, the original publication of the experimental affinity was checked to verify the reported value. The crude collection was refined by deleting improper entries such as:

- Entries lacking a carbohydrate ligand
- Entries where the carbohydrate molecule was not the biologically relevant ligand
- Entries where the ligand is covalently bound to the protein
- Entries for which the reported affinity measurement for the carbohydrate–protein pair was incorrectly linked to the carbohydrate ligand in the complex, and the correct value was not found
- Entries for which the reported affinity was EC_{50}
- Redundant entries

In the end, the database included *526* entries of reviewed experimental affinities for carbohydrate–protein complexes. Binding affinities in our collection were K_d's (dissociation constants), K_a's (association constants), K_i's (inhibition constants), or IC_{50}'s (concentrations resulting in 50% enzyme inhibition). It is important to note that the number *526* includes redundant entries for the complexes measured by isothermal titration calorimetry (ITC);

since for these complexes the values of thermodynamic components of ΔG_{bind}; i.e. ΔH and TΔS, were stored in the database as separate entries. All binding affinity values were converted to binding free energies (ΔG, kcal.mol^{-1}) using the thermodynamic master equation ΔG = – RT*ln*K (equation I-2, page 24). The non-redundant set comprised *353* unique carbohydrate–protein complexes with experimental binding affinities (Figure I-6); which were carried forward to the structure preparation step (below). The complete listing of the 353 non-redundant entries is given in Appendix 1, along with references to the primary literature of the affinity measurements and important preparation notes. Structures of the carbohydrate ligands in the studied complexes are listed in Appendix 2.

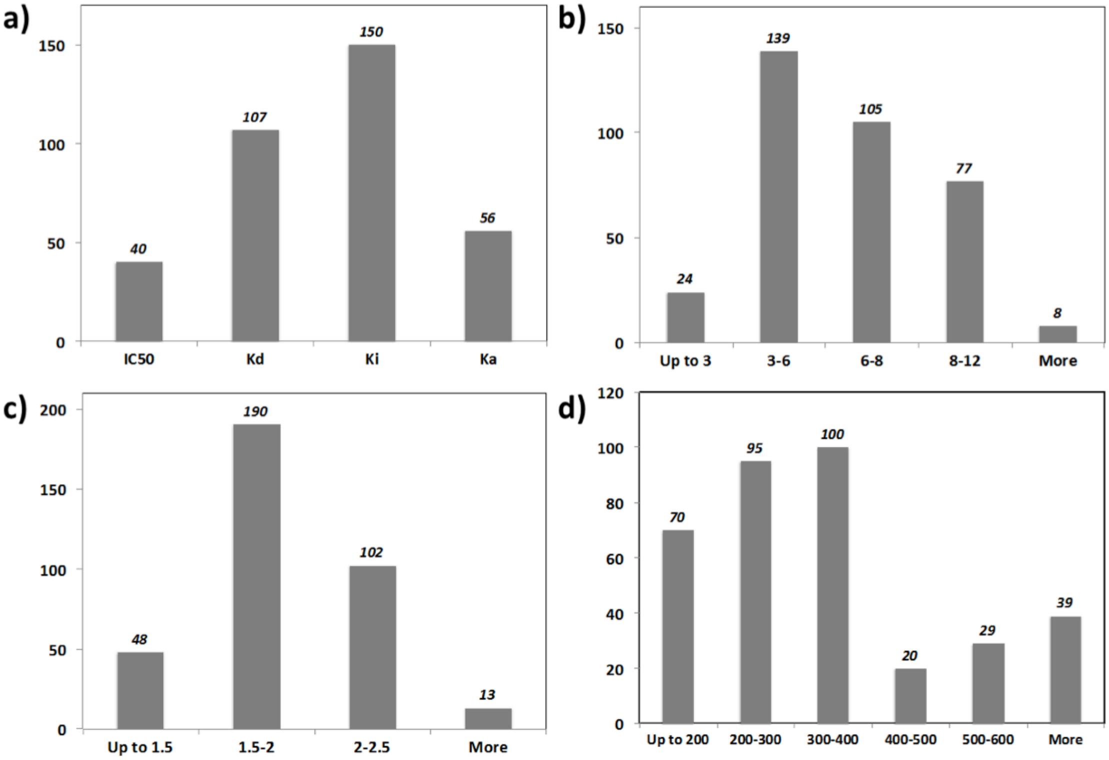


Figure I-6: Histograms of the non-redundant set of carbohydrate–protein binding affinity data (N=353) prior to the structure preparation stage: (a) type of experimental binding affinity, (b) experimental –ΔG_{bind} in kcal/mol, (c) PDB resolution in Å, (d) molecular weight of the ligand

In the final preparation step, where the complexes were scrutinized *one-by-one*, some problems that had not been noticed earlier were revealed:

- Multiple copies of the same ligand in complex differing significantly in conformation and orientation in the binding site, as judged by their heavy-atom RMSD's (see 'Preparing ligand–protein complexes' section for details)
- Polyvalent carbohydrate-binding protein with several ligand recognition subsites, e.g. Fucose-specific lectin (PDB: 1OFZ) and Concanavalin A (PDB: 2D7F)
- Very large ligands (MW > 1'000) that will be troublesome to handle in subsequent modeling stages, e.g. β-cyclodextrin (PDB: 1DMB) and Eritoran (PDB: 2Z65)

- Ligands with missing atoms, i.e. atoms not resolved in the X-ray crystal structure, e.g. digoxin (PDB: 1IGJ)
- Rare atom types; e.g. Selenium in glycosidic linkage of the ligand (PDB: 1O9V) and Copper bound to Belomycin (PDB: 2A4W), that cannot be properly handled by most force fields due to lack of appropriate parameterization
- The same ligand–protein combination deposited in the PDB more than once, e.g. D-gluconhydroximo-1,5-lactam in complex with myrosinase (PDBs: 1E72 and 1E6S), in which case the complex with the higher resolution was used

Nevertheless, these problems do not mean that the binding affinity associated with a particular carbohydrate–protein complex was invalid. Instead, these problems indicate uncertainties in geometry or inability of common force fields to handle some ligand atoms. Thus, these 'problematic' entries were kept in the database but excluded from the *final data set*. The *final data set*, thus, consisted of *316* unique carbohydrate–protein complexes (Figure I-7), which were used to test, develop, and validate all scoring functions discussed here for predicting affinities of putative carbohydrate–protein complexes.

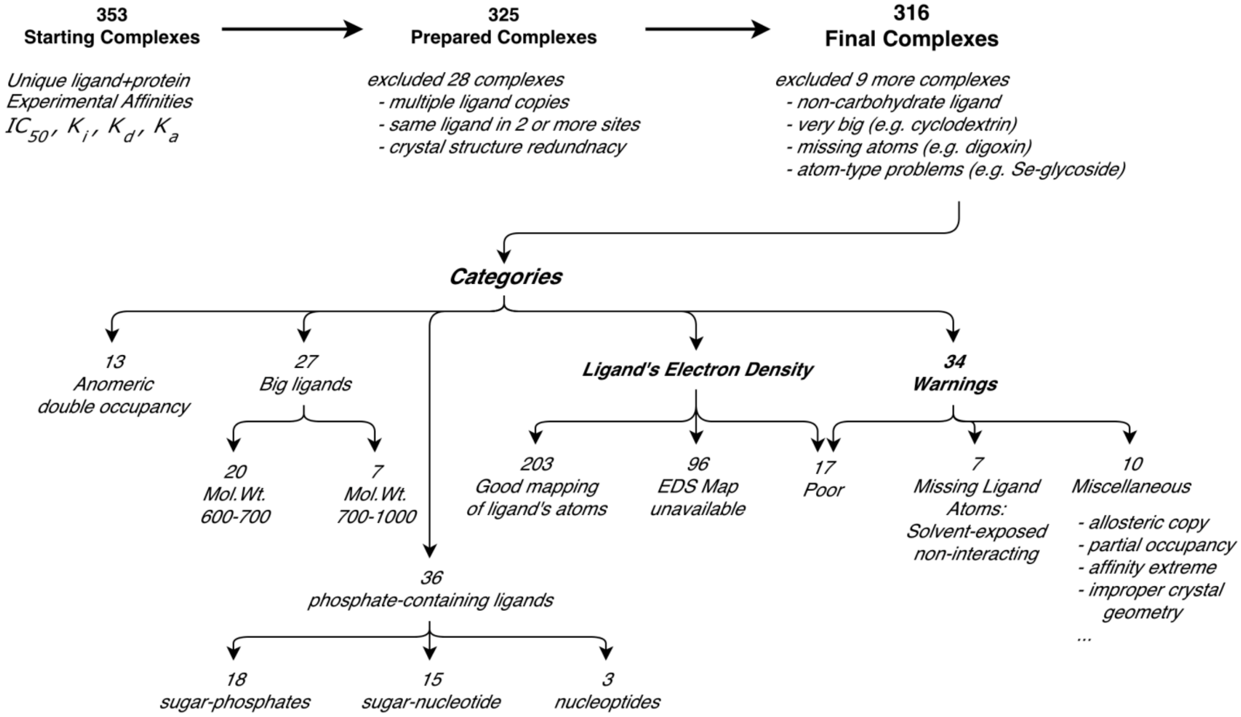


Figure I-7: Categorization of the final data set of prepared carbohydrate–protein complexes highlighting important structural and crystallographic sub-classes.

Electron-density maps for all crystal structures were interactively checked using the online Astex Viewer™ Viewer (Hartshorn, 2002). Ligand's electron density was described as 'poor' if not all ligand atoms in the crystal structure had corresponding electron densities in the EDS map. We found no trends in the relationship between ligand size (molecular weight), protein size, or crystal-structure resolution of a given complex and the quality of the electron density mapping of its ligand atoms (Figure I-8).

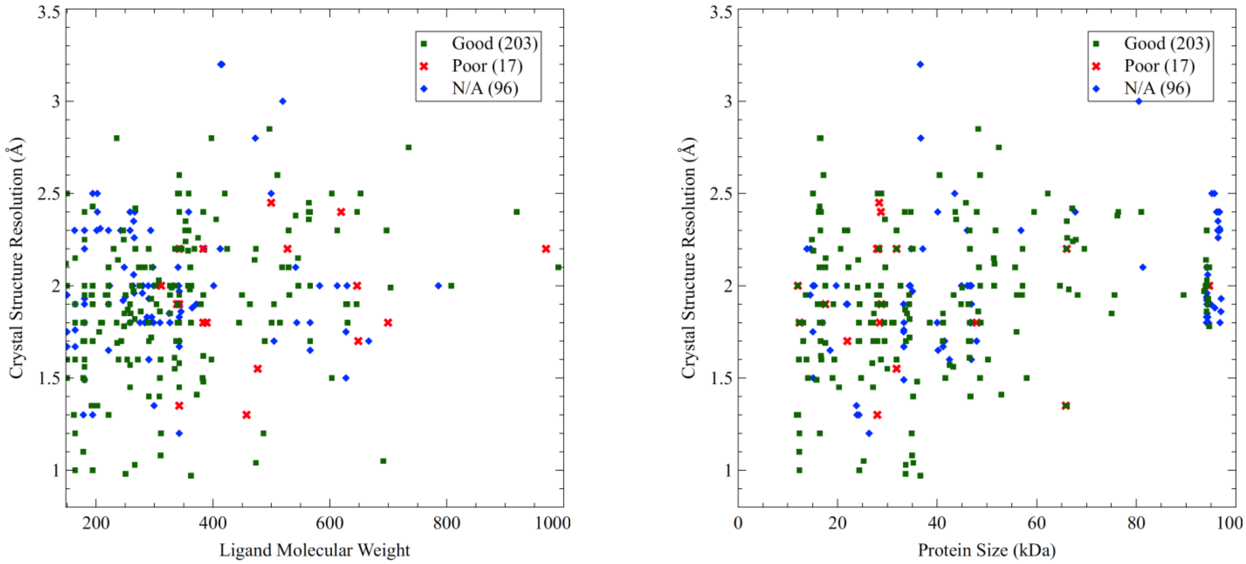


Figure I-8: Quality of EDS mapping of ligand atoms for the studied carbohydrate–protein complexes and its relation to ligand's molecular weight, protein size, and crystal structure resolution. The protein size was calculated for the final prepared complex used in this study.

I.3.2 Preparing ligand–protein complexes

Preprocessing. All ligand protein complexes were downloaded from the Protein Data Bank (www.pdb.org) and processed using Maestro's Protein Preparation Wizard (Maestro, 2011). All hydrogens in input structures were deleted, then bond orders were automatically assigned and hydrogens were added accordingly. Water molecules within 5.0 Å from non-standard residues (e.g. ligands, cofactors, metals) were kept and all other water molecules were deleted. Missing side chains were completed and optimized using Prime (Prime, 2011). Metals in some complexes had zero-order bonds to their ligands; in such cases these bonds were deleted.

Multiple ligand copies. When a complex exhibited multiple chains with several copies of the ligand molecule in the asymmetric unit, the individual chains were superimposed and heavy-atom RMSDs were computed for the ligand and the surrounding residues. In most complexes all the copies had RMSD values within 1.0 Å; in which case the first chain having a resolved ligand was used and its chain identifier was noted. Complexes where ligand copies differed significantly in conformation and/or orientation in the binding site, i.e. RMSD > 1.0 Å were discarded (examples: 1A0T and 1JZ7). In some complexes, the ligand had two overlapping representations, mostly resulting from the α- and β-anomers being simultaneously resolved in the binding pocket. Unless the affinity measurement explicitly refers to the β-anomer, the α-anomer was used in subsequent computations and the β-anomer copy was deleted. In some complexes there was a ligand copy in an allosteric binding site, as indicated in the original publication of the PDB structure. In such cases, we confirmed that the measured affinity was *competitive* by revisiting the respective publication, and subsequently deleted the allosteric copy of the ligand (examples:2QN8 and 2QNB). Before proceeding, we made sure that each complex had one, and *only one*, ligand copy with a unique residue number. Not all complexes had a single chain, however; since in some cases the ligand lies in close proximity to and, hence, could interact with residues from two or more adjacent chains. All processing notes —e.g. retained chains in case of multiple-chain PDB's, deleted ligand copies, etc. — are given in Appendix 1.

Covalent structure and protonation. Each ligand's chemical structure was cross-checked against the corresponding primary citation and inconsistencies resulting from incorrect bond order assignments were corrected manually. Protonation and tautomeric states for all HET groups were automatically assigned using Epik (Shelley *et al.*, 2007). We used the protonation state of the ligand whenever it was explicitly mentioned in the original publication; otherwise the top-ranked state from Epik was used. At this stage we have a fully-atomistic model of the ligand–protein complex, each with a unique ligand molecule with revised chemical structure and protonation state.

Geometry optimization. The last processing step was to optimize the geometry of the ligand–protein complexes. First, the geometry and orientation of all added hydrogen atoms was exhaustively sampled for optimal H-bond formation, including any necessary flipping of glutamine, asparagine, and histidine side chains. Finally, each complex was refined by full minimization using OPLS_2005 force field as implemented in Schrödinger's MacroModel (MacroModel, 2011). Minimization was set to converge within heavy-atom RMSD of 0.3 Å from the input geometry to avoid significant deviations from the experimental geometry.

I.3.3 The NeoScore platform

In the course of this study, several empirical functions varying in degree of sophistication were assessed, aiming to improve our understanding of the relationship between binding affinities of carbohydrates–protein complexes and their geometries, and hence provide a way to reliably predict the former from the latter. We built a computation and statistical analysis platform tailored specifically for this purpose, the NeoScore platform. NeoScore has two main modules; the first is the NeoScore Computation Engine, which processes ligand–protein complexes to extract useful *descriptors* quantifying the relevant geometric and energetic features. The second module, the NeoScore Analysis Engine, is equipped with a variety of statistical methodologies to facilitate building models for predicting binding affinities from the aforementioned *descriptors*, as well as assessment and validation of these models. The NeoScore platform was programmed using Python Programming Language (www.python.org) and employing object-oriented programming style to present a simple calling interface to users.

I.3.3.1 The NeoScore Computation Engine

The input for the NeoScore Computation Engine (Figure I-9) is a ligand–protein complex in any recognized format. The PDB, SD, and Maestro structure file formats are supported natively, and more formats are supported through automated conversion into one of the native formats using the Open Babel Package (Babel, 2012; O'Boyle *et al.*, 2011). Basic information about the processed ligand–protein complex (e.g. PDB ID, ligand identifier, experimental binding affinity type and value) can be automatically extracted from the input file, or provided by the user. The NeoScore Engine stores complexes and associated properties in a special format; the Complex Descriptors file (cdr file).

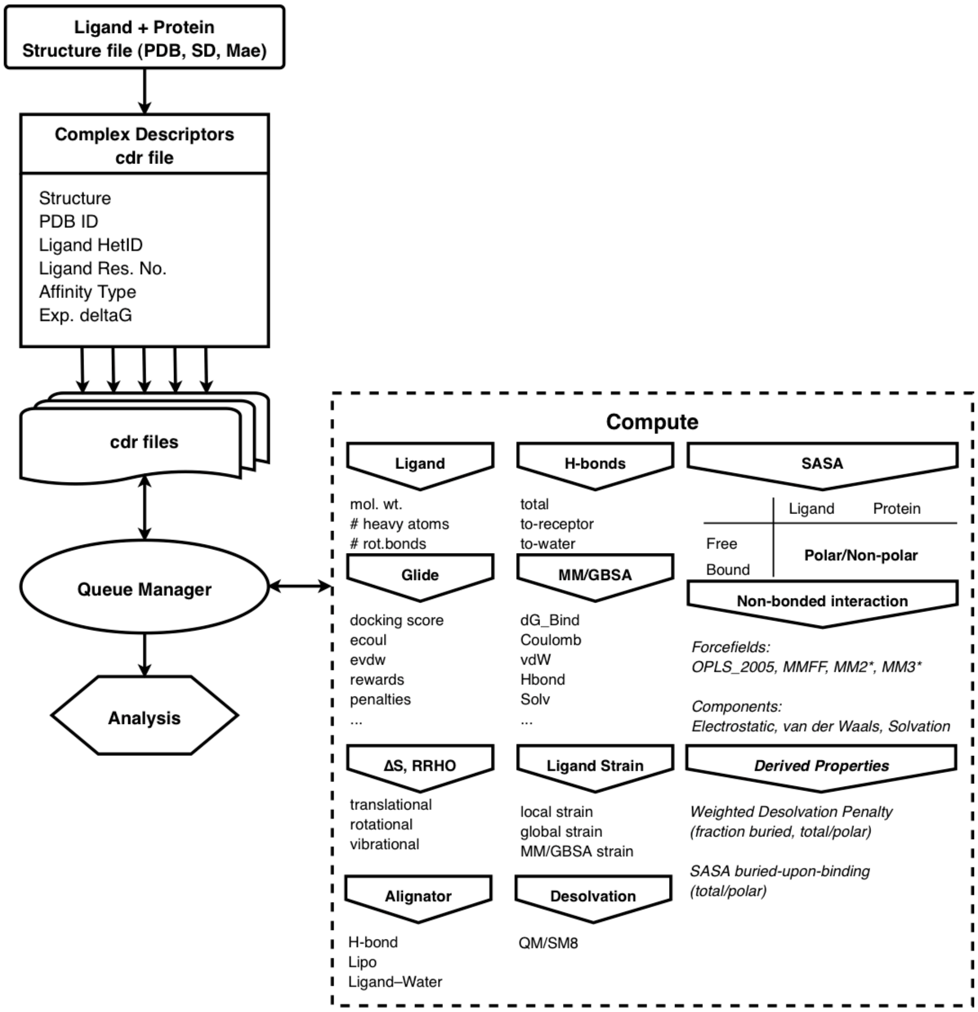


Figure I-9: Flowchart of the NeoScore Computation Engine. The flow starts by reading structure file of the ligand–protein complex to be processed and storing the coordinates and associated properties in a Complex Descriptors file (CDR) native to NeoScore. The NeoScore Computation engine provides interfaces to a multitude of geometric and energetic properties of the studied complex.

The Computation Engine offers simple interfaces to calculate numerous geometric and energetic properties of ligand–protein complexes. The currently supported properties are summarized in Figure I-9, and they will be discussed in more details in a subsequent section. Architecture of the NeoScore platform, including the Computation Engine, is highly modular, allowing for easy extension to support more *structure-based* descriptors or additional statistical analysis tools. Thus, virtually any property calculated from the coordinates of ligand–protein complexes can be easily plugged into our NeoScore platform. Moreover, NeoScore Computation Engine has a built-in queue manager capable of running multiple computations jobs in parallel. NeoScore can, thus, achieve higher computational efficiency on multi-core processors available in most affordable computers nowadays.

I.3.3.2 The NeoScore Analysis Engine

In the realm of developing scoring function for carbohydrate–protein modeling, the complex descriptors calculated by the NeoScore Computation Engine can be considered as predictor (independent) variables, while the experimental affinity (ΔG) plays the role of the dependent variable. The aim of this work could be summarized as trying to find a mathematical relationship (an *equation*) capable of predicting the latter quantity (ΔG's) given the former (complex descriptors). In addition to its predictive power, this *equation* should be statistically valid and physicochemically meaningful, in order to have any useful application in structure-based drug design. All the *equations* examined in the course of this study consist of a linear combination of terms, or complex descriptors, characterizing components of the free energy of the ligand–protein binding process.

To achieve this goal, it was necessary to evaluate many, thousands in fact, mathematical formulations varying in composition (terms) and degree of sophistication. The central idea of the NeoScore Analysis Engine is to make this process of *hypothesis testing* as fast and convenient as possible. It does this by providing a simple interface for statistical investigations on a combination of (a subset of) ligand–protein complexes and (a subset of) their descriptors (Figure I-10). Moreover, the Analysis Engine can compute new descriptors from existing ones, such as buried-SASA-weighted desolvation penalties. It can also create categories by applying user-defined classification criteria based on available descriptors, e.g. surface-exposed vs. buried binding sites based on receptor SASA buried-upon-binding. Additionally, the Analysis Engine has integrated support for genetic algorithm optimization, which is discussed in more details below.

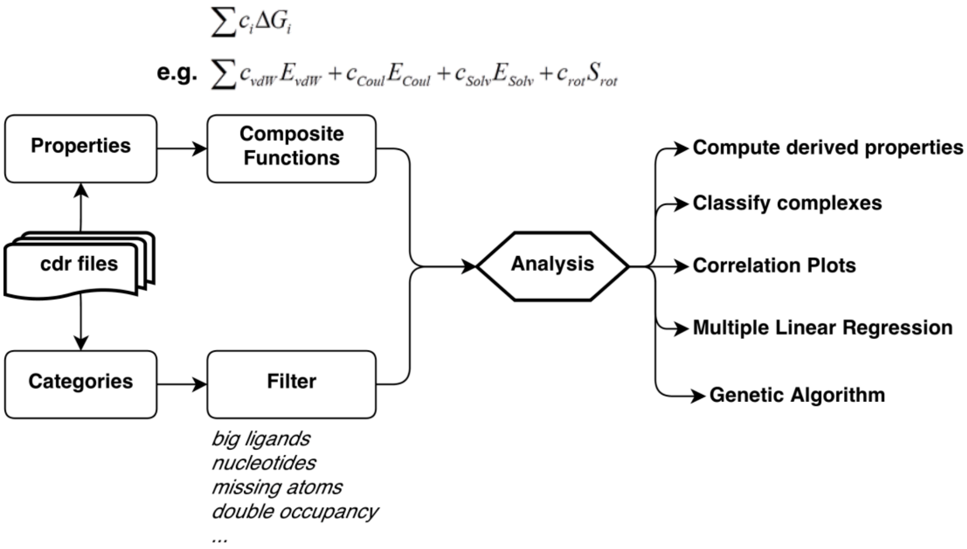


Figure I-10: Flowchart of the NeoScore Analysis Engine.

I.3.3.3 Genetic algorithm

The search space in our study could be described as a collection of entries (ligand–protein complexes) and their associated properties (complex descriptors). The *solution* would, thus, be an equation incorporating a suitable number of terms that can satisfactorily describe the variation in binding affinities of the set of complexes being considered. We can demonstrate the complexity of trying to find this solution by drawing an example. The pool of descriptors we have at our disposal is close to 250; each is a viable candidate term in the free-energy equation we want to construct. Assuming that this equation has 10 terms, the number of possible combinations from the pool of 250 terms would be $^{250}C_{10}$, roughly 200 million billion possibilities. Thus, an exhaustive *brute force* search, i.e. evaluating each and every one of these possibilities, is practically intractable. Luckily, there are a number of ways to approach similar problems; genetic algorithm is one of them.

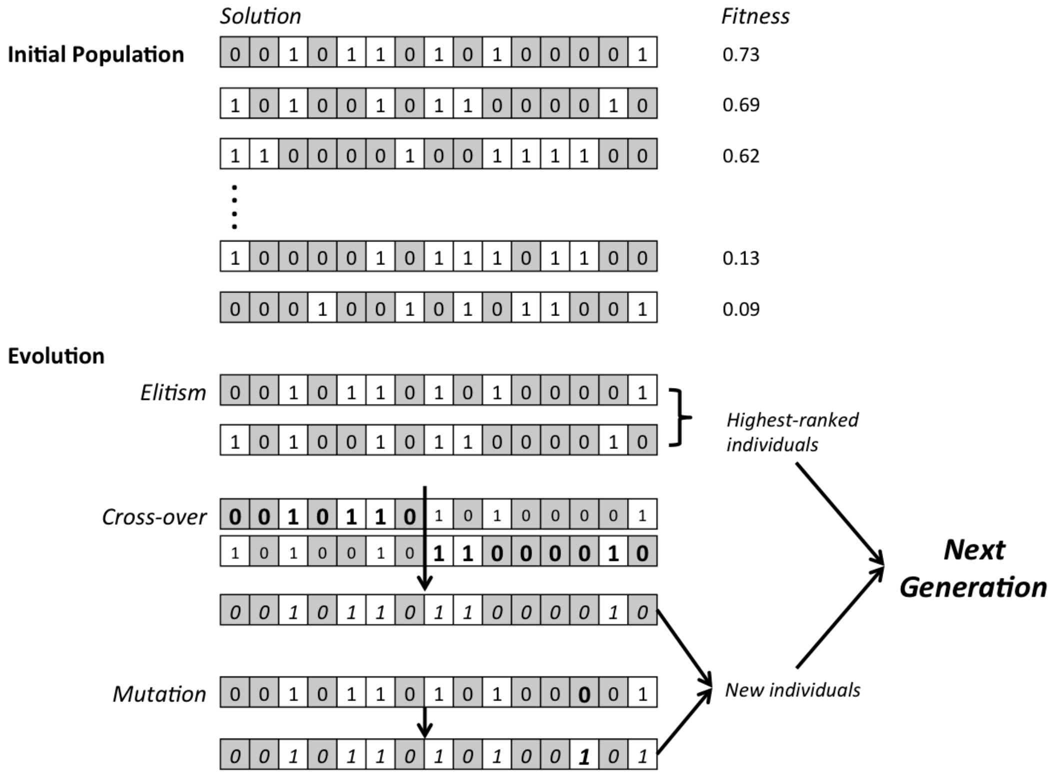


Figure I-11: Genetic algorithm starts by a randomly generated population of chromosomes representing the problem being solved. This population is evolved through several generations to improve quality of the solution. Selecting the best-fit individuals creates the necessary evolutionary pressure. New child members are introduced from fit parents by crossover and mutation mimicking natural genetic evolution.

Genetic algorithm (GA) is an iterative optimization technique inspired by the natural evolutionary processes associated with passing genetic material from parents to their offspring. The basic idea is to randomly generate an initial *population*, whose members are candidate solutions to the problem, and *evolve* that population under appropriate *selection pressure* to obtain a better solution. The process starts by representing the search domain by *chromosomes* that can be mutated and altered (Figure I-11). In our case the chromosomes were binary, i.e. consisting of strings of zero's and one's, with the zero denoting an ignored descriptor and/or ligand–protein complex. Putative solutions (also called individuals) are rank-ordered using a *fitness function*. A new population is then created starting by the top-ranked individuals from the previous population. This process is called *elitism*, and it insures survival of the fittest parents from one generation to the next.

Although *elitism* is sufficient to create the necessary evolutionary pressure, it does not introduce any diversity to newer generations. Novel members are introduced in a new generation by mutating and breeding highly-ranked members of the preceding generation. *Mutation* is small simple random change involving one *gene* in the chromosome, e.g. by changing a zero into a one or vice versa. Breeding, on the other hand, is accomplished by combining two parent chromosomes to create a new child chromosome by a *crossover* operation, where a number of genes is taken from one parent and the rest of the genes is taken from the other parent. The whole process of selection, mutation, and crossover repeats either for a preset number of cycles, or until no improvement in the fitness function is observed for a number of generations (Figure I-11).

Fitness functions. The NeoScore Analysis Engine provides two options for the fitness function; the adjusted-*r*^{2} ($\bar{r}^2$) and the root-mean-square-error (RMSE). The adjusted-*r*^{2} penalizes models that have more explanatory terms than necessary (equation I-6). Unlike *r*^{2} (coefficient of determination), the adjusted-*r*^{2} does not increase unless a new term in the equation improves the model more than would be expected by chance. The adjusted-*r*^{2} can have negative values, and its value is always less than or equal to the corresponding *r*^{2}. In the context of GA, evolutionary pressure is applied in the direction of *maximizing* the adjusted-*r*^{2}, whereas it aims to *minimize* the RMSE (equation I-7).

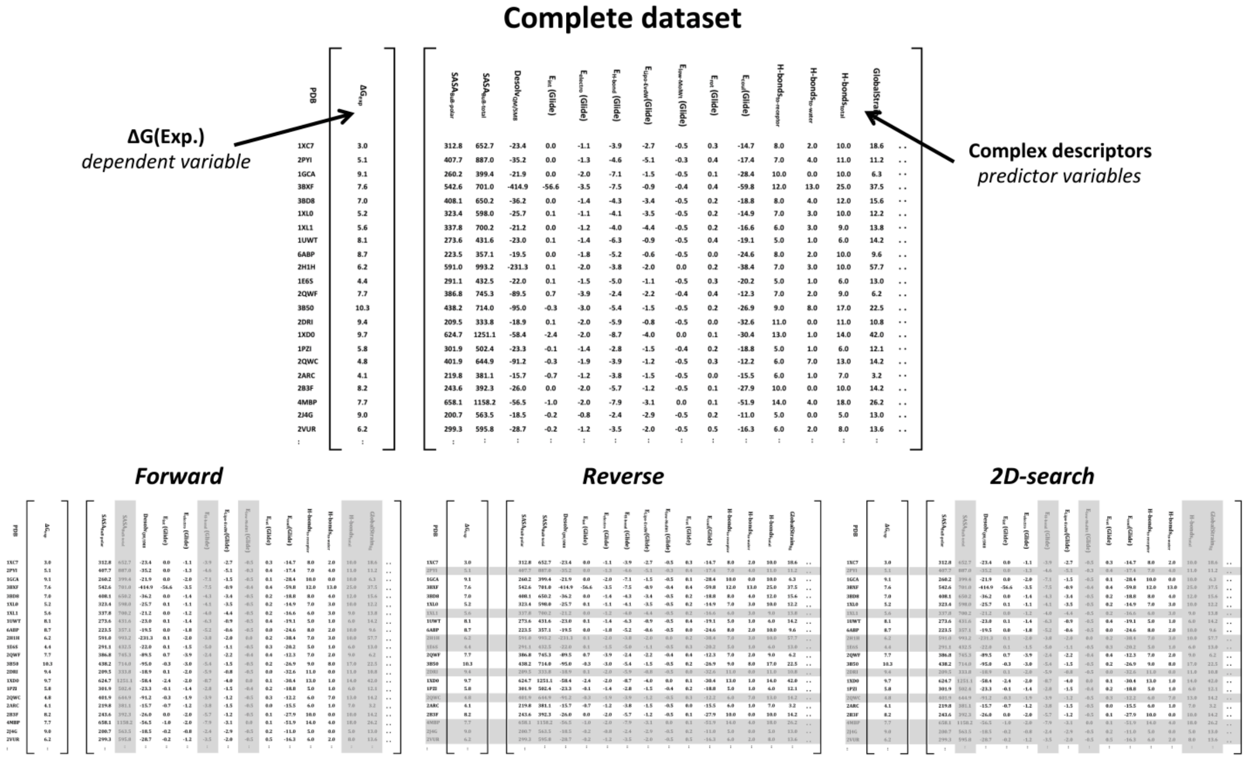
$$\bar{r}^2 = 1 - \frac{N-1}{N-k-1}(1-r^2) \qquad \textbf{I-6}$$

K *Total number of independent (estimator) variables*
N *Sample size*

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\Delta G_{i,exp} - \Delta G_{i,calc}\right)^2}{N}} \qquad \textbf{I-7}$$

$\Delta \boldsymbol{G}_{\boldsymbol{i,exp}}$ *Actual value for the i-th observation*
$\Delta \boldsymbol{G}_{\boldsymbol{i,calc}}$ *Estimated value for the i-th observation*
N *Sample size*

Application of GA in our study. In essence, GA is a solution-finding algorithm. In this study, the *solution* is a combination of equation terms (complex descriptors) capable of predicting binding affinities for a specific set of carbohydrate–protein complexes. GA is implemented in the framework of the NeoScore Analysis Engine to perform model mining in three different modes or directions. The first "forward mode" reads a fixed set of complexes (observations) and searches the pool of available complex descriptors (predictor variables) for the best subset that describes the variation in observed affinities. Additionally, we customized the GA module to perform a "reverse mode" search, i.e. find the best subset of complexes (observations) whose affinity can be accurately predicted by a pre-defined set of predictor variables. The third mode, the "2D-search mode", is a combination of the two previous modes, where the selections of a subset of complexes and a subset of complex descriptors are optimized simultaneously.



PDB	ΔG_{exp}	$SASA_{BuB\text{-}polar}$	$SASA_{BuB\text{-}total}$	$Desolv_{QM/SM8}$	E_{int} (Glide)	$E_{electro}$ (Glide)	$E_{H\text{-}bond}$ (Glide)	$E_{Lipo\text{-}EvdW}$(Glide)	$E_{low\text{-}MoWt}$ (Glide)	E_{rot} (Glide)	E_{coul}(Glide)	H-bonds$_{to\text{-}receptor}$	H-bonds$_{to\text{-}water}$	H-bonds$_{total}$	GlobalStrain	
1XC7	3.0	312.8	652.7	-23.4	0.0	-1.1	-3.9	-2.7	-0.5	0.3	-14.7	8.0	2.0	10.0	18.6	. .
2PYI	5.1	407.7	887.0	-35.2	0.0	-1.3	-4.6	-5.1	-0.3	0.4	-17.4	7.0	4.0	11.0	11.2	. .
1GCA	9.1	260.2	399.4	-21.9	0.0	-2.0	-7.1	-1.5	-0.5	0.1	-28.4	10.0	0.0	10.0	6.3	. .
3BXF	7.6	542.6	701.0	-414.9	-56.6	-3.5	-7.5	-0.9	-0.4	0.4	-59.8	12.0	13.0	25.0	37.5	. .
3BD8	7.0	408.1	650.2	-36.2	0.0	-1.4	-4.3	-3.4	-0.5	0.2	-18.8	8.0	4.0	12.0	15.6	. .
1XL0	5.2	323.4	598.0	-25.7	0.1	-1.1	-4.1	-3.5	-0.5	0.2	-14.9	7.0	3.0	10.0	12.2	. .
1XL1	5.6	337.8	700.2	-21.2	0.0	-1.2	-4.0	-4.4	-0.5	0.2	-16.6	6.0	3.0	9.0	13.8	. .
1UWT	8.1	273.6	431.6	-23.0	0.1	-1.4	-6.3	-0.9	-0.5	0.4	-19.1	5.0	1.0	6.0	14.2	. .
6ABP	8.7	223.5	357.1	-19.5	0.0	-1.8	-5.2	-0.6	-0.5	0.0	-24.6	8.0	2.0	10.0	9.6	. .
2H1H	6.2	591.0	993.2	-231.3	0.1	-2.0	-3.8	-2.0	0.0	0.2	-38.4	7.0	3.0	10.0	57.7	. .
1E6S	4.4	291.1	432.5	-22.0	0.1	-1.5	-5.0	-1.1	-0.5	0.3	-20.2	5.0	1.0	6.0	13.0	. .
2QWF	7.7	386.8	745.3	-89.5	0.7	-3.9	-2.4	-2.2	-0.4	0.4	-12.3	7.0	2.0	9.0	6.2	. .
3B50	10.3	438.2	714.0	-95.0	-0.3	-3.0	-5.4	-1.5	-0.5	0.2	-26.9	9.0	8.0	17.0	22.5	. .
2DRI	9.4	209.5	333.8	-18.9	0.1	-2.0	-5.9	-0.8	-0.5	0.0	-32.6	11.0	0.0	11.0	10.8	. .
1XD0	9.7	624.7	1251.1	-58.4	-2.4	-2.0	-8.7	-4.0	0.0	0.1	-30.4	13.0	1.0	14.0	42.0	. .
1PZI	5.8	301.9	502.4	-23.3	-0.1	-1.4	-2.8	-1.5	-0.4	0.2	-18.8	5.0	1.0	6.0	12.1	. .
2QWC	4.8	401.9	644.9	-91.2	-0.3	-1.9	-3.9	-1.2	-0.5	0.3	-12.2	6.0	7.0	13.0	14.2	. .
2ARC	4.1	219.8	381.1	-15.7	-0.7	-1.2	-3.8	-1.5	-0.5	0.0	-15.5	6.0	1.0	7.0	3.2	. .
2B3F	8.2	243.6	392.3	-26.0	0.0	-2.0	-5.7	-1.2	-0.5	0.1	-27.9	10.0	0.0	10.0	14.2	. .
4MBP	7.7	658.1	1158.2	-56.5	-1.0	-2.0	-7.9	-3.1	0.0	0.1	-51.9	14.0	4.0	18.0	26.2	. .
2J4G	9.0	200.7	563.5	-18.5	-0.2	-0.8	-2.4	-2.9	-0.5	0.2	-11.0	5.0	0.0	5.0	13.0	. .
2VUR	6.2	299.3	595.8	-28.7	-0.2	-1.2	-3.5	-2.0	-0.5	0.5	-16.3	6.0	2.0	8.0	13.6	. .
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	

Figure I-12: Genetic algorithm can mine data for solutions in three different directions depending on what's being omitted (grey shading) to improve prediction; Forward: omitting predictor variables (complex descriptors), Reverse: omitting observations (complexes), and 2D-search: omitting both. The solution is a combination of observations and predictor variables, in which the latter can accurately predict values of the former.

Excluding outliers and risk of over-fitting. The "Reverse" and "2D-search" modes in NeoScore GA implementation are somewhat non-typical ways of using GA. They were employed in this work for one purpose: detecting outliers. In statistics, an outlier is a data point that deviates significantly from the other members of the studied population or sample. In the framework of our study, however, and *outlier* could be better defined as *a carbohydrate–protein complex that causes significant deterioration in the performance of an otherwise predictive model*. Outliers are automatically identified and excluded during the GA evolution as an *intended* side-effect of the applied pressure towards improving adjusted-r^2 and/or RMSE. All statistics textbooks sound a clear warning against the practice of excluding outliers to boost model performance or enhance data consistency. In our case, automated exclusion of outliers poses the risk of generating artificially *over-fit* models (scoring functions) lacking predictive power and/or physicochemical soundness.

One of the common reasons for existence of outliers are experimental or measurement errors. When such errors are confirmed, e.g. by repeating the measurement, the corresponding observation (the outlier) can be safely excluded without fear of being frowned upon by statisticians. In this study, however, reevaluating any data point is rather infeasible, since this would require repeating the X-ray crystallography and/or the binding affinity measurement. That's why we needed to implement the non-classical GA modes (reverse and 2D-search) as efficient means of detecting outliers. To minimize the associated risk of data over-fitting, we used GA for outlier identification only as a last resort and always adhered to the following self-restraints:

- First and foremost, the primary objective for using GA to find outliers was to *investigate* rather than to *improve*. Many thermodynamically sound equations that were tested showed only modest ability to predict binding affinities of the provided data set. The question of 'which complexes are not adequately explained by this model/equation?' cannot be answered by trial and error, since the number of possible combinations resulting from all possible omissions is prohibitively large. Therefore, GA offered an attractive and efficient alternative for identifying these "outliers". The *outlier* ligand–protein complexes were interactively inspected to look for any peculiar structural components or special inter- or intra-molecular interactions both favorable and unfavorable, which gave us insight into deficiencies of the scoring function being investigated.
- The number of excluded complexes was kept to a minimum. In most cases, a 10% threshold was used for outlier exclusion.
- All models were subjected to rigorous validation using traditional statistical methods; including cross-validation r^2 (q^2), scrambling of response variable (binding affinity), as well as random allocation of complexes to sub-categories (see Topological classification of binding sites section, page 66).
- In all cases, models lacking physicochemical sense were not accepted. Examples include models where binding affinities and favorable ligand–protein interactions (e.g. number of hydrogen-bonds) are negatively correlated, as well as models where unfavorable binding events (e.g. desolvation) have positive impact on binding affinities.
- Performance of the final models was validated using an external test set comprising 106 FimH ligands and compared against a well-established predictive 6D-Quasar model published recently (Eid *et al.*, 2013).

I.3.4 Complex descriptors

A *complex descriptor* is a quantity measuring some energy-based or geometric feature of a given ligand–protein complex. In the context of this study, they serve as the building blocks of our empirical scoring functions. Such *terms* are abundantly available from well-established force fields, scoring functions, and free-energy functions; and they can be employed as complex descriptors right out-of-the-box. These *terms* cover a fairly wide range of physiochemical and energetic features of ligand–protein complexes, and many of them have been explored in similar binding affinity studies with varying degrees of success. In our investigation, we employed several terms describing components of the binding free energy according to the thermodynamic decomposition shown in Figure I-4 (page 35).

More time was spent *investigating* (combinations of) existent descriptors, varying in theoretical derivation and degree of sophistication, rather than *inventing* novel or more sophisticated descriptors. This decision was made in an attempt to seek an *improved understanding* of the factors governing the strength of carbohydrate–protein binding, which is likely more lacking than *deriving* new terms quantifying one or more aspects of that binding process. Complex descriptors employed in our study will be discussed in the following sections, highlighting their underlying theoretical foundations and details of the employed computational procedures.

I.3.4.1 Non-bonded interactions from force fields

Employment of the non-bonded interaction terms from molecular mechanics force fields in docking and scoring could be traced back to the earliest docking program, DOCK (Kuntz, *et al.*, 1982). The AMBER force field (Weiner, *et al.*, 1984), for instance, provides the basis for scoring functions in DOCK 4.0 (Ewing, *et al.*, 2001) and AutoDock (Huey, *et al.*, 2007; Rosenfeld, *et al.*, 2003). Molecular mechanics force fields typically model non-bonded interactions as a sum of van der Waals and Columbic terms:

$$E_{non-bonding} = \sum_{i \neq j}^{non-bonded} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + \left(\frac{q_i q_j}{\varepsilon r_{ij}} \right)$$

where A_{ij} and B_{ij} are the vdW parameters, r_{ij} is the distance between the non-bonded atoms i and j, q_i and q_j are the atomic charges, and ε_{ij} is the dielectric constant of the medium. Force fields differ in the way they account for solvent effects; some use a distance-dependent dielectric constant, while others employ more rigorous implicit solvent models, e.g. generalized-Born/surface area (GB/SA) model (Still, *et al.*, 1990).

Over the past few decades, an increasing number of force fields developed and/or optimized for modeling carbohydrates have been reported (Foley, *et al.*, 2012). Carbohydrate force fields can be roughly classified according to their design philosophy; some force fields are developed for the purpose of simulating large biomolecular systems, e.g. OPLS-AA (Damm *et al.*, 1997; Kony *et al.*, 2002) and GLYCAM06 (Kirschner, *et al.*, 2008), while others are more tuned for small organic molecules, e.g. MM3 (Allinger *et al.*, 1990). The former class of force fields is obviously more suited for the purpose of our study.

The first force field employed in our study was OPLS_2005, the MacroModel implementation of the OPLS-All-Atom force field (Kaminski *et al.*, 2001; MacroModel, 2011). The OPLS (Optimized Potentials for Liquid Simulations) force field was originally optimized for protein simulations (Jorgensen and Tirado-Rives, 1988) and updated later to an All-Atom variant, OPLS-AA (Jorgensen *et al.*, 1996). Shortly afterwards, it was extended to carbohydrates by refitting some of the parameters to *ab initio* results for complete hexopyranoses (Damm, *et al.*, 1997) and by applying additional scaling factors for the 1,5 and 1,6 electrostatic interactions (Kony, *et al.*, 2002). OPLS-AA-driven molecular dynamics simulations have been successfully employed for studying carbohydrate-protein interactions (Margulis, 2005; Sharma and Vijayan, 2011). It is important to note, however, that the OPLS force field does not use explicit terms to describe hydrogen bonds and has no additional interaction sites for lone pairs (Jorgensen and Tirado-Rives, 1988). Neverthless, the proven success of OPLS in reproducing experimental and *ab initio* parameters of small organic molecules as well as larger biomolecules gives confidence in the reliability of the non-bonded interaction terms employed in the potential energy function, especially when used in conjuction with TIP3P and TIP4P water models (Damm, *et al.*, 1997; Jorgensen *et al.*, 1983; Jorgensen and Madura, 1985; Jorgensen, *et al.*, 1996; Jorgensen and Tirado-Rives, 1988).

Moreover, the non-bonded interaction terms from MMFFs, the MacroModel implementation of the MMFF94s force field, were included (Halgren, 1996a; Halgren, 1996b; Halgren, 1996c; Halgren, 1996d; Halgren, 1999a; Halgren, 1999b; Halgren and Nachbar, 1996). The Merck molecular force field (MMFF) was parameterized using a wide variety of chemical systems, and targets simulation of small molecules as well as proteins and biological systems. The MMFF94s variant enforces planarity around sp^2 hybridized nitrogens. The chemical classes included in MMFF94 core parameterization do not include carbohydrates, though. We included MMFFs as a general-utility biomolecular force field to compare performance against OPLS-AA, which is optimized for carbohydrates. In analogy to the OPLS-AA force field, the MMFFs force field describes hydrogen-bonding interactions by adjusting key vdW and electrostatic parameters to better fit scaled intermolecular-interaction energies and geometries obtained from high level *ab initio* calculations (Halgren, 1996a; MacKerell and Karplus, 1991).

Non-bonded interaction energy descriptors employed in our study are listed in Table I-2. Each component was calculated by performing a single-point energy calculation using the respective force field on the ligand–protein complex, the protein alone, and the ligand alone. Subsequently, the non-bonded interaction energies were calculated from the single-point energy components (electrostatic, van der Waals, and solvation) using the formula:

$$E_{non-bonded} = E_{complex} - (E_{ligand} + E_{receptor})$$

Table I-2: Complex descriptors from the non-bonded interaction energy components from two force fields, OPLS_2005 and MMFFs. Energy values are in kcal/mol.

Descriptor	Meaning
NB_OPLS_2005_Electrostatic_Energy	Coulomb interaction component from OPLS_2005 force field
NB_OPLS_2005_Van_der_Waal_Energy	Van der Waals interaction component from OPLS_2005 force field
NB_OPLS_2005_Solvation_Energy	Solvation component from OPLS_2005 force field, computed using GS/SA implicit solvent approximation and water as the solvent
NB_OPLS_2005_Total	Sum of electrostatic, van der Waals and solvation non-bonded interaction energies from OPLS_2005
NB_MMFFs_Electrostatic_Energy	Coulomb interaction component from MMFF94s force field
NB_MMFFs_Van_der_Waal_Energy	Van der Waals interaction component from MMFF94s force field
NB_MMFFs_Solvation_Energy	Solvation component from MMFF94s force field, computed using GS/SA implicit solvent approximation and water as the solvent
NB_MMFFs_Total	Sum of electrostatic, van der Waals and solvation non-bonded interaction energies from MMFF94s

I.3.4.2 MM/GBSA free-energy function

Treatment of solvent effects in ligand–protein simulations is a major challenge for force field-based methods (Huang, *et al.*, 2010). The solvent effect could be modeled implicitly using a distance-dependent dielectric constant, e.g. in the Amber-based DOCK scoring function (Meng, *et al.*, 1992). Although computationally efficient, this approach neglects the desolvation effect and could potentially over-estimate binding of highly charged ligands. At the other end of the spectrum, rigorous methods such as free energy perturbation (FEP) (Jorgensen and Thomas, 2008; Zwanzig, 1954) or thermodynamic integration (TI) (Rodinger, *et al.*, 2005) simulate the solvent molecules explicitly. The high computational demand of these methods, however, makes them inadequate for high-throughput virtual screening.

The combined Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) approach offers a good compromise between computational efficiency and prediction accuracy. In MM/GBSA, the solvent is considered as a continuum electrostatic. Polar interactions of the solutes (ligand and protein) are computed using the generalized Born model of Still *et al.* (Still, *et al.*, 1990), where solute molecules (bearing discrete atomic charges) as considered as regions of low dielectricity embedded in a medium of high dielectric constant (Still, *et al.*, 1990; Warshel and Papazyan, 1998). The non-polar contribution to desolvation is considered to be proportional to the solvent-accessible surface area of the binding partners buried upon binding. Finally, the free energy of binding is calculated according to the following equation:

$$\begin{aligned}\Delta G_{bind} &= \Delta E_{MM} + \Delta G_{GBSA} - T.\Delta S \\ &= \Delta E_{bonded} + \Delta E_{vdW} + \Delta E_{coul} + \Delta G_{solv,p} + \Delta G_{solv,np} - T\Delta S\end{aligned}$$

where ΔE_{MM} is the difference in molecular-mechanics energy between the complex and its individual components, ligand and protein, decomposed into bonded contributions, ΔE_{bonded} from bond, angle, and torsion terms of the employed force field and non-bonded contributions, ΔE_{vdW} and ΔE_{coul}. The $\Delta G_{solv,p}$ and $\Delta G_{solv,np}$ are the polar and non-polar contributions, respectively, from the GBSA solvation model. The entropy term, T.Δ*S*, is sometimes discarded in congeneric series and in relative free-energy calculations (Hou *et al.*, 2011a; Hou, *et al.*, 2011b; Rastelli, *et al.*, 2010).

The MM/GBSA approach has been employed as an *end point* approach to post-score docking poses (Ewing, *et al.*, 2001; Moustakas *et al.*, 2006; Zou, *et al.*, 1999). Recently, Greenidge *et al.* assessed the quality of binding free-energy predictions of MM/GBSA on a set of 855 ligand–protein complexes extracted from the PDBbind data set (Greenidge *et al.*, 2013). They showed that, despite the huge diversity in the employed set, MM/GBSA achieved reasonable correlation (r^2=0.63) with experimental binding affinities. However, the study highlighted some limitations of applying MM/GBSA to estimate binding affinity; including sensitivity to imperfections in input structures, need for accurate treatment of entropy and ligand strain, and inability to handle metal ions and water-bridged interactions. The use of ensemble averages from MD simulations could help overcome some of these limitations. There are numerous examples of successful employment of molecular-dynamics in conjunction with MM/GBSA to study the thermodynamics of ligand–protein binding (Hou, *et al.*, 2011a; Hou, *et al.*, 2011b; Rastelli, *et al.*, 2010; Sadiq *et al.*, 2010; Srivastava and Sastry, 2012).

In this study, MM/GBSA calculations were done in Prime (Prime, 2011) considering residues within 8.0 Å zone from the ligand as flexible. Prime implementation calculates the GBSA contribution using the VSGB 2.0 energy model (Li *et al.*, 2011) and calculates the molecular-mechanics energy using the OPLS-AA force field (Damm, *et al.*, 1997; Jorgensen, *et al.*, 1996; Kony, *et al.*, 2002). The VSGB 2.0 model includes physics-based correction terms for improved handling of π-π stacking, hydrogen-bonding interactions, hydrophobic interactions, and self-contacts of the side chains of certain residues. The VSGB 2.0 model employs the Surface Generalized Born (SGB) model (Ghosh, *et al.*, 1998; Yu, *et al.*, 2006) in conjunction with a variable dielectric (VD) treatment to account for polarization effects from protein side chains by varying the internal dielectric constants from 1.0 to 4.0 (Zhu, *et al.*, 2007). The MM/GBSA-derived ΔG_{bind} and its components were included as complex descriptors in this study (Table I-3).

Table I-3: Complex descriptors from Prime MM/GBSA calculation (Prime, 2011) employing corrections in the VSGB 2.0 energy model (Li, *et al.*, 2011).

Descriptor	**Meaning**
mmgbsa-dG_Bind	MM-GBSA free energy of binding, including strain energy
mmgbsa-dG_Bind(NS)	MM-GBSA free energy of binding, without including strain energy
mmgbsa-dG_Bind_Coulomb	Contribution of Coulombic interactions to the MM-GBSA free energy of binding
mmgbsa-dG_Bind_Covalent	Contribution of covalent interactions (i.e. bonded force field terms) to the MM-GBSA free energy of binding
mmgbsa-dG_Bind_vdW	Contribution of van der Waals interactions to the MM-GBSA free energy of binding
mmgbsa-dG_Bind_Solv	Contribution of the generalized Born electrostatic solvation energy to the MM-GBSA free energy of binding
mmgbsa-dG_Bind_Lipo	Contribution of the solvation energy from non-polar surface area to the MM-GBSA free energy of binding
mmgbsa-dG_Bind_Hbond	Hydrogen-bonding corrections to the MM-GBSA free energy of binding (Li, *et al.*, 2011)
mmgbsa-dG_Bind_Packing	Corrections for π-π interactions to the MM-GBSA free energy of binding (Li, *et al.*, 2011)
mmgbsa-dG_Bind_SelfCont	Corrections to the MM-GBSA free energy of binding due to self-contacts between side chains of Asn, Gln, Ser, and Thr and with their own backbone nitrogen or oxygen atoms (Li, *et al.*, 2011)
mmgbsa-Lig_Strain_Energy	Ligand strain energy (compared to the energy minimized geometry)
mmgbsa-Rec_Strain_Energy	Receptor strain energy (compared to the energy minimized geometry)
mmgbsa_total_strain	Sum of Ligand + Receptor strain energies

I.3.4.3 Glide XP Score

Glide (Grid-based Ligand Docking with Energetics) is a widely used software package for docking small-molecule ligands to macromolecular protein targets (Friesner, *et al.*, 2004; Glide, 2011; Halgren, *et al.*, 2004). Several studies have shown Glide's superiority in predicting and properly ranking binding configurations of carbohydrate ligands to their protein targets (Agostino, *et al.*, 2009; Alexacou *et al.*, 2008; Nurisso *et al.*, 2008). Glide employs a rigid-receptor approximation where ligands are docked into a single static representation of the binding site residues. Glide allows for induced-fit effects either modestly via down-scaling of the van der Waals radii of non-polar ligand/protein atoms or through a specialized induced-fit protocol (Farid *et al.*, 2006; Sherman *et al.*, 2006). The Glide docking algorithm uses a series of hierarchical filters to search for optimal poses of the ligand in the active-site region of the receptor (protein), collectively referred to as the "Glide docking funnel" (Friesner, *et al.*, 2004). The automatically generated set of initial conformations is pre-screened to reduce the number of poses undergoing the subsequent energy and gradient-evaluations steps, which are more computational expensive. The authors claim that this design approximates an exhaustive systematic search within acceptable computation time (Friesner, *et al.*, 2004).

The starting point for scoring in Glide is the empirical ChemScore scoring function, which encompasses lipophilic, hydrogen-bonding, and metal terms as well as a penalty term for freezing rotatable bonds of the ligand (Eldridge, *et al.*, 1997). Glide implements a modified and expanded version of the ChemScore scoring function, *GlideScore*, to predict binding affinity and to rank order ligands in database searches. As shown in equation I-8, GlideScore embodies ChemScore components in the first six terms. The lipophilic term is defined as in ChemScore, while the hydrogen-bonding term is separated into three differently weighted components based on the nature of the hydrogen-bonding partners. The authors claimed that hydrogen-bonds between neutral donors and neutral acceptors contribute more strongly to binding than the when both partners are charged; and that the neutral-charged one lies in-between (Friesner, *et al.*, 2004). GlideScore also modifies the form of the metal interaction term in ChemScore to ensure rewarding those metal–ligand contacts that contribute favorably to binding.

$$\begin{aligned}\Delta G_{bind,GlideScore} = {} & C_{lipo-lipo}\sum f\left(r_{lp}\right) + \\ & C_{hbond-neut-neut}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-neut-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{hbond-charged-charged}\sum g(\Delta r)h(\Delta\alpha) + \\ & C_{max-metal-ion}\sum f(r_{lm}) + \\ & C_{rotb}H_{rotb} + \\ & C_{polar-phob}V_{polar-phob} + \\ & C_{coul}E_{coul} + C_{vdW}E_{vdW} + E_{solv}\end{aligned} \qquad \textbf{I-8}$$

$\boldsymbol{\Delta G_{bind,GlideScore}}$	*GlideScore predicted Gibb's free energy of binding*
$\boldsymbol{C_{lipo-lipo,rotb,\ldots}}$	*Empirical weighting coefficients*
$\boldsymbol{r_{lp}}$	*Sum over all ligand-atom/protein-atom lipophilic pairs according to ChemScore atom classification*
$\boldsymbol{r_{lm}}$	*Sum over all ligand-atom/metal pairs*
$\boldsymbol{f, g, h}$	*Functions that give a full score (1.00) for distances or angles that lie within nominal limits and a partial score (1.00-0.00) for distances or angles that lie outside those limits but inside larger threshold values*
$\boldsymbol{H_{rotb}}$	*Entropic penalty for frozen rotatable bonds (ChemScore)*
$\boldsymbol{V_{polar-phob}}$	*Reward for polar but non-hydrogen-bonding atoms in a hydrophobic region*
$\boldsymbol{E_{coul}}$	*Weighted OPLS-AA ligand–protein Columbic interactions*
$\boldsymbol{E_{vdW}}$	*Weighted OPLS-AA ligand–protein van der Waals interactions*
$\boldsymbol{E_{solv}}$	*GlideScore desolvation penalty*

The seventh term ($C_{polar-phob}V_{polar-phob}$) is calculated by the Schrödinger's active-site mapping facility to reward instances in which a polar but non-hydrogen-bonding atom (as classified by ChemScore) is found in a hydrophobic region (Friesner, *et al.*, 2004; Glide, 2011). Moreover, GlideScore encompasses appropriately weighted gas-phase non-bonded interactions (Columbic and van der Waals) derived from the OPLS-AA force field

(Jorgensen, *et al.*, 1996). Finally, Glide computes the desolvation penalty term (E_{solv}) using a robust solvation model, where waters are docked explicitly to competitive ligand poses. Subsequently, these waters are evaluated by a specialized scoring function, which measures exposure of various groups to the explicit waters. The use of explicit waters has obvious advantages over implicit solvation models, but the required sampling is time-consuming.

Glide employs three related classes of scoring functions to achieve different goals. Selecting the correct docking poses is done via the *Emodel* score, which incorporates GlideScore, the ligand–receptor molecular mechanics interaction energy, and the ligand strain energy. Additionally, Glide employs two forms of GlideScore; namely, Glide SP (for Standard Precision) and Glide XP (for eXtra Precision). Although both *Glide SP* and *Glide XP* use similar terms, they are optimized to serve two different roles in Glide. Glide SP is softer and more forgiving; hence it finds ligands that could potentially bind to the target, even if their poses have slight imperfections. Glide SP's tendency to minimize false negatives and fast performance makes it more suitable for use in database screening. In contrast, Glide XP is a harder function that employs more terms, enhanced sampling, and exacts severe penalties for poses that violate established physical principles. Glide XP, thus, is more efficient at minimizing false positives and is more suitable for lead optimization or for a higher quality prediction on smaller number of ligands.

The goal of the XP Glide methodology is to semi-quantitatively rank the ability of candidate ligands to bind to a specified conformation of the protein receptor (Friesner, *et al.*, 2006). In comparison to Glide SP, the XP Glide scoring is characterized by two key features, (1) it applies a large desolvation penalties to both ligand and protein polar and charged groups in appropriate cases and (2) it has specialized routines to identify specific structural motifs that provide exceptionally large contributions to enhanced binding affinity. Glide XP special terms are derived from theoretical physical chemistry of ligand–protein interactions and tuned to match observations from a wide range of ligand–protein test cases. The beneficial contributions of these terms are not fully accounted for by the generic terms frequently employed scoring function, including those of GlideScore. For example, the standard score assigned by ChemScore lipophilic protein/ligand atoms pair function underestimates the favorable contribution of lipophilic ligand groups *enclosed* in a tight binding pocket. In such cases, a special *hydrophobic enclosure* term is necessary to confer a higher contribution to this motif.

Special terms included in Glide XP and the physicochemical principles and experimental results that led to their inclusion in the scoring function are described in the original article (Friesner, *et al.*, 2006). *Glide XP* is tuned for prediction of binding affinities, and hence it was *chosen for the purpose of this study*. In contrast to the Glide SP or Emodel scoring functions, Glide XP has numerous specialized reward and penalty terms and covers a wider range of ligand–protein interaction motifs. Table I-4 lists the complex descriptors calculated by Glide employed in our study. This includes the total GlideScore and its components, which give a more detailed picture of the features of ligand–protein interactions. Moreover, Glide supports two modes for scoring ligand–protein complexes: (1) the *in place* mode, where the input ligand coordinates are used directly for scoring, and (2) the *refine input* mode, where the input ligand coordinates are optimized in the field of the receptor prior to scoring (Glide, 2011). Both modes were employed in our study; the *in place* mode to assess Glide XP performance on unperturbed experimentally determined structures and the *refine input* mode to assess its sensitivity to relatively imperfect geometries.

Table I-4: Complex descriptors from the Glide XP scoring function. The X denotes the Glide docking mode employed to calculate the respective values, either 'inplace' or 'refineinput'.

Descriptor	Meaning
Glide_X-score	Total Glide XP score, sum of GlideScore components, desolvation penalty, and special penalty and reward terms.
Glide_X-ecoul	Non-bonded Coloumb energy from OPLS-AA force field. Calculated with 50% reduced net ionic charges on groups with formal charges, such as metals, carboxylates and guanidiniums.
Glide_X-evdw	Non-bonded van der Waals energy from OPLS-AA force field. Reduced for the atoms with formal charges (as in Glide_X-ecoul).
Glide_X-energy	Glide_X-ecoul + Glide_X-evdw
Glide_X-einternal[a]	Internal torsional energy of the ligand.
Glide_X-emodel[a]	Emodel score, used to rank order docking poses.
Glide_X-HBond	Hydrogen-bonding term from ChemScore, divided into three differently weighted components depending on nature of the donor and acceptor: neutral–neutral, neutral–charged, or charged–charged.
Glide_X-nbrot	Number of rotatable bonds in the ligand.
Glide_X-RotPenal	Penalty for freezing rotatable bonds of the ligand, defined in ChemScore (Eldridge, *et al.*, 1997) by the function: $H_{rot} = 1 + (1 - 1/N_{rot})\sum_r [P_{nl}(r) + P'_{nl}(r)]/2$ where N_{rot} is the number of frozen rotatable bonds, the summation is over frozen rotatable bonds and $P_{nl}(r)$ and $P'_{nl}(r)$ are the percentages of non-lipophilic heavy atoms on either side of the rotatable bond.
Glide_X-formalcharge	Net charge on the ligand.
Glide_X-nsalt	Number of ionized acidic or basic groups in the ligand.
Glide_X-LowMW	Reward for the ligands with low molecular weight
Glide_X-LipophilicEvdW	Lipophilic term from the hydrophobic grid potential and fraction of the total ligand–protein van der Waals energy
Glide_X-PhobEn	Reward for hydrophobic enclosure.
Glide_X-PhobEnHB	Reward for hydrophobically packed hydrogen-bond.
Glide_X-PhobEnPairHB	Reward for hydrophobically packed correlated hydrogen-bonds.
Glide_X-Electro	Electrostatic rewards; includes Coloumb and metal terms from ChemScore.
Glide_X-Sitemap	Reward polar but non-hydrogen-bonding atom found in a hydrophobic region.
Glide_X-EposPenal	Penalty for solvent exposed ligand groups; cancels van der Waals terms.
Glide_X-Penalties	Polar atom burial and desolvation penalties and penalty for intra-ligand contacts.
Glide_X-HBPenal[b]	Penalty for ligands with large hydrophobic contacts and low H-bond scores
Glide_X-ClBr[b]	Reward for Cl or Br in a hydrophobic environment that pack against Asp or Glu
Glide_X-PiStack[b]	Reward for π-π stacking
Glide_X-PiCat	Reward for π-cation interactions

[a] Not applicable in case of 'in place' scoring, since in this mode neither conformational sampling nor pose generation are performed.

[b] Special structural motifs pertaining to these rewards/penalties were not encountered in any of the complexes in our data set

I.3.4.4 Entropic penalty

Change in entropy upon ligand–protein association is probably the most elusive component of the binding free energy. Thermodynamically robust methods for calculating entropy, e.g. mining minima (Head *et al.*, 1997), typically require efficient sampling of the conformational space e.g. by molecular dynamics. Nonetheless, the high computational cost renders such approaches impractical for database screening or lead optimization purposes. On the other hand, scoring functions usually employ an empirical form to estimate entropic penalties of ligand binding. These empirical formulations vary according to the underlying physical model and degree of approximation, and hence in computational cost (Chang *et al.*, 2007). At the most basic level only the restriction of internal conformational degrees of rotatable bonds is considered. Consequently, a constant penalty is assigned for each freely rotatable bond in the ligand, ranging in value from 0.4 to 1.0 kcal/mol (Galzitskaya *et al.*, 2000; Gilson and Zhou, 2007; Hossain and Schneider, 1999; Page, 1973; Page, 1977a; Page, 1977b; Page and Jencks, 1971; Pickett and Sternberg, 1993; Raha and Merz, 2005; Searle and Williams, 1992). However, this approximation is valid only under three basic assumptions: (1) loss in translational and rotational degrees of freedom (DoF's) of the ligand (and protein) are of little or no consequence on entropy, (2) the bound ligand retains no residual mobility, and (3) there is no change in vibrational DoF upon binding. We implemented three more treatments for ligand entropy addressing one or more of these assumptions; below is a brief discussion thereof.

The first assumption could be valid in a congeneric series of similarly-sized ligands, where the losses in translational and rotational DoF's upon binding are of comparable magnitudes. However, this is not always the case in structure-based design. Consequently, an empirical treatment of translational and rotational DoF's is highly desirable. For instance, the carbohydrate-specific free-energy function proposed by Hill and Reilly (Hill and Reilly, 2008) estimates ΔS_{bind} according to the following formula:

$$\Delta S_{bind} = -k\left[ln(6+\xi N_{tors}) - \frac{\xi N_{tors}}{6+\xi N_{tors}} ln\xi\right]$$

where k is the Boltzmann constant, N_{tors} is the number of torsional DoF (i.e. rotatable bonds), the 6 denotes the loss of three translational and three rotational DoF's, and ξ is an empirical coupling constant introduced by the authors to link the movement in transformational DoF to that in torsional DoF. We implemented the four values of ξ investigated by the authors (0.1, 0.33, 0.67 and 1.0) in our study. We also included the entropic penalty term employed in Glide scoring function, which accounts for the residual ligand mobility by applying the penalty only to bonds expected to be frozen in the bound conformation (Eldridge, *et al.*, 1997). Finally, we used the rigid-rotor harmonic oscillator approximation to estimate the changes in vibrational, rotational, and translational components of ligand's entropy upon binding (MacroModel, 2011).

Table I-5: Complex descriptors estimating the entropic penalty associated with the degrees of freedom of the ligand upon binding.

Descriptor	Meaning
lig_rot_bonds	Number of freely rotatable bonds in the ligand
Hill_dS_0_10	Entropic penalty for freezing ligand's movement (translational, rotational, conformational) according to Hill and Reilly (Hill and Reilly, 2008), ξ=0.10
Hill_dS_0_33	*As above*, ξ=0.33
Hill_dS_0_67	*As above*, ξ=0.67
Hill_dS_1_00	*As above*, ξ=1.00
Glide_X-RotPenal	GlideXP's penalty for freezing rotatable bonds of the ligand, defined in ChemScore (Eldridge, *et al.*, 1997) by the function: $H_{rot} = 1 + (1 - 1/N_{rot})\sum_r [P_{nl}(r) + P'_{nl}(r)]/2$ where N_{rot} is the number of frozen rotatable bonds, the summation is over frozen rotatable bonds and $P_{nl}(r)$ and $P'_{nl}(r)$ are the percentages of non-lipophilic heavy atoms on either side of the rotatable bond.
rrho-TdS_rot	Entropic penalty for the *rotational* degrees of freedom, calculated using the rigid-rotor harmonic oscillator approximation
rrho-TdS_trans	Entropic penalty for the *translational* degrees of freedom, calculated using the rigid-rotor harmonic oscillator approximation
rrho-TdS_vib	Entropic penalty for the *vibrational* degrees of freedom, calculated using the rigid-rotor harmonic oscillator approximation
rrho-TdS_total	Sum of RRHO entropic penalty components

I.3.4.5 Solvent-accessible surface area descriptors

Changes in the solvent-accessible surface areas (SASA) of two binding partners could contribute to the association free energy in a number of ways. Most notably, burial of hydrophobic surfaces is associated with a gain in free energy. This effect, dubbed the *hydrophobic effect*, is a major driving force for one of the most important biological events, protein folding (Chothia, 1974; Chothia and Janin, 1975). Several studies employed the buried (hydrophobic) surface area to compute free energies pertinent to the hydrophobic effect and solvation free energies in general (Eisenberg and McLachlan, 1986; Ooi *et al.*, 1987; Pace, 1992; Wang *et al.*, 2001b). On the other hand, burial of a polar/hydrophilic surface upon binding is associated with an enthalpic cost. Removal of a polar/hydrophilic surface from bulk solvent into the ligand–protein interface involves breaking of established hydrogen-bonds from the protein and the ligand sides alike. In rational design this cost is typically compensated by functionalizing the ligand, so that it can reestablish the same, or even better, H-bond networks upon binding.

In this study, SASA descriptors for the studied complexes were computed to measure the relevant SASA changes upon ligand–protein association. The solvent-accessible surface is defined as the surface traced by the center of spherical probe rolled over the analyzed molecule such that it touches but does not overlap with the van der Waals radii of atoms in that molecule. In our study, SASA components were calculated using a water-sized spherical probe (radius1.4 Å) scanning the surface of the analyzed molecule(s) at 0.1 Å spaced grid points. The NeoScore platform, however, supports adjustment of probe radius and scanning resolution in SASA computations. We took advantage of this option in developing a classification scheme for binding site geometries (see Topological classification of binding sites section below). SASA descriptors employed in our study are listed in Table I-6 and schematically illustrated in Figure I-1.

Table I-6: Solvent-accessible surface area (SASA) descriptors, computed using 1.4 Å probe and a grid spacing (resolution) of 0.1 Å. An atom is classified as polar if the absolute value of its partial charge is more than 0.25e.

Descriptor	**Meaning**
sasa-complex_total	$SASA_{all\ atoms}^{complex}$
sasa-complex_non_polar	$SASA_{non-polar\ atoms}^{complex}$
sasa-complex_polar	$SASA_{polar\ atoms}^{complex}$
sasa-complex_receptor	$SASA_{receptor\ atoms}^{in\ complex}$
sasa-complex_receptor_non_polar	$SASA_{non-polar\ receptor\ atoms}^{in\ complex}$
sasa-complex_receptor_polar	$SASA_{polar\ receptor\ atoms}^{in\ complex}$
sasa-receptor_free	$SASA_{receptor\ atoms}^{free}$
sasa-receptor_free_non_polar	$SASA_{non-polar\ receptor\ atoms}^{free}$
sasa-receptor_free_polar	$SASA_{polar\ receptor\ atoms}^{free}$
sasa-receptor_buried_total	$SASA_{receptor\ atoms}^{free} - SASA_{receptor\ atoms}^{in\ complex}$
sasa-receptor_buried_non_polar	$SASA_{non-polar\ receptor\ atoms}^{free} - SASA_{non-polar\ receptor\ atoms}^{in\ complex}$
sasa-receptor_buried_polar	$SASA_{polar\ receptor\ atoms}^{free} - SASA_{polar\ receptor\ atoms}^{in\ complex}$
sasa-complex_ligand	$SASA_{ligand\ atoms}^{in\ complex}$
sasa-complex_ligand_non_polar	$SASA_{non-polar\ ligand\ atoms}^{in\ complex}$
sasa-complex_ligand_polar	$SASA_{polar\ ligand\ atoms}^{in\ complex}$
sasa-ligand_free	$SASA_{ligand\ atoms}^{free}$
sasa-ligand_free_non_polar	$SASA_{non-polar\ ligand\ atoms}^{free}$
sasa-ligand_free_polar	$SASA_{polar\ ligand\ atoms}^{free}$
sasa-ligand_buried_total	$SASA_{ligand\ atoms}^{free} - SASA_{ligand\ atoms}^{in\ complex}$
sasa-ligand_buried_non_polar	$SASA_{non-polar\ ligand\ atoms}^{free} - SASA_{non-polar\ ligand\ atoms}^{in\ complex}$
sasa-ligand_buried_polar	$SASA_{polar\ ligand\ atoms}^{free} - SASA_{polar\ ligand\ atoms}^{in\ complex}$

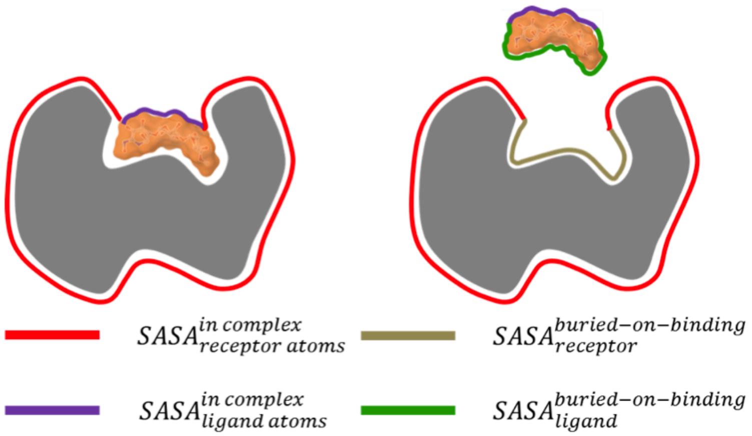

$SASA^{in\ complex}_{receptor\ atoms}$
$SASA^{buried-on-binding}_{receptor}$
$SASA^{in\ complex}_{ligand\ atoms}$
$SASA^{buried-on-binding}_{ligand}$

Figure I-13: Schematic representation of SASA components employed in our study for binding affinity prediction.

I.3.4.6 Ligand descriptors

A number of ligand-derived descriptors were included to represent potentially relevant structural and energetic features, in our descriptor pool (Table I-7). Below are a brief description of them and an explanation of their relevance to scoring ligand–protein interactions. Two basic measures of molecular size were considered: molecular weight and number of heavy atoms of the ligand. Their significance was first underlined in the classical work by Kuntz *et al.* who demonstrated that for strong binders the free energy of binding increases by ~ −1.5 kcal/mol for each non-hydrogen atom up to a limit of 15, where it reaches a plateau (Kuntz *et al.*, 1999). Employing additive terms in scoring functions is known to have a biasing effect; larger ligands receive higher scores (Balius *et al.*, 2011; Ferrara, *et al.*, 2004). It might, therefore, be necessary to compensate for this bias by penalizing large ligands and/or rewarding relatively smaller ligands (Friesner, *et al.*, 2006).

We also included descriptors to account ligand internal strain, which is defined as the energetic cost paid for forcing the relaxed unbound conformation of the ligand to assume the bioactive conformation. The relaxed conformation could be taken to be the *nearest local minimum* found in by typical energy minimization or to the *global minimum* (Perola and Charifson, 2004). The global minima for the studied carbohydrate ligands, were obtained through an exhaustive conformational search using MacroModel (MacroModel, 2011), setting the maximum number of generated conformers to 5000 and a wide energy window (40.0 kcal/mol) for conformer rejection. In addition, the SM8 quantum mechanical aqueous continuum solvation model (Marenich *et al.*, 2007) was employed to estimate ligands' desolvation penalties. The computation was carried out on the crystallographic ligand conformation using B3LYP density functional and the 6-31G** basis set in Jaguar (Jaguar, 2011). We also employed SM8 solvation free energy weighted according to the ligand's buried surface area to account for partial ligand desolvation, particularly for ligands bound close to the surface.

Table I-7: Complex descriptors measuring important structural and energetic features of the ligand molecule.

Descriptor	**Meaning**
lig_mol_wt	Molecular weight of the ligand
lig_heavy_atoms	Number of non-hydrogen atoms in the ligand
lig_rot_bonds	Number of freely rotatable bonds in the ligand, used as estimate for the conformational entropic cost of binding
lig_local_strain	Difference in force field potential energy (OPLS-AA) between the bound conformation of the ligand and its nearest local energy minimum
lig_global_strain	Difference in force field potential energy (OPLS-AA) between the bound conformation of the ligand and its global minimum obtained from an exhaustive conformational search
desolvation_qm_sm8	Solvation free energy from quantum mechanical SM8 model (Marenich, *et al.*, 2007)
fraction_ligand_exposed	$SASA_{ligand}^{in\ complex}/SASA_{ligand}^{free}$
wt_desolv_qm_sm8_buried	*(1– fraction_ligand_exposed)·desolvation_qm_sm8*
hbonds-to_receptor	Number of hydrogen bonds between ligand and protein in the energy-minimized crystal structure
hbonds-to_water	Number of hydrogen bonds between ligand and binding site waters resolved in the crystal structure, zero if the crystal structure had no waters
hbonds-total	*hbonds-to_receptor + hbonds-to_water*

I.3.4.7 Dynamic Properties

The use of static representation of ligand–protein complexes to calculate bulk properties such as binding free energies is a widely used approximation, primarily due to computational efficiency. In reality, however, thermodynamic observables arise from of an ensemble of microstates of the system, not a single configuration. Rigorous methods for prediction of binding affinity, such as free energy perturbation (Jorgensen and Thomas, 2008; Zwanzig, 1954) and thermodynamic integration (Rodinger, *et al.*, 2005), require converged and sufficient sampling of the system's configurational space. Such configuration ensembles are typically generated from physics-based simulations of the system such as molecular dynamics (Deng and Roux, 2009) or Monte Carlo simulations (Price and Jorgensen, 2001). Alternatively, frames from molecular dynamics trajectories are used as input for end point methods such as MM/GBSA or MM/PBSA (Hou, *et al.*, 2011a). Evidently, adequate conformational sampling is particularly important in case of carbohydrate–protein complexes due to their natural high mobility (Bradbrook *et al.*, 2000). Moreover, it has been reported that water molecules bridging ligand–protein interactions in relatively open binding pockets are being exchanged constantly (Caffarena, *et al.*, 2002; Tempel, *et al.*, 2002), which makes modeling them explicitly imperative.

To account for the dynamic nature of molecular interactions in our carbohydrate–protein data set, all ligand–protein complexes were subjected to 5.0 ns long molecular-dynamics simulations. Desmond software package from D. E. Shaw's group (Bowers *et al.*, 2006) was used to perform these simulations. For the purpose of this study, we employed the OPLS_2005 force field as implemented in the Schrödinger 2011 Suite (Damm, *et al.*, 1997; Jorgensen, *et al.*, 1996; Jorgensen and Tirado-Rives, 1988; Kaminski, *et al.*, 2001; Kony, *et al.*, 2002; Maestro, 2011). Each system was solvated using an orthorhombic, TIP3P water box (Jorgensen, *et al.*, 1983) extending at least 10 Å away from the complex. Sodium and chloride ions (0.15 M) were added to neutralize the charges and to approximate physiological conditions. Long-range electrostatic interactions were handled using the Particle mesh Ewald summation (Darden *et al.*, 1993). All systems were equilibrated using the default relaxation protocol (Desmond, 2011) and simulated over the span 5.0 ns with a time step of 2.0 fs. The SHAKE algorithm (Ryckaert *et al.*, 1977) was applied to all heavy-atom bound hydrogens. Production runs were carried out in the Martyna-Tobias-Klein (Martyna *et al.*, 1994) NPT ensemble (constant pressure, temperature, and total number of particles) and employing the Nose-Hoover barostat (Nosé, 1984) to maintain a constant temperature of 300 K. Molecular energies and trajectories (atomic coordinates and velocities) were recorded in 5.0 ps intervals resulting in a total of 1000 frames per simulation. When the simulation was finished, we extracted 25 equally spaced frames (200 ps apart) from the output trajectory. These frames were used to compute the time-dependent values for non-bonded force field interactions (OPLS_2005 and MMFFs) and MM/GBSA free energy functions following the same procedures used to compute the static descriptors. Finally, dynamic averages of these descriptors were added to the complex descriptors pool.

I.3.5 Topological classification of binding sites

Binding sites differ in the shape and degree of exposure to the bulk solvent; some are shallow depressions on the protein surface while others are completely buried pockets. *(Note: We will use the terms 'receptor' and 'protein' interchangeably in the following discussion to refer to the macromolecular target to which a small molecule ligand binds, which includes enzymes, membrane-bound receptors, nuclear receptors, etc.)*. In the course of our investigation, we observed interesting trends in the relationship between binding affinity and certain topological features of binding sites. The plot in Figure I-14 is an illustrative example of the clues we encountered for the existence of such trends. A simple relationship between the receptor SASA that becomes buried upon ligand binding and the exposed portion of ligand's SASA in the complex roughly separates ligand–protein complexes into clusters with different relative distributions of high/low binding affinities. This relationship (along with several comparable relationships) indicated that SASA components could be, somehow, useful in our quest for a reliable free-energy function.

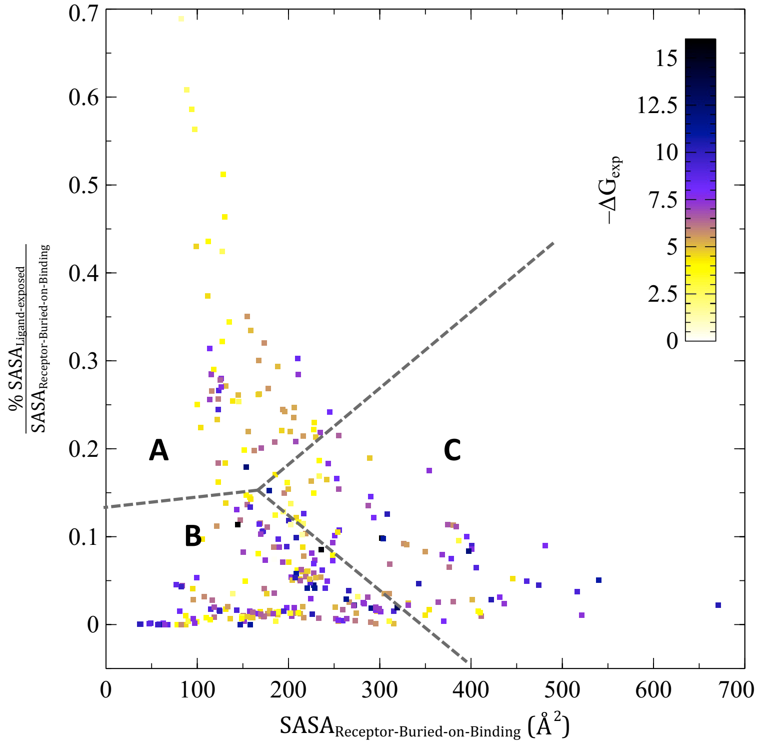


Figure I-14: The ratio of ligand exposed SASA (%) to receptor SASA buried on binding (vertical axis) versus the receptor SASA buried on binding on the (horizontal axis). Points are colored according to experimental ΔG of the corresponding complex (color bar). Zone A is enriched in complexes with low to moderate affinity, while zone B is enriched in moderate to high-affinity complexes. Zone C doesn't show any specific affinity-based enrichment.

Molecular surface area components are widely used in modeling molecular properties such as hydrophobic interactions (Chothia, 1974; Chothia and Janin, 1975) and solvation effects (Wang, *et al.*, 2001b). Despite the fairly predictable nature of the relationship between SASA components and these properties, no such relationship exists between SASA and binding affinities. In our data set, experimental binding affinities were not correlated to SASA buried-on-binding of the ligand (r=0.32) nor the receptor (r=0.17). This lack of correlation precludes the possibility of directly employing SASA components as terms in free energy function; and consequently calls for a more thorough investigation to understand the nature of the trends described in the previous paragraph. The investigation was continued under the assumption that SASA components could –in a useful way– categorize the studied carbohydrate–protein complexes based on binding site shape and degree of solvent exposure.

I.3.5.1 Preliminary investigation

We started the investigation by using a rather simple descriptor to measure the degree of solvent exposure of the binding site; namely the percent ligand exposed (PLE):

$$PLE = \left(\frac{SASA_{ligand}^{in\ complex}}{SASA_{ligand}^{free}}\right).100(\%)$$

Complexes with higher PLE values are expected to have a binding site fairly close to the protein surface allowing for a larger portion of the bound ligand to be solvent exposed. We then examined the use of various cut-offs to distinguish *surface-accessible pockets* (high PLE) from *buried inaccessible cavities* (low PLE) (Figure I-15). It became readily apparent that this approach is inadequate, as complexes with PLE as low as 5% have fairly surface-accessible binding sites (Figure I-16). Moreover, the use of ligand-based property only obscures relevant protein features, such the size of the binding site.

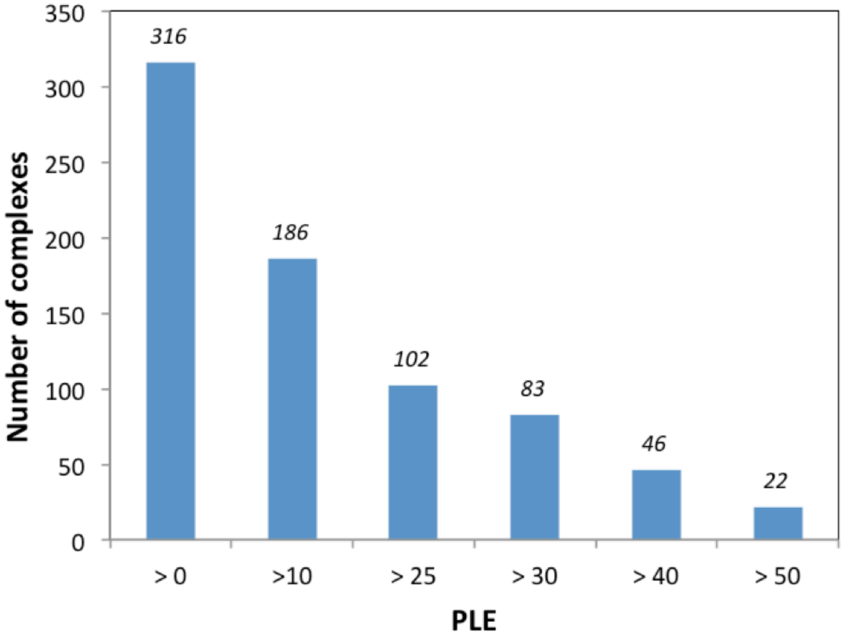
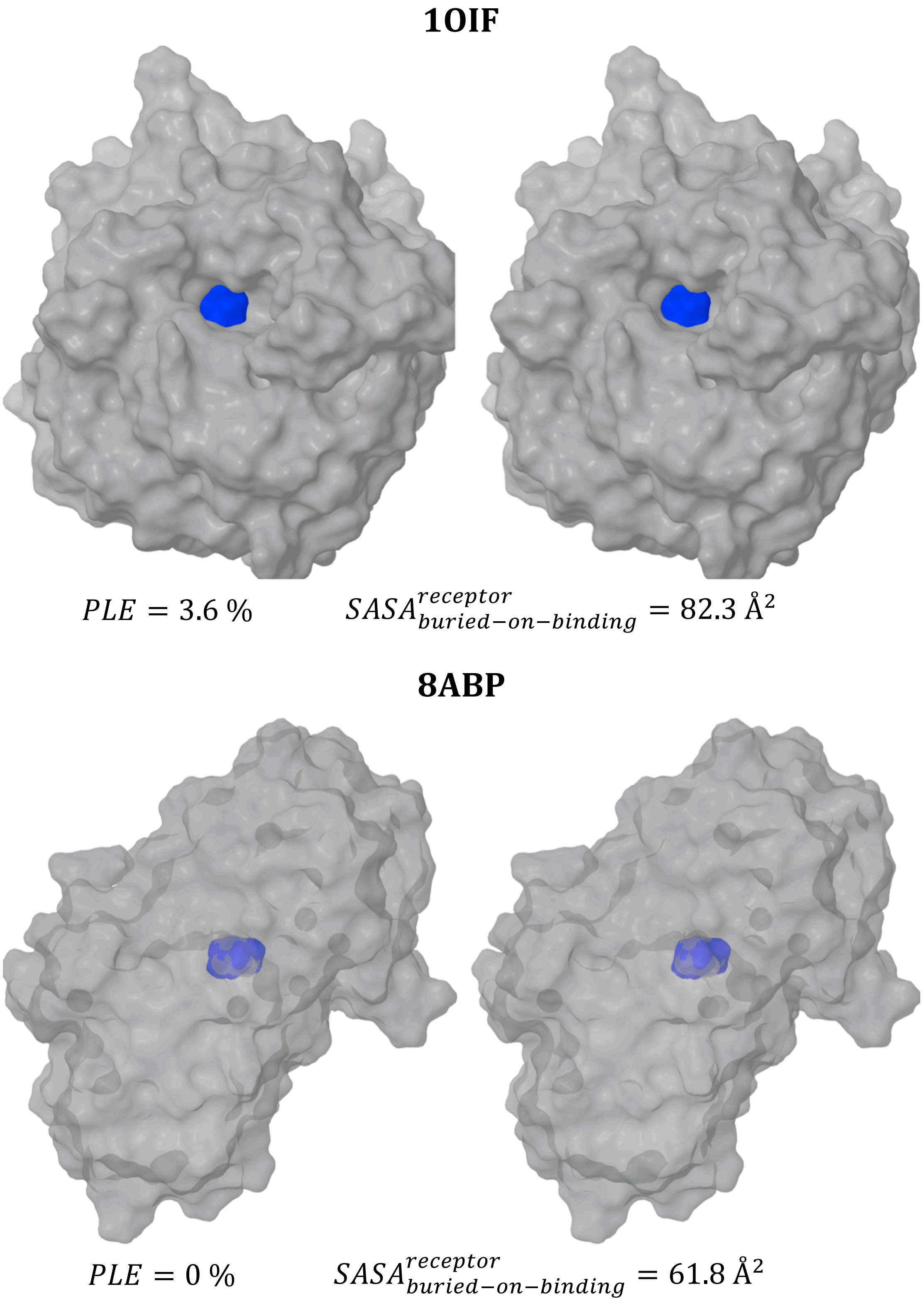


Figure I-15: Number of proteins with a surface-accessible binding site using different cutoffs for the percentage of ligand SASA exposed in complex.



1OIF

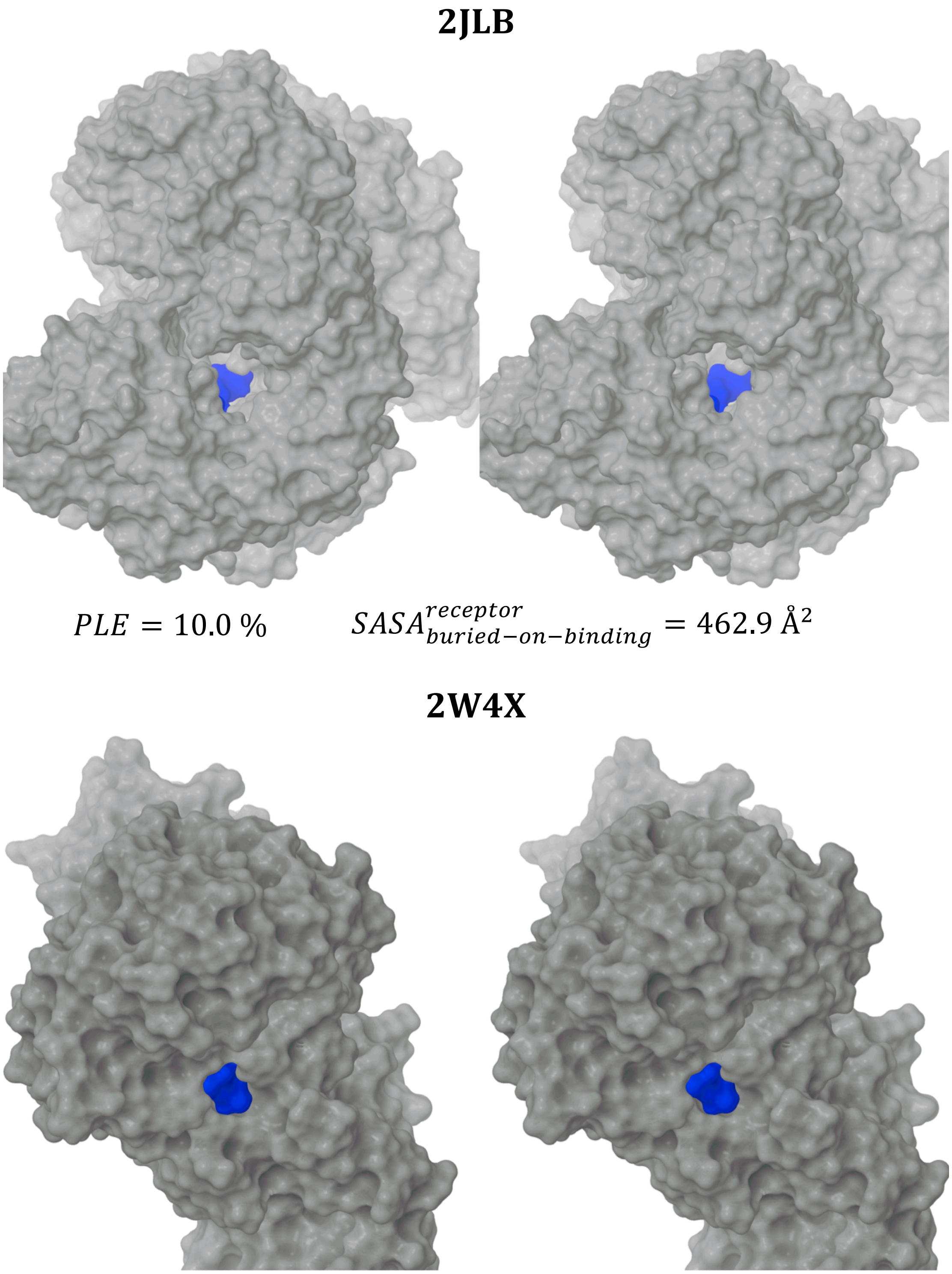
$PLE = 3.6\ \%$ $SASA^{receptor}_{buried-on-binding} = 82.3\ \text{Å}^2$

8ABP

$PLE = 0\ \%$ $SASA^{receptor}_{buried-on-binding} = 61.8\ \text{Å}^2$

Figure I-16: Stereo view of the solvent-accessible surface representation of 1OIF (top) and 8ABP (bottom), grey: protein, blue: ligand. Although a small fraction of the ligand SASA in 1OIF is exposed in complex, the binding site of the protein is fairly surface-accessible in comparison to 8ABP. Protein surface for 8ABP is rendered transparent to show the buried ligand.

The receptor SASA buried upon ligand binding ($SASA_{receptor}^{buried-on-binding}$) could serve as a direct measure of the size and solvent-accessibility of the binding site. Using this value alone, however, is not sufficient to decide whether the binding site is surface-accessible or buried as can be seen from comparison in Figure I-17. Although the 2JLB complex has a relatively big binding surface (462.0 Å^2), the entrance of the binding site is comparatively smaller and, thus, only a small portion of the ligand surface (10%) is solvent accessible in the complex. On the other hand, the 2W4X complex has a significantly smaller binding surface (218.0 Å^2) despite the obvious similarity shared between its binding-site topology and that of the 2JLB complex (both are accessible via a relatively small opening). Therefore, it became obvious that the use of simple ligand-based and/or protein-based properties is not sufficient for geometrical classification of protein binding sites. In addition, visual inspection of the gross topological features of ligand–protein complexes confirmed that it is not accurate to think of binding sites according to the surface-accessible vs. buried dichotomy; and it is necessary to introduce more categories to account for the in-between cases. With the failure of simple approaches, we turned to literature searching for more rigorous approaches for deriving topological classifications of binding sites.



2JLB

$PLE = 10.0\ \%$ $\qquad SASA^{receptor}_{buried-on-binding} = 462.9\ \text{Å}^2$

2W4X

$PLE = 6.6\ \%$ $\qquad SASA^{receptor}_{buried-on-binding} = 218.0\ \text{Å}^2$

Figure I-17: Stereo view of the solvent-accessible surface representation of 2JLB (top) and 2W4X (bottom), grey: protein, blue: ligand. There is a substantial difference between the two complexes in the surface area buried-on-binding (size of the binding pocket), although the two binding sites are comparable in shape and degree of surface exposure.

I.3.5.2 Reported methods for binding site characterization

A number of methods for characterization of binding sites are described in literature. Early approaches focused on the morphological classification of antibody combining sites. Rees *et al.* identified –by visual inspection– three classes of antibody-combining-site topology: cavity, groove, and planar (Webster *et al.*, 1994; Webster and Rees, 1995). Thornton *et al.* employed a more quantitative approach to classify antibody-binding sites using Kuhn's fractal atomic density measurement as an estimate of the surface curvature (MacCallum *et al.*, 1996). Based on the relative degree of surface concavity and convexity antibodies structures were sorted into four categories: concave, moderately concave, ridged, and plane. Lee *et al.* employed a larger data set of antigen–antibody complexes to develop a heuristic classification algorithm (Lee *et al.*, 2006). Antibodies were allocated to one of five topographic classes based on computerized analysis of certain geometric characteristics of the binding site. The topographical classes were named after familiar geological features: cave, crater, canyon, valley, and plain.

Furthermore, several tools are available analysis of binding pockets and cavities that are not limited to surface binding pockets. CASTp server, for instance, is an online tool for identification and measurement of surface accessible pockets and interior cavities (Dundas *et al.*, 2006). It employs an analytical approach to calculate the area and volume of each pocket and cavity, and reports relevant features of mouth openings, e.g. number, area, and circumference of mouth lips. The current version of CASTp, however, does not encompass shallow depressions. The creators of BindingMOAD online database (Benson, *et al.*, 2008) developed a tool, GoCAV (Smith *et al.*, 2006), to calculate and display molecular surfaces for the ligand and for protein cavities and incorporated it into their the online portal (http://www.BindingMOAD.org). More programs have been developed to calculate surfaces and cavities, including POCKET (Levitt and Banaszak, 1992), SURFNET (Laskowski, 1995), CAST (Liang *et al.*, 1998), and PASS (Brady and Stouten, 2000).

All of the aforementioned programs have definite advantages, such as the ability to identify and describe pockets without needing bound ligands to locate them. However, they are primarily aimed at the characterization of geometrical features of these pockets, rather than classifying them into defined categories. Moreover, they cannot handle certain binding site configurations properly, e.g. antibody-specific approaches are tuned for surface-exposed binding sites and not applicable for buried binding sites while CASTp excludes shallow depressions from calculation. In all cases, special software is required to perform binding site analysis, which might not be convenient for all users. Therefore, we decided to develop a SASA-based empirical method for binding site characterization and classification that can be easily implemented using commonly used modeling software. This method, the BOB-PR plots, is described it in the following section.

I.3.5.3 The BOB-PR plots

To construct the BOB-PR plots, the receptor SASA that becomes buried upon binding ($SASA_{receptor}^{buried-on-binding}$) was calculated using probes with varying radii, from 1.0 to 10.0 Å. Then, we plot the values of the SASA buried-on-binding (BOB) vs. the corresponding probe radius (PR). Receptor BOB-SASA is calculated as a difference between two receptor-based SASA components:

$$SASA_{receptor}^{buried-on-binding} = SASA_{receptor\ atoms}^{free\ receptor} - SASA_{receptor\ atoms}^{in\ complex}$$

This difference maps the patch of the receptor surface accessible to the solvent in the free receptor, and when the ligand binds the patch is no longer solvent accessible. The basic idea of the BOB-PR plots is that smaller probes can roll into deeper and less accessible binding sites while larger probes cannot. As the probe size increases, its ability to gain access to partially buried pockets is reduced and consequently the difference in the equation above decreases. When the probe is too big to access the interior surface of ligand binding site, it will essentially map the same surface in presence and absence of the ligand; i.e. the receptor exterior surface, and the difference above becomes zero (Figure I-18). On the other hand, a fully solvent exposed binding site, e.g. a shallow surface depression, can be equally sampled by small and large probes (Figure I-19).

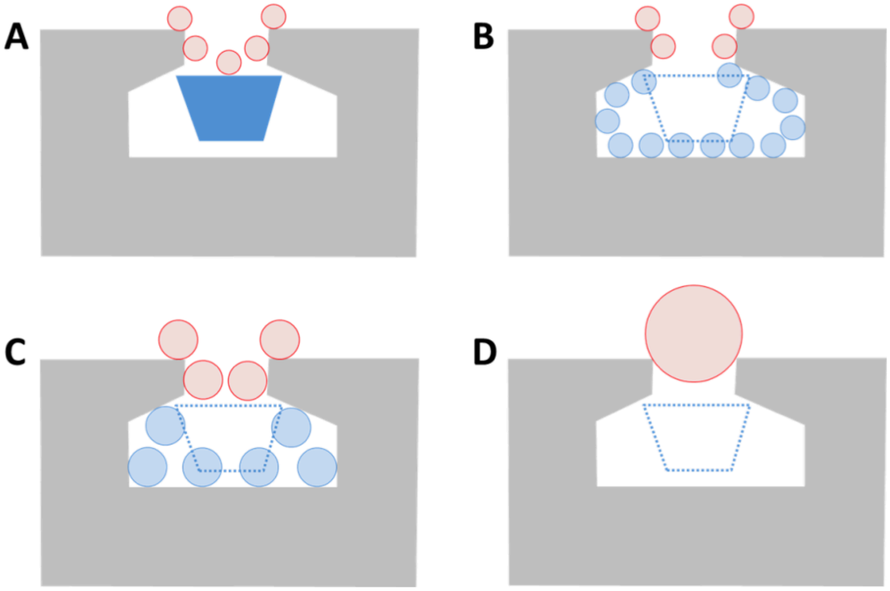


Figure I-18: Schematic representation of the principle of BOB-PR plots. (A) When the ligand is bound the solvent probe (red circles) cannot sample the entire surface of the receptor. (B) When the ligand is not present the probe maps a new surface (blue circles) that was not visible, i.e. buried, when the ligand was bound in A. (C) A medium-sized probe maps a smaller interior surface of the receptor which is buried when the ligand is bound (fewer blue circles). (D) A large probe cannot map the interior receptor surface, i.e. the presence or absence of the ligand has no effect on the calculated receptor SASA. Consequently, the difference between receptor SASA with and without ligand becomes zero.

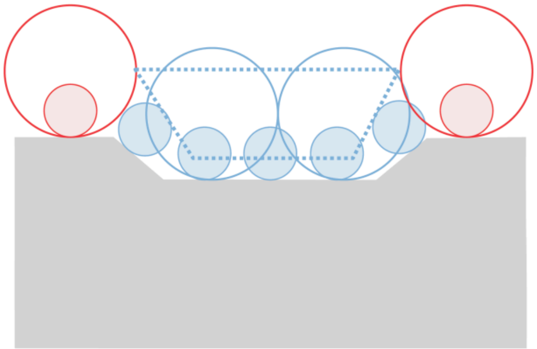


Figure I-19: BOB-PR in case of shallow surface exposed binding site. The receptor SASA buried when the ligand (blue dotted trapezoid) binds can be mapped, virtually equally, by small (filled blue circles) and fairly large (empty blue circles) probes.

We visually inspected BOB-PR plots of a few examples cases and interactively analyzed the molecular surface representations of the corresponding complexes Figure I-20. There are a number of ways to describe BOB-PR curves: a given curve is roughly either pointing up (e.g. 2VMC) or down (e.g. 3HDQ); it might be steep (e.g. 1W9W) or flat (e.g. 1BB8); and it might go down to zero relatively fast (e.g. 1UWU), slowly (e.g. 2JJO), or not at all (e.g. 1RDK). From the many geometrical features that can be derived from BOB-PR plots, two were found to be most useful for developing scheme to classify binding sites: where the curve BOB curve reaches zero and the slope of the longest uniform tail of the curve (Figure I-21). The probe size at which value of receptor BOB vanishes can be obtained directly from the plot if it is less than 10 Å, otherwise it is calculated by extrapolating the curve's tail portion.

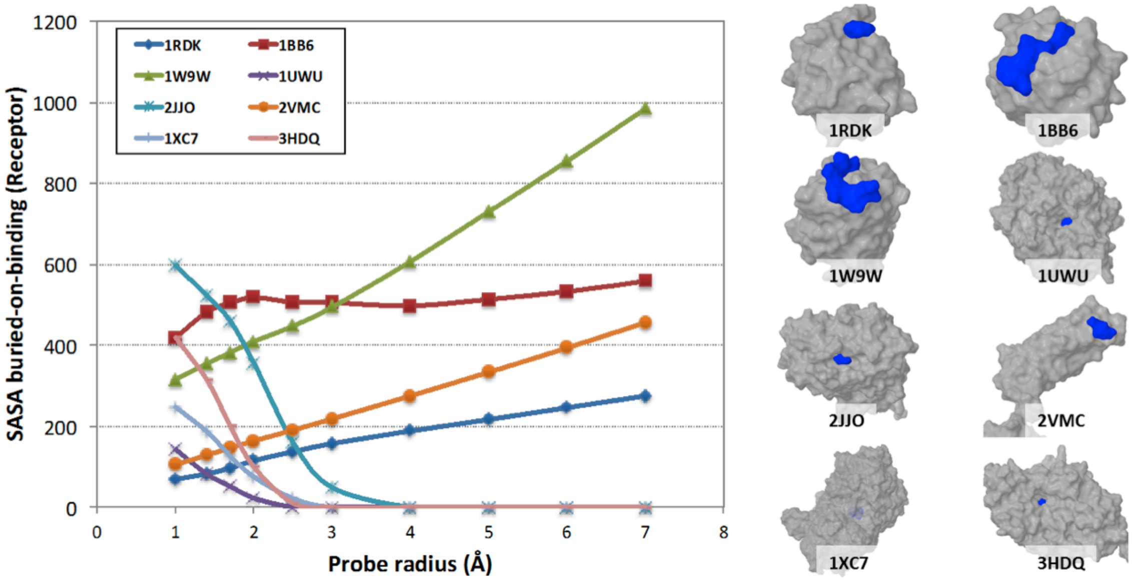


Figure I-20: BOB-PR plots for eight test cases (PDB codes in the legend) and the solvent-accessible surface representations of the corresponding ligand–protein complexes.

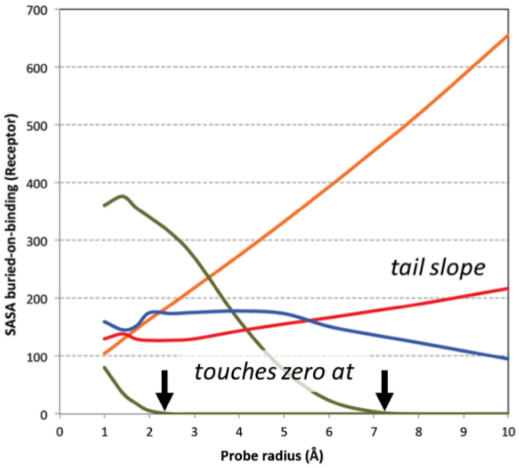


Figure I-21: Two key features of BOB-PR plots employed for categorization of binding sites.

The rules for defining binding site anatomical classes and allocating complexes to them were derived and optimized iteratively. Initially three binding site classes were proposed: shallow, partially buried, and fully buried. We then realized that the partially buried category is too heterogeneous, so we introduced a fourth class, groove, which is halfway between shallow and partially buried. Finally, the groove can be further divided into two classes based on the relative size of groove mouth opening: big mouth and small mouth. The process of defining the classes and setting the boundaries between them was guided by (1) visual inspection of complexes and comparing binding site morphological features to those reported in literature, and (2) assessment of the influence of the proposed set of rules on the performance of several empirical free-energy functions, by fitting the empirical function to the resultant categories separately. Eventually, our data set was sorted into five non-overlapping classes based on geometry and degree of solvent exposure of the binding site: fully buried, partially buried, small mouth groove, big mouth groove, and shallow (Table I-8 and Figure I-22).

Table I-8: Topological classification of binding sites.

Category	Members	touches_zero_at	tail_slope	$SASA_{ligand}^{in\ complex}$
Fully buried	72	≤ 4 Å	n.a.	< 11 Å^2
Partially buried	52	≤ 4 Å	n.a.	> 11 Å^2
Small mouth (groove)	43	< 8 Å	< 0	n.a.
Big mouth (groove)	63	> 8 Å	< 0	n.a.
Shallow	86	< 0 Å	> 0	n.a.

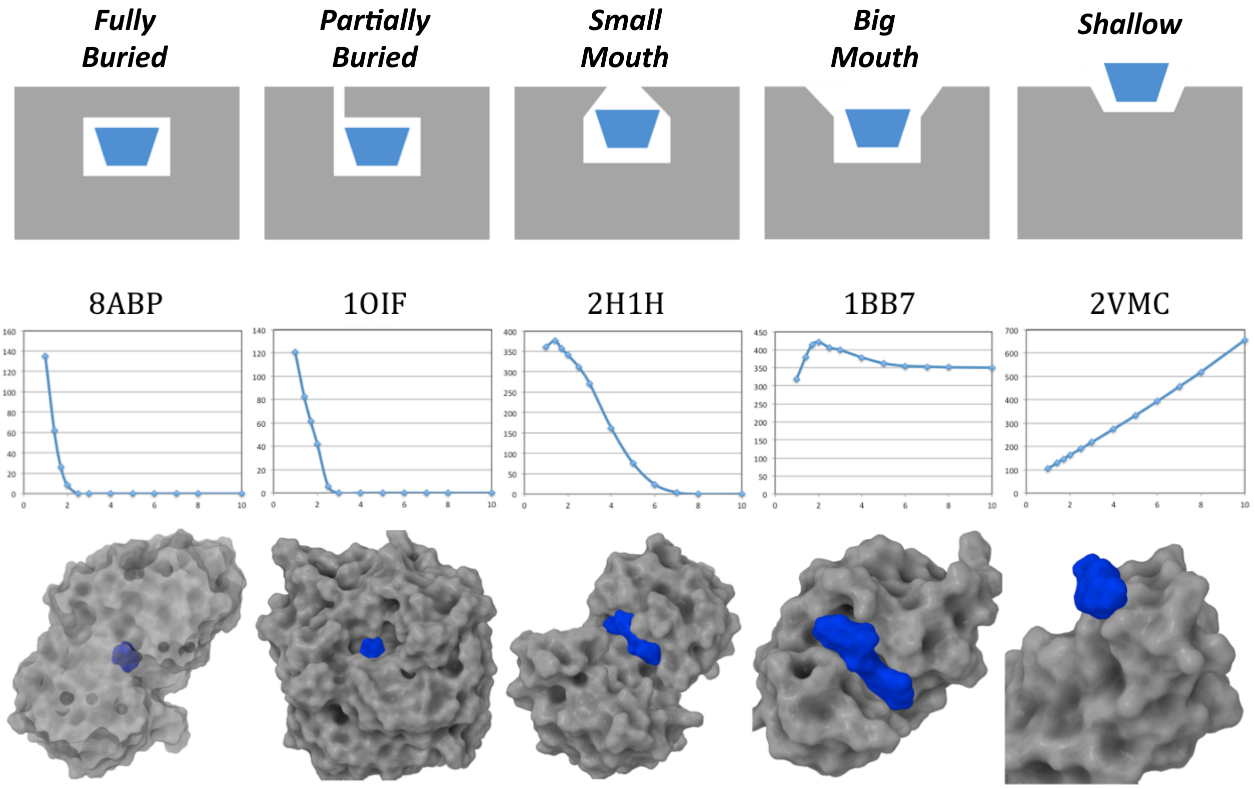


Figure I-22: Complexes were classified into five categories based on topology and solvent exposure of the carbohydrate-binding site. From top to bottom the figure shows: category name; schematic representation of the category; PDB code for an example carbohydrate–protein complex; BOB-PR plot: Receptor SASA buried on binding (Å^2, vertical axis) vs. probe size (Å, horizontal axis) for the example complex; and the solvent-accessible surface representation of the example complex (blue: ligand, grey: protein). In the left-most complex, the protein surface is rendered transparent to show the completely buried ligand.

I.3.6 Statistical validation

Empirical free-energy models investigated in this study are linear combinations of terms each represents a component of the free-energy change associated with binding.

$$\Delta G_{bind} = c_1 \Delta G_1 + c_2 \Delta G_2 + \cdots$$

The experimental binding affinity, ΔG_{bind}, is the dependent (or response) variable (y) while the complex descriptors, ΔG_i's, constitute the independent (or predictor) variables (x's). Standard multiple linear regression was used to derive the weighting coefficients, c_i's, by fitting the linear equation(s) to experimental binding affinities. In the following paragraphs we briefly outline the metrics employed to assess the quality and statistical validity of the developed models.

***r*²**

The coefficient of determination, r^2, is the portion of variance in the observations explained by the model. It is calculated using the formula:

$$r^2 = 1 - \frac{SS_{Err}}{SS_{Tot}} = 1 - \frac{\sum_{i=1}^{N} \left(\Delta G_{i,calc} - \Delta G_{i,exp}\right)^2}{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \langle \Delta G \rangle\right)^2}$$

where SS_{Err} is the sum of squared errors (residuals), SS_{Tot} is the total sum of squares, $\Delta G_{i,exp}$ is the experimental (actual) binding affinity, $\langle \Delta G \rangle$ is the average of binding affinities in the training set and $\Delta G_{i,calc}$ is the binding affinity predicted by the model. A perfect model, one that explains all the variance in the experimental affinities using the proposed equation, has an r^2 of 1.0. The value of r^2 should be interpreted carefully, since it only measures the ability of the model to *reproduce* the binding affinities in the training set, and has no relation to the predictive power of the model. A variant of r^2, the adjusted-r^2, was used as a fitness function to rank models in genetic algorithm searches (cf. page 47). The adjusted-r^2 is more suitable for that purpose because it penalizes overly complex models.

RMSE and MUE

The root mean square error (RMSE) and mean unsigned error are measures of absolute accuracy of estimation, i.e. how accurately can the model predict the experimental binding free energies.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \Delta G_{i,calc}\right)^2}{N}}$$

$$MUE = \frac{\sum_{i=1}^{N} \left|\Delta G_{i,exp} - \Delta G_{i,calc}\right|}{N}$$

q^2

In cross-validation, a number of modified subsets is created by repeatedly removing one or more points (complexes) from the original training set, such that each complex is removed only once. Each modified subset is then used to estimate the coefficients of the equation, and the resultant model is used to predict the binding affinity of the complexes that were left out. The cross-validation r^2, r^2_{cv} or q^2, is calculated according to the equation:

$$q^2 = 1 - \frac{PRESS}{TSS} = 1 - \left(\frac{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \Delta G_{i,calc} \right)^2}{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \langle \Delta G \rangle \right)^2} \right)$$

where *PRESS* is the prediction error sum of squares, and *TSS* is total sum of squares of response variable in the training set. If one data point is removed in each cycle, the process is termed leave-one-out (LOO) cross-validation. In such case, the data structure might be perturbed only slightly, especially if the original data set is large. An alternative approach could be used to perturb the data more significantly, namely the “leave-several-out” or “leave-*k*-out”. It is recommended to choose the number of points to be left out such that *seven* subgroups are created to around 7 (Lavine, 1996). It is important to note that q^2 should be regarded as a measure of internal consistency of the derived model rather than as a true indicator of the predictability (Verma, *et al.*, 2010).

p^2

The predictive correlation coefficient, r^2_{pred} or p^2, measures the ability of the model to forecast the affinities of an external test set of complexes using the model derived from the training set.

$$p^2 = \frac{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \langle \Delta G \rangle \right)^2 - \sum_{i=1}^{N} \left(\Delta G_{i,calc} - \Delta G_{i,exp} \right)^2}{\sum_{i=1}^{N} \left(\Delta G_{i,exp} - \langle \Delta G \rangle \right)^2}$$

In our study, the external test set comprised 106 FimH adhesin inhibitors with known experimental binding affinities. We also compared the predictive quality of the proposed free-energy functions to the recently reported mQSAR model for this data set (Eid, *et al.*, 2013).

I.4. Results and Discussion

I.4.1 Layout

The primary aim of this thesis is to develop and validate a predictive scoring function to calculate the binding affinity of carbohydrate–protein complexes from their three-dimensional structures. Typically, scoring functions and free-energy functions model the binding free energy as a mathematical relationship to the atomic coordinates of the ligand–protein system under investigation.

$$\Delta G_{bind} = f(x, y, z)_{3N-6}$$

The question dealt with in this thesis could, therefore, be viewed as a classical quantitative structure activity relationship (QSAR) problem. In such a theme, atomic configurations of carbohydrate–protein complexes are first transformed into meaningful quantities referred to as *descriptors*. The required scoring function uses a combination of these descriptors to predict binding affinities. The solution space of hypothetical free-energy functions can be schematically represented as shown in Figure I-23. A scoring function, which can be used in structure-based applications, should employ a minimal number of descriptors and predict binding affinities with acceptable accuracy (e.g. within 1 order of magnitude from experiment); and they should do so in a physically meaningful and statistically robust manner. In good QSAR practice, models relying on large number of descriptors to achieve satisfactory performance should be avoided. Such models are potentially over-fitted and usually have no useful predictive power.

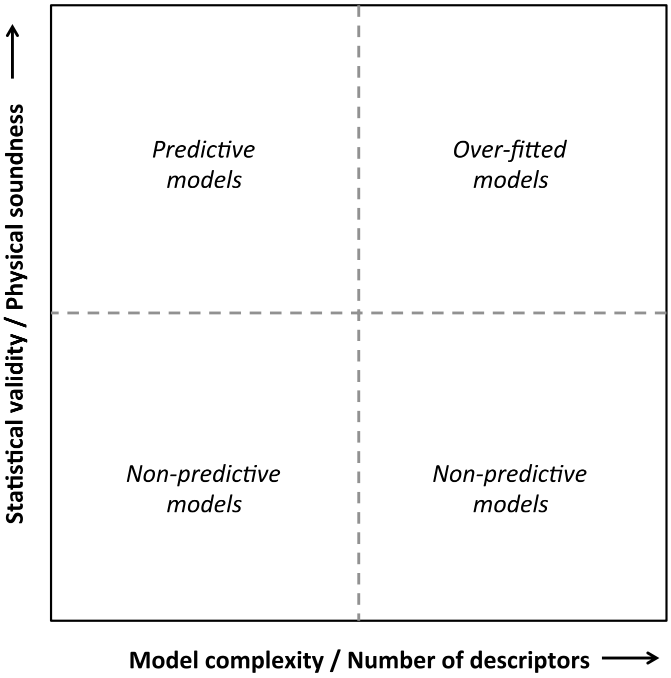

Statistical validity / Physical soundness
Predictive models
Over-fitted models
Non-predictive models
Non-predictive models
Model complexity / Number of descriptors

Figure I-23: Solution space for potential carbohydrate–protein free-energy functions.

To achieve the aforementioned goal, a diverse data set of carbohydrate–protein complexes with known three-dimensional structure and experimental binding affinities was collected and refined. A large number of molecular descriptors were calculated to quantify relevant geometrical and physicochemical features of the assembled complexes. The gathered pool of carbohydrate–protein binding affinities and the corresponding complex descriptors should, in principle, contain the solution to the problem at hand, i.e. the relationship between binding affinity and atomic configuration. The process of uncovering and understanding this relationship went through iterative stages of investigation and analysis (Figure I-24). This section discusses the fundamental issue(s) dealt with at each stage and the corresponding outcomes in the same arrangement presented in Figure I-24.

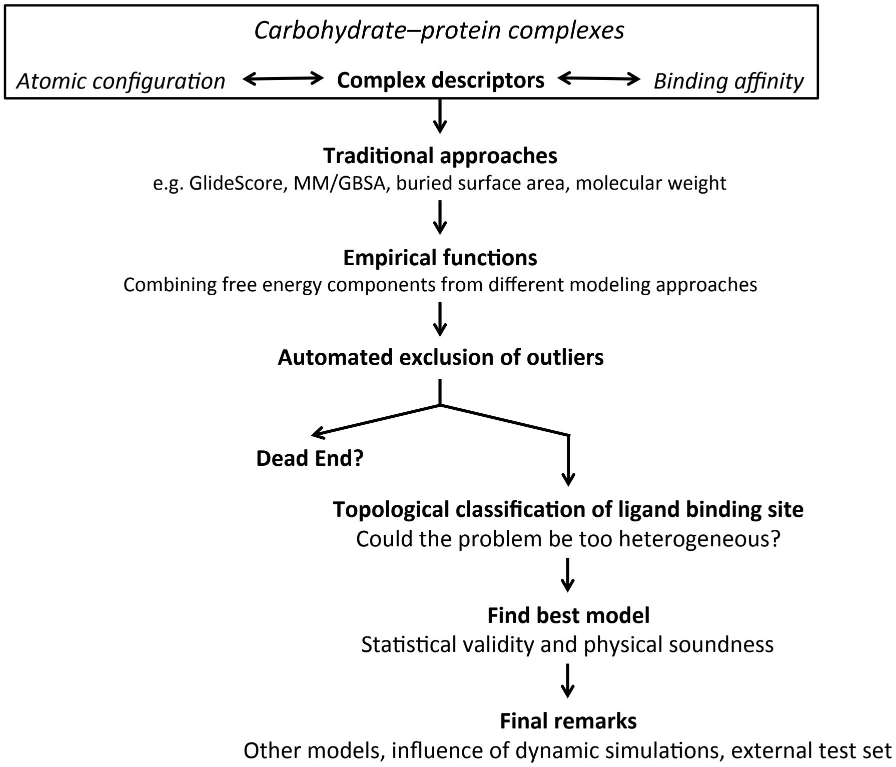

Carbohydrate–protein complexes
Atomic configuration
Complex descriptors
Binding affinity
Traditional approaches
e.g. GlideScore, MM/GBSA, buried surface area, molecular weight
Empirical functions
Combining free energy components from different modeling approaches
Automated exclusion of outliers
Dead End?
Topological classification of ligand binding site
Could the problem be too heterogeneous?
Find best model
Statistical validity and physical soundness
Final remarks
Other models, influence of dynamic simulations, external test set

Figure I-24: Principal stages of the study and layout of the results section. Complex descriptors were used as a proxy to investigate the relationship between binding affinity and three-dimensional structure of carbohydrate–protein systems.

I.4.2 Preliminary investigation using traditional approaches

Two well-established methods presumably capable of predicting binding affinities of ligand–protein systems were employed in this study; namely the Glide XP scoring function and the Molecular Mechanics/Generalized-Born surface area (MM/GBSA) free-energy function. Comparative docking studies on carbohydrate ligands have shown Glide to outperform other docking programs in reproducing the experimentally determined binding modes (Alexacou, *et al.*, 2008; Nurisso, *et al.*, 2008). On the other hand, several studies have demonstrated the usefulness of MM/GBSA predicting binding affinities of a relatively homogenous set of protein systems (Rastelli, *et al.*, 2010; Srivastava and Sastry, 2012) as well as in large data sets comprising diverse protein families (Greenidge, *et al.*, 2013). Therefore, it seemed reasonable to start our investigation by assessing the performance of both approaches on our carbohydrate-specific data set.

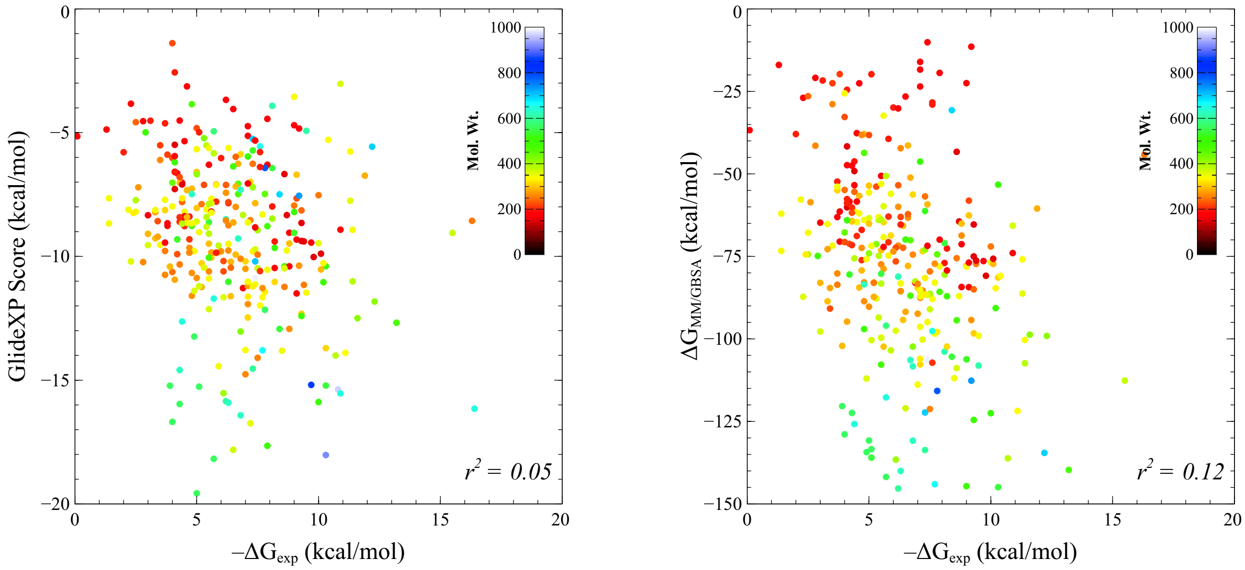


Figure I-25: Correlation plots of experimental free energies in the carbohydrate–protein data set vs. GlideXP scoring function (left) and MM/GBSA free-energy function (right), points are color-coded according to the ligand's molecular weight (*N*=316).

As shown in Figure I-25, scores of both free-energy functions do not correlate well with the experimental binding affinities in our carbohydrate data set (r^2=0.05 and 0.12 for Glide XP score and MM/GBSA, respectively). Although this finding is disappointing, it is not by any means surprising. Despite the reported success of Glide in reproducing crystallographic conformations and database screening, it was shown to yield inaccurate binding affinity predictions in several protein families (Warren, *et al.*, 2006). In general, the prediction accuracy of scoring functions employed in widely used docking programs is known to be system-dependent (Ferrara, *et al.*, 2004; Mooij and Verdonk, 2005; Perola, *et al.*, 2004; Warren, *et al.*, 2006). On the other hand, performance of MM/GBSA in free-energy predictions was in most cases assessed on uniform data sets of ligands binding to the same protein (Rastelli, *et al.*, 2010; Srivastava and Sastry, 2012) or on relatively small data set of different proteins (Hou, *et al.*, 2011a). In the latter case MM/GBSA was shown to exhibit target-dependent variation in prediction accuracy in a manner similar to scoring functions employed in docking (Guimarães and Mathiowetz, 2010; Hou, *et al.*, 2011a).

However, the apparent lack of correlation in Figure I-25 is not dependent on the molecule size; i.e. the Glide XP and MM/GBSA energies incorrectly describe small rigid ligands and larger and more flexible ligands alike. It is worth noting that despite the lack of trends in the horizontal direction (experimental binding affinity), there is a clear trend in the vertical one (calculated interaction energy). Obviously, both energy models were, at least to some extent, biased towards larger ligands awarding them higher scores (i.e. more negative values) in comparison to smaller ligands. This bias would be useful for the purpose of affinity prediction if a paralleling correlation existed between the sizes of the ligands (or the ligand–protein contact areas) and binding affinities. It has been reported, for instance, that binding free energy increases by ~ –1.5 kcal/mol for each non-hydrogen atom in the ligand up to a limit of 15, where it reaches a plateau (Kuntz, *et al.*, 1999). The binding free energy of carbohydrate ligands in particular has been shown to increase quickly for the first few carbohydrate subunits then tapers off (Neumann, *et al.*, 2004). On the other hand, the solvent accessible surface that becomes buried when the ligand and protein associate (i.e. contact area) is a major determinant of strength of interaction (Eisenberg and McLachlan, 1986; Kerzmann, *et al.*, 2008; Ooi, *et al.*, 1987; Pace, 1992; Wang, *et al.*, 2001b). In our data set, however, such correlations between binding affinity and ligand size or contact area were not found (Figure I-26). This was somehow expected given the large diversity and the wide affinity range of the studied carbohydrate–protein complexes.

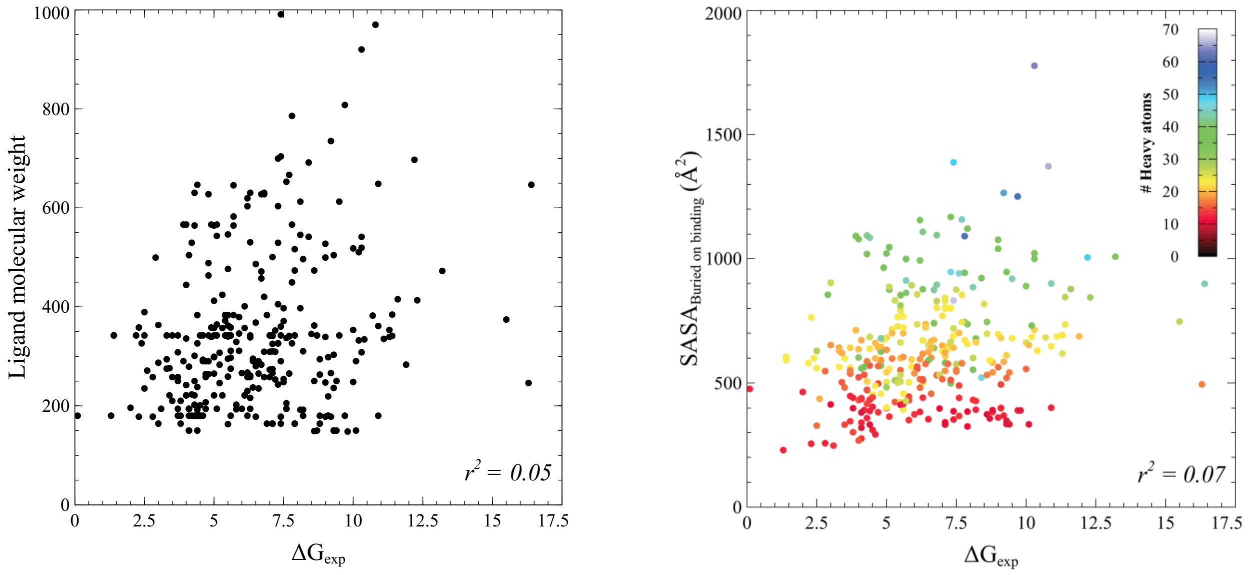


Figure I-26: Correlation plots of experimental free energies in the carbohydrate–protein data set vs. ligand's molecular weight (left) and total SASA (ligand + protein) buried upon binding (right), points are color-coded in the plot on the right according to the number of non-hydrogen atoms in the ligand (*N*=316).

The underlying physical model and mathematical formulation of the empirical scoring function in GlideXP differ significantly from those in the MM/GBSA free-energy function (details are given in the Methods section). Surprisingly, however, the energy scores of both methods correlate well with each other and suffer similarly from size-dependent bias in the calculated energies (Figure I-27). As pointed out previously, both methods, as examples of "mature" free-energy functions, failed to predict binding affinities in carbohydrate–protein systems (Figure I-25). It is important to note, however, that in the preliminary assessments above both methods were used as black boxes and the calculated energies were used "as is" without parameter fitting to the carbohydrate data set. Previous studies on similar problems highlighted the difference in relative importance of certain components of binding free energy in carbohydrate–protein interactions. For example, Laederach and Reilly (2003) reported that electrostatic interactions play a more important role in determining the affinity between a carbohydrate and a protein. Since the MM/GBSA model uses equal weights for the different energy components (electrostatic, vdW, etc.), it is crucial to introduce empirical weighting coefficients when applying it for carbohydrate–protein systems. Similarly, the coefficients employed in GlideXP scoring function were optimized to reproduce the experimental affinities of a training set of 198 complexes (Friesner, *et al.*, 2006). Since the proteins employed to train GlideXP energy model are not necessarily carbohydrate binders, it might also be beneficial to recalibrate the GlideXP coefficients for our carbohydrate-specific set.

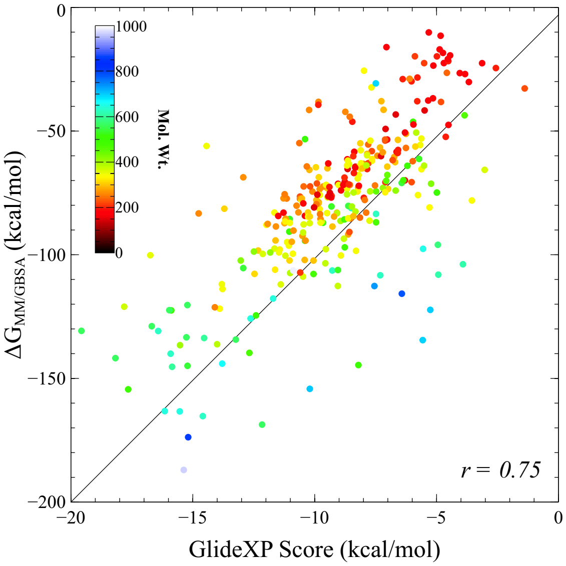


Figure I-27: Correlation plot of GlideXP scores vs. MM/GBSA free-energy estimates in the carbohydrate–protein data set, points are color-coded according to the ligand's molecular weight (N=316).

In addition to the mismatch of the empirical weighting coefficients, GlideXP and MM/GBSA free-energy models have other limitations that must be addressed in order to have a reliable free-energy model for carbohydrate ligands. First of all, both methods would seem to lack corrections for the size-dependent bias described above. Moreover, MM/GBSA does not have a term to estimate the entropic changes associated with ligand–protein binding. Additionally, it might be necessary to correct the solvation treatment by accounting for partial desolvation of ligand and protein upon binding. The poor prediction could also be caused by inaccuracies in the underlying force-field model, which would warrant the employment of alternative force fields to compute non-bonded interactions. Finally, the highly dynamic nature of carbohydrate–protein interactions might necessitate calculating energies as an ensemble average from configurations generated MD simulations for instance. Combining all the possible formulations of the free-energy components leads to a large number of empirical formulations, each having the potential of accurately describing the binding affinities in our carbohydrate–protein data set. These empirical formulations are discussed in the following section.

I.4.3 Empirical free-energy functions

A common theme in development of scoring and free-energy functions is the representation of the binding free energy as a sum of weighted energy terms each representing a physically meaningful component of the binding process, e.g. electrostatics, vdW interactions, hydrogen bonds, entropy, etc.

$$\Delta G_{bind} = \sum_{i} w_i . \Delta G_i$$

The empirical weighting coefficients, w_i's, are typically derived by least-squares fitting to a training set of ligand–protein complexes for which the three-dimensional structures and experimental binding affinities are known. All reported carbohydrate-specific scoring functions are, in fact, empirical functions derived by recalibrating the terms of an existing scoring functions on training sets of carbohydrate–protein complexes, with the addition of terms to improve treatment of special interaction motifs such as C–H···π interactions (Hill and Reilly, 2008; Kerzmann, *et al.*, 2008; Kerzmann, *et al.*, 2006; Laederach and Reilly, 2003).

The following Master Equation was employed as a testing device in the search for a carbohydrate-specific free-energy model. The Master Equation is basically a generic formulation of empirical scoring functions:

Master Equation

$$\Delta G_{bind} = c_1 \Delta G_{inter} + c_2 \Delta G_{solv} + c_3 \Delta G_{strain} + c_4 \Delta S_{lig} + c_5 \Delta G_{reward/penalty}$$

Here ΔG_{inter} is the ligand–protein interaction energy, ΔG_{solv} is the desolvation penalty associated with binding, ΔG_{strain} is the conformational strain penalty, ΔS_{lig} is the entropy lost by the ligand upon binding, and $\Delta G_{reward/penalty}$ represent special rewards and penalties, e.g. $SASA_{ligand}^{buried-on-binding}$. The empirical weighting coefficients c_i's are obtained by fitting the equation, using multiple linear regression, to the experimental binding affinities in our carbohydrate–protein data set.

To broaden the coverage of the potential solution space, we employed various well-established methodologies to describe each term in the Master Equation. The ligand–protein interaction energy, for instance, was estimated using GlideXP scoring function, MM/GBSA free-energy model, and the non-bonded interaction terms from MMFF (general utility force field) and OPLS-AA (force field optimized for carbohydrates). The GlideXP scores were calculated for the unchanged input structures (*in place* mode) and after short geometry optimization (*refine input* mode). Additionally, MM/GBSA and non-bonded force field energies were calculated from the static input structure as well as from an ensemble of configurations extracted from 5.0 ns long MD simulations. The conformational strain was calculated for the ligand alone or for ligand and receptor; and the former was calculated with respect to the closest local energy minimum or to the global minimum obtained from an extensive conformational search. The same applies for the remaining terms in the Master Equation: each term was represented by descriptors varying in the underlying theoretical approximation, degree of sophistication, as well as associated computational cost. The goal was to investigate, as thoroughly as possible, the ability of the available repertoire of methodologies for modeling molecular interactions to formulate a reliable free-energy model for carbohydrate–protein systems. Figure I-28 shows all possible permutations obtainable using different *complex descriptors* at all the positions of the Master Equation. Computational protocols for the employed complex descriptors are discussed in greater detail in the Methods section.

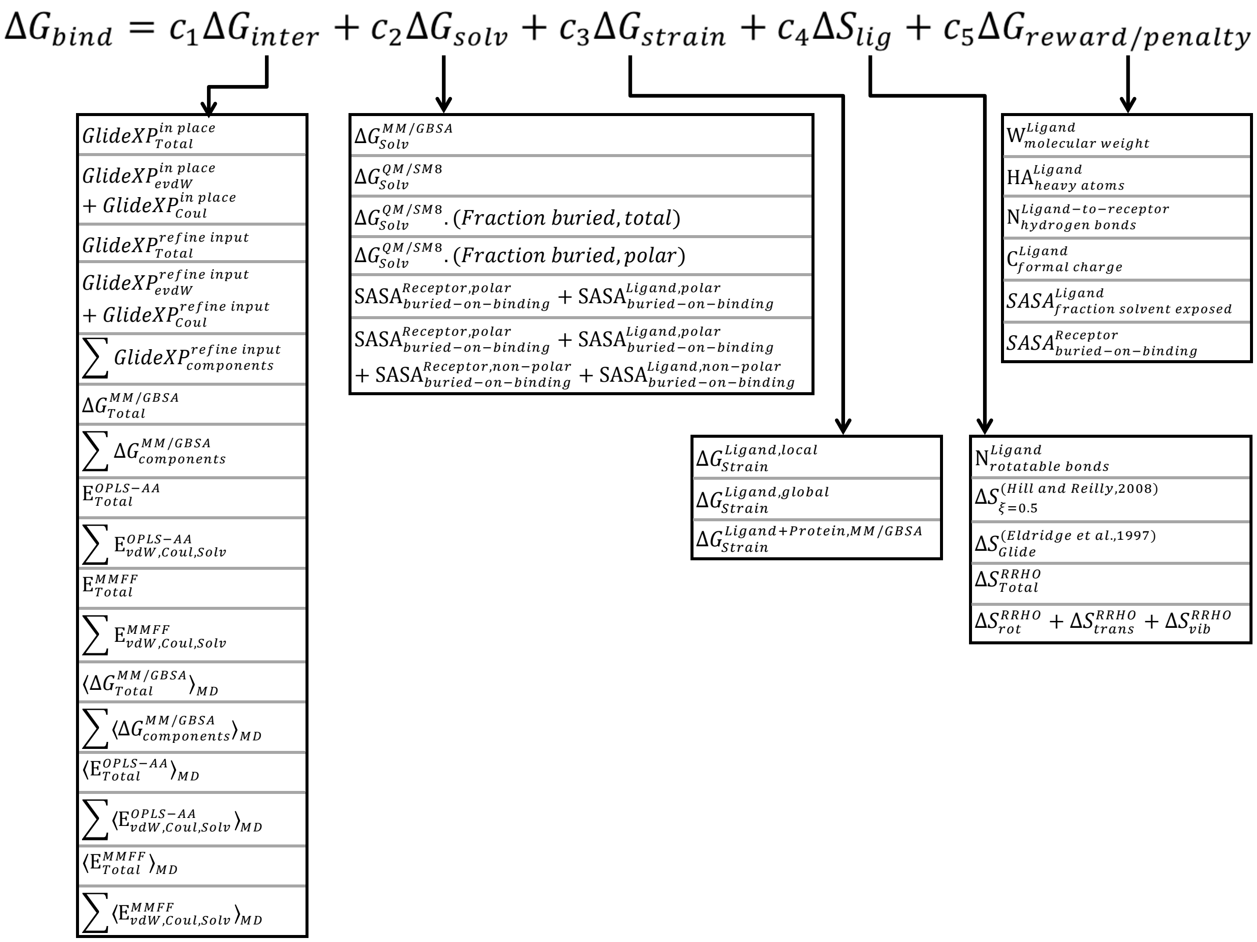

$\Delta G_{bind} = c_1 \Delta G_{inter} + c_2 \Delta G_{solv} + c_3 \Delta G_{strain} + c_4 \Delta S_{lig} + c_5 \Delta G_{reward/penalty}$
$GlideXP_{Total}^{in\ place}$
$GlideXP_{evdW}^{in\ place} + GlideXP_{Coul}^{in\ place}$
$GlideXP_{Total}^{refine\ input}$
$GlideXP_{evdW}^{refine\ input} + GlideXP_{Coul}^{refine\ input}$
$\sum GlideXP_{components}^{refine\ input}$
$\Delta G_{Total}^{MM/GBSA}$
$\sum \Delta G_{components}^{MM/GBSA}$
$\mathrm{E}_{Total}^{OPLS-AA}$
$\sum \mathrm{E}_{vdW,Coul,Solv}^{OPLS-AA}$
$\mathrm{E}_{Total}^{MMFF}$
$\sum \mathrm{E}_{vdW,Coul,Solv}^{MMFF}$
$\langle \Delta G_{Total}^{MM/GBSA} \rangle_{MD}$
$\sum \langle \Delta G_{components}^{MM/GBSA} \rangle_{MD}$
$\langle \mathrm{E}_{Total}^{OPLS-AA} \rangle_{MD}$
$\sum \langle \mathrm{E}_{vdW,Coul,Solv}^{OPLS-AA} \rangle_{MD}$
$\langle \mathrm{E}_{Total}^{MMFF} \rangle_{MD}$
$\sum \langle \mathrm{E}_{vdW,Coul,Solv}^{MMFF} \rangle_{MD}$
$\Delta G_{Solv}^{MM/GBSA}$
$\Delta G_{Solv}^{QM/SM8}$
$\Delta G_{Solv}^{QM/SM8}.(Fraction\ buried, total)$
$\Delta G_{Solv}^{QM/SM8}.(Fraction\ buried, polar)$
$\mathrm{SASA}_{buried-on-binding}^{Receptor,polar} + \mathrm{SASA}_{buried-on-binding}^{Ligand,polar}$
$\mathrm{SASA}_{buried-on-binding}^{Receptor,polar} + \mathrm{SASA}_{buried-on-binding}^{Ligand,polar} + \mathrm{SASA}_{buried-on-binding}^{Receptor,non-polar} + \mathrm{SASA}_{buried-on-binding}^{Ligand,non-polar}$
$\Delta G_{Strain}^{Ligand,local}$
$\Delta G_{Strain}^{Ligand,global}$
$\Delta G_{Strain}^{Ligand+Protein,MM/GBSA}$
$\mathrm{N}_{rotatable\ bonds}^{Ligand}$
$\Delta S_{\xi=0.5}^{(Hill\ and\ Reilly,2008)}$
$\Delta S_{Glide}^{(Eldridge\ et\ al.,1997)}$
ΔS_{Total}^{RRHO}
$\Delta S_{rot}^{RRHO} + \Delta S_{trans}^{RRHO} + \Delta S_{vib}^{RRHO}$
$\mathrm{W}_{molecular\ weight}^{Ligand}$
$\mathrm{HA}_{heavy\ atoms}^{Ligand}$
$\mathrm{N}_{hydrogen\ bonds}^{Ligand-to-receptor}$
$\mathrm{C}_{formal\ charge}^{Ligand}$
$SASA_{fraction\ solvent\ exposed}^{Ligand}$
$SASA_{buried-on-binding}^{Receptor}$

Figure I-28: Overview of the combinations of complex descriptors used to represent empirical free-energy terms in the Master Equation.

A total of 27,215 models were exhaustively enumerated and evaluated by linear fitting to the training set comprising 316 carbohydrate–protein complexes. The adjusted coefficient of determination (adjusted-*r*^{2}) was used to assess the quality of the resultant models. The examined empirical models ranged in complexity from simple equations using a single predictor variable to complex equations using 21 variables. Models using relatively small number of predictor variables (complex descriptors) and exhibiting high adjusted-*r*^{2} would be good candidates for more thorough investigation and statistical validation. Given the reasonable coverage of the solution space provided by the employed pool of complex descriptors, we had high hopes of finding a reliable free-energy model for carbohydrate–protein systems. However, the results did not match the initial expectations.

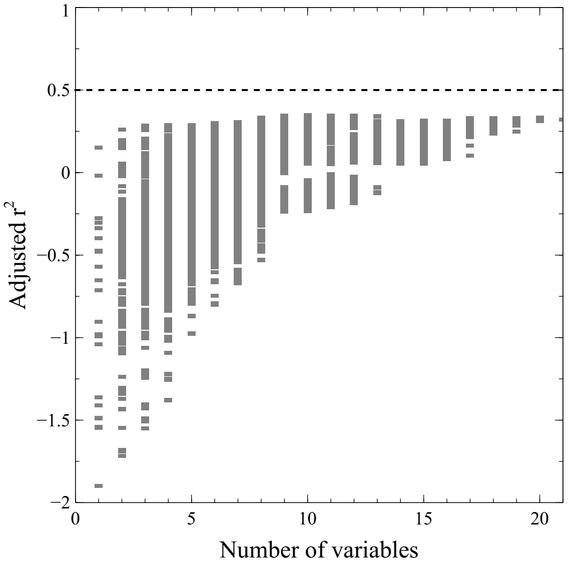


Figure I-29: Statistical assessment of the free-energy models resulting from the combinations of complex descriptors shown in Figure I-28 in the Master Equation. Number of independent variables in the model is plotted on the horizontal axis, while the adjusted-r^2 as a measure of model predictive quality is plotted on the vertical axis. The dotted line marks the value of adjusted-r^2=0.5, which can be used as an arbitrary threshold delineating potentially predictive models from non-predictive models.

The picture portrayed by Figure I-29 would seem to be quite clear; none of the assessed free-energy models, not even those employing more than 20 descriptors, could predict the carbohydrate-binding affinities in our data set. In other words, the search for a free-energy function to quantify carbohydrate–protein binding, which was conducted by mining most of the well-established, highly-regarded, and widely-used computational approaches, turned up empty. It could be argued that in the world of scientific research a negative result is *a result* nonetheless. In fact, peer-reviewed journals dedicated to publishing negative results do exist; the *Journal of Negative Results in Biomedicine* serving as an example. Therefore, it would be reasonable (and rather tempting) to state the following as the outcome of this study, at least up to this stage: the contemporary molecular modeling methodologies with low to moderate computational cost, i.e. excluding thermodynamic integration and free-energy perturbation methods, cannot be used reliably to predict binding affinity of carbohydrate–protein complexes.

Nevertheless, scientific research does not stop at the first negative outcome. The study, thus, continues to investigate the potential causes responsible for the inability of the evaluated methodologies, and combinations thereof, to predict the binding free energy in the studied carbohydrate-specific data set. This lack of explanatory power could be attributed to one (or more) of following reasons:

1. **Inadequacy of computational methods.** All computational approaches for quantification of intermolecular interactions, except for robust statistical mechanical methods such as TI and FEP, rely to varying degrees on approximations in the

employed free-energy functions. Although these approximations typically have sound theoretical foundation, they always involve simplifying the description of, or even completely ignoring crucial aspects of molecular interactions to reduce computational costs. As a consequence, these approximations could eventually lead to fundamentally inaccurate physical description of molecular interactions in particular molecular systems. Nonetheless, it can be demonstrated by examining the literature that a significant number of studies have successfully employed these methods, along with the implied *approximations*, for binding affinity predictions in a wide variety of molecular systems. Therefore, although the impact of these approximations on the prediction accuracy is undeniable, this factor alone cannot explain the observed lack of predictive power of all free-energy models evaluated in this study. For example, inaccuracies in modeling electrostatic interactions could, in principle, be resolved by including a directional hydrogen-bonding term and a special metal-center function (Vedani and Huhta, 1990) or by using a polarizable force field (Halgren and Damm, 2001). However, the anticipated gain in accuracy from improved treatments (of electrostatics) is probably insufficient to compensate for the substantial prediction errors observed in the assessed free-energy models.

2. **Uncertainty of experimental measurements.** The development of usable free-energy models relies to a great extent on the accuracy of the two experimental measurements associated with each ligand–protein complex; namely its geometry and binding affinity. In a critical assessment of errors associated with several crystal structures, DePristo *et al.* (2004) stated that ‘the accuracy of X-ray crystal structures has been widely overestimated’ and consequently that ‘analyses depending on small changes in atom position may be flawed’. Moreover, Søndergaard *et al.* (2009) reported that 36% of ligands in the PDBbind refined data set (Wang, *et al.*, 2005) have artifacts resulting from crystal contacts and that these artifacts detrimentally influence the performance of scoring functions. On the other hand, experimentally determined binding affinities are infamous for the substantial variability resulting from inter-person as well as inter-lab variations (Brown *et al.*, 2009; Kramer *et al.*, 2012). In a recent study by Kramer *et al.*, the experimental uncertainty of heterogeneous public K_i data available in public databases was reported to have a standard deviation of 0.54 log units (Kramer, *et al.*, 2012). Since the data set employed in this study is large and heterogeneous, it is unavoidably prone to *outliers* due to uncertainties in crystal geometry and/or binding affinity.
3. **Heterogeneity of the carbohydrate–protein data set.** An approximate treatment, or even complete disregard of some components of binding free energy, e.g. entropy, have less negative impact on prediction quality in homogeneous data set of related ligands binding to the same target as compared to heterogeneous sets. The well-documented success of methods such as Linear Interaction Energy (LIE) in binding affinity prediction (Nicolotti, *et al.*, 2012) is a good example that commonly-used computational tools, despite the approximations employed therein, can yield useful

predictions especially in homogeneous problems. On the contrary, heterogeneity in the data set for which a predictive scoring function is required could aggravate the consequences of the physical approximations employed in the free-energy models. Development of a scoring function of general applicability is, therefore, a formidable challenge. In case of heterogeneous sets it is more crucial to properly account for all the parameters of molecular interactions. The use of robust methods such as TI and FEP, which employ minimal approximations, is however unaffordable in applications such as virtual database screening and lead optimization. In this case improving the free-energy model by including more terms might be a viable solution. In our case, however, this was found to be of little help as discussed above.

The remaining part of this thesis deals with the investigation of the hypotheses described in points 2 and 3 above, i.e. the potential deleterious effect of outliers and the possible existence of subgroups within the potentially non-homogeneous data set of carbohydrate–protein complexes used in this study.

I.4.4 Exclusion of outliers

In an ideal world, one where financial and human resources are unlimited, the formally correct course of action when outliers are suspected is to *repeat the experiment*. Applying this to our case would mean producing sufficient amounts of the 316 carbohydrate-binding proteins, synthesizing or purchasing all carbohydrate ligands, repeating the biological assays, growing the co-crystals, and finally solving the structures of the 316 complexes. Clearly, this approach is unfeasible. Alternatively, we resorted to mathematics, more specifically to data mining techniques in statistics, for a cheaper and more practical solution. The Genetic Algorithm module in the NeoScore analysis engine described in the Methods section (page 47) was used for automated detection of outliers. In the context of this study, outliers are perceived to be carbohydrate–protein complexes with inherent flaws in geometry and/or affinity measurement. Such complexes poison the data set and preclude the possibility of building meaningful free-energy relationships. Statistical assessment of the scoring functions resulting from combinations of free-energy terms according to the scheme described in Figure I-29 was repeated in conjunction with GA-guided exclusion of outliers (Figure I-30).

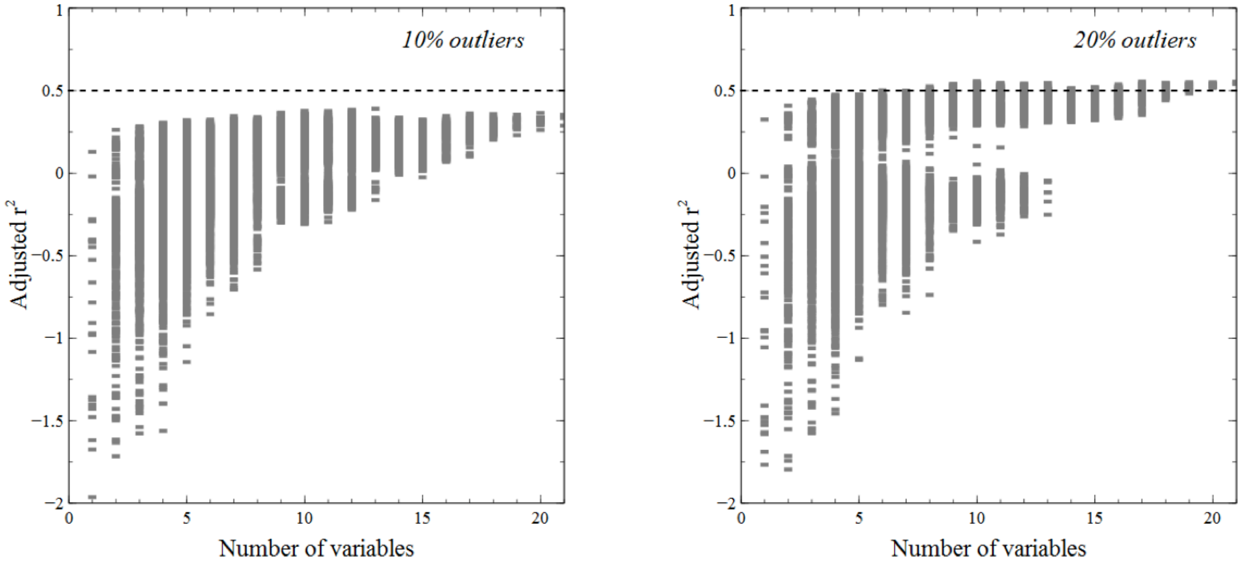


Figure I-30: Results of statistical assessment with automated exclusion of 10% and 20% outliers via the GA module of NeoScore Analysis Engine. Plot description matches that of Figure I-29.

The results shown in Figure I-30 (with outlier exclusion) are not essentially different from those of the models evaluated on the entire data set (Figure I-29). The few models exhibiting adjusted-r^2 slightly values above 0.5 wither did not pass statistical validation tests such as cross-validation or were not physically sensible, e.g. positively correlating entropic penalties with binding affinity. A higher threshold of 30% for outlier exclusion resulted in numerous statistically valid free-energy models for the remaining 70% of the data set (222 carbohydrate–protein complexes). These models, however, carry a substantial risk of representing artificial relationships with no real predictive power, so they are not presented here and will not be discussed further. At this point, the study seemed to have reached a dead end; a fairly large pool of well-established computational methods for modeling molecular interaction failed to produce valid free-energy models for the studied carbohydrate–protein systems, even when potential outliers were eliminated from the training set. In other words, the problem seems to be too heterogeneous to be adequately described by common free-energy models and/or within tractable computational time.

I.4.5 Topological classification of carbohydrate-binding sites

Accounting for solvation effects in molecular interactions of is one of the most challenging issues in structure-based design. Most scoring and free-energy functions rely on certain approximations to account for solvation effects thereby trading accuracy for computational efficiency. Methods combining force fields with implicit solvation model such as MM/PBSA and MM/GBSA are examples of rigorous methods with numerous successful applications in a variety of ligand–protein systems. Their performance, however, is known to be largely system-dependent (Kuhn, *et al.*, 2005; Pearlman, 2005). The physical model employed by both methods pictures the interacting molecules as zones of low dielectricity embedded in a continuum of high-dielectricity, i.e. the solvent. One of the factors that limit the accuracy of this model, among others, is the difficulty in accurately defining the boundary between the two zones of differing dielectric properties (Bordner and Huber, 2003; Boschitsch and Fenley, 2004; Davis and McCammon, 1991; Fogolari *et al.*, 2002; Neves-Petersen and Petersen, 2003). Moreover, Hou *et al.* demonstrated that MM/GBSA predictions are quite sensitive to the solute dielectric constant (Hou, *et al.*, 2011a). The authors recommended that the dielectric parameter 'should be carefully determined according to the characteristics of the protein/ligand binding interface'. Inaccuracy in treatment of dielectric properties could result in errors in the final estimates of solvation contribution to the binding free energy. In principle, these errors would be more or less uniform in homogeneous sets and consequently have less negative impact on final free-energy estimates. In heterogeneous sets, however, binding sites exhibit larger variations in shape and solvent-accessibility. In such cases, the errors introduced by inaccurate dielectric boundary assignment will significantly vary with the topological features of the binding site, and hence have more detrimental effect on accuracy of the calculated free energies.

The next stage of this thesis was based on the following assumption: the extent to which the carbohydrate-binding site is in continuity with the solvent bulk is governed by its shape and solvent accessibility. This, in turn, influences key parameters of the micro-environment where the intermolecular interaction takes place, e.g. dielectric properties. Nevertheless, analytical treatment of these parameters requires long converged conformational sampling in explicit solvent, e.g. by molecular dynamics simulations, which are practically unfeasible. However, the complexity of the free-energy landscape could, in principle, be simplified by defining families of binding site topologies within which the binding micro-environments are roughly identical. Such topological classification could reduce the large and heterogeneous problem to a set of smaller more homogenous problems, for which simple free-energy formulations could be applied.

Therefore, a heuristic method was developed for allocating complexes in our data set into non-overlapping categories based on the geometry of the carbohydrate-binding site. The classification scheme employed in this study was inspired by previously reported methods for characterization of binding site geometries and refined by visual inspection of binding sites in the studied complexes (cf. page 66 for details). Carbohydrate–protein complexes were allocated to one of five topological categories based on shape and degree of surface exposure of the binding site: fully buried, partially buried, small-mouth groove, big-mouth groove, and shallow (cf. Figure I-22, page 76). Figure I-31 shows the distribution of important properties within the different binding site categories in our data set.

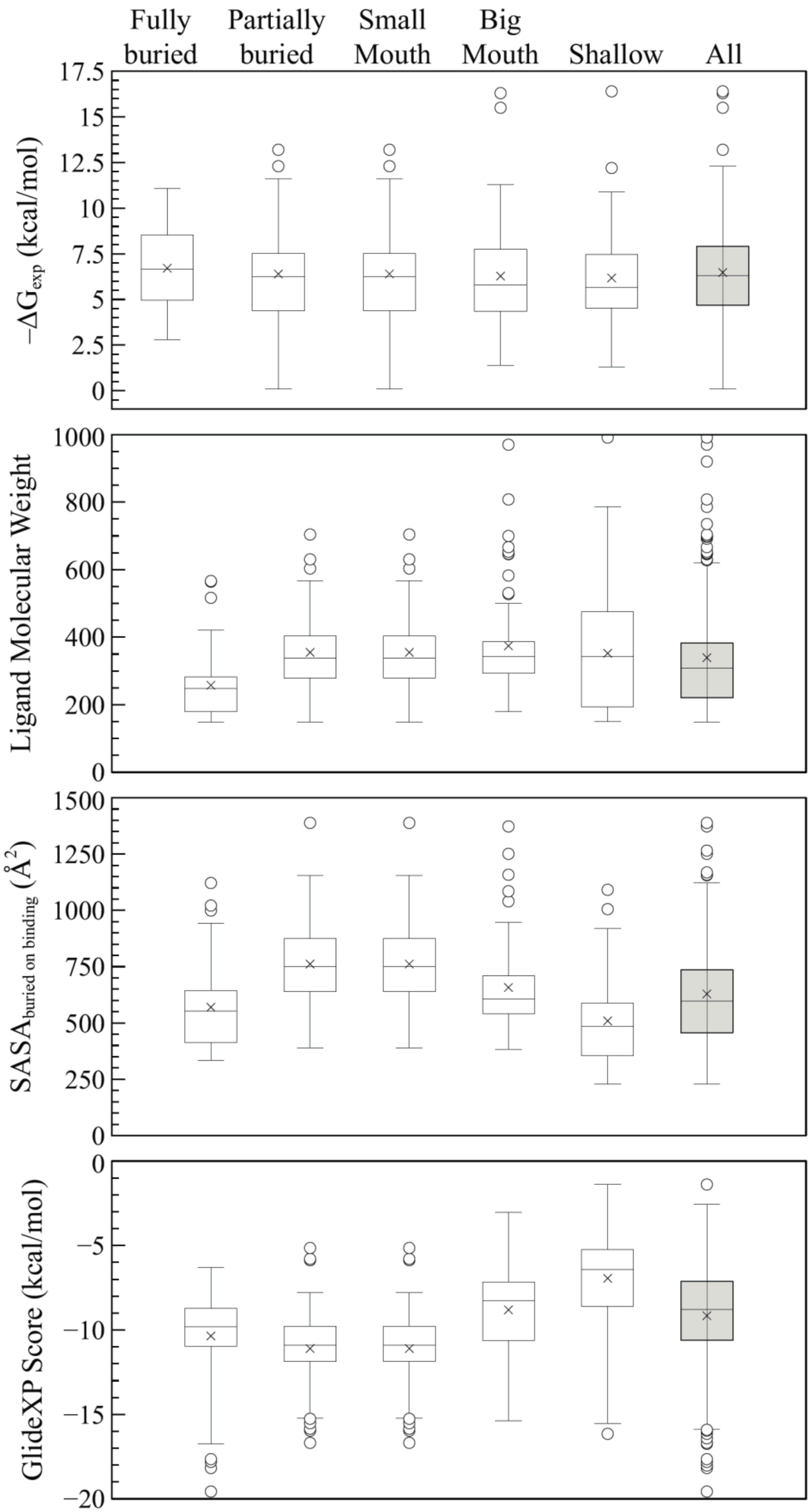


Figure I-31: Distribution of key properties within binding-site categories of the studied carbohydrate data set (non-shaded box plots) and the entire uncategorized data set (shaded box plot). Median indicated by black bar, average indicated by the (×) marker. Boxes indicate the first (25%) and third (75%) quartiles. Whiskers plotted at 1.5 × interquartile range, roughly encompassing 99.7% of the data (mean ± 3σ). Circles represent individual outliers larger than the upper/lower whiskers.

As seen from the topmost plot in Figure I-31, the proposed classification did not segregate complexes according to binding affinity, i.e. carbohydrate ligands could exhibit high or low affinity to their targets regardless of the binding-site topology. Complexes in the fully-buried category span similar range of binding affinities to those in the shallow category. There are, however, differences in molecular weight distributions among the different categories. Fully-buried binding sites tend to accommodate smaller ligands while the three middle categories bind medium-sized ligands. On the other hand, fully exposed shallow binding sites can accommodate a wide range of ligand sizes including relatively large molecules. The size of the contact surface, however, follows a qualitatively different trend with the middle three binding categories exhibiting relatively larger interaction surfaces. The smaller average contact surfaces in fully buried binding sites could be justified by the small sizes of bound ligands in this category. Surprisingly, shallow binding sites show on average contact surfaces of the same scale observed in case of fully buried sites, although the former bind larger ligands. This could be an indication that in shallow carbohydrate-recognition sites, ligands require relatively smaller contact areas to bind to their targets. This observation matches the picture of carbohydrate-binding proteins involved, for instance, in cell-cell communication, e.g. lectins, where the carbohydrate ligand is typically a large biopolymer interacting via a small di- or tri-saccharide motif at its tip. Finally, GlideXP seems to mirror the trends seen in molecular weights and contact surface areas. GlideXP tends to assign lower scores on average to ligands in the fully buried category (smaller ligands) and to those in the shallow category (small contact surface). This trend matches our earlier observation of the size-dependent bias in GlideXP scores.

The influence of categorization on the prediction accuracy of empirical scoring functions is presented in Figure I-32 (right) and compared to the models developed for the entire data set without categorization (left, basically summarizing the results shown in Figure I-29 and Figure I-30). It is quite apparent that independent training of the empirical free-energy functions for individual categories results in substantial improvement in prediction accuracy. A significant proportion of evaluated empirical scoring functions (nearly 20% at 10%-outlier threshold) were capable of reproducing binding affinities of the training set with acceptable accuracy (adjusted-r^2 > 0.7). This result agrees with our initial assumption: the problem we are looking at; predicting carbohydrate–protein binding affinities, is likely a collectively heterogeneous problem of smaller internally more homogeneous sub-problems.

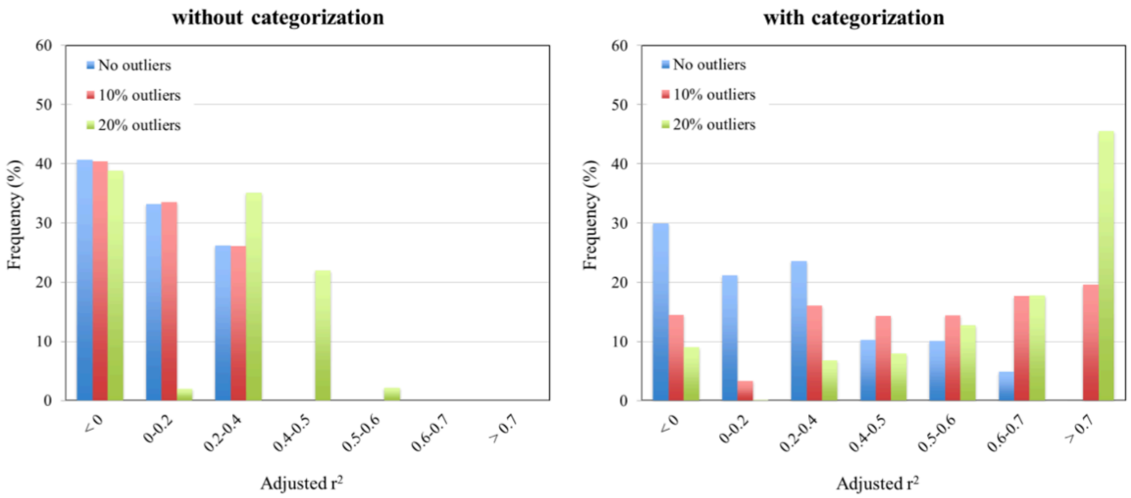
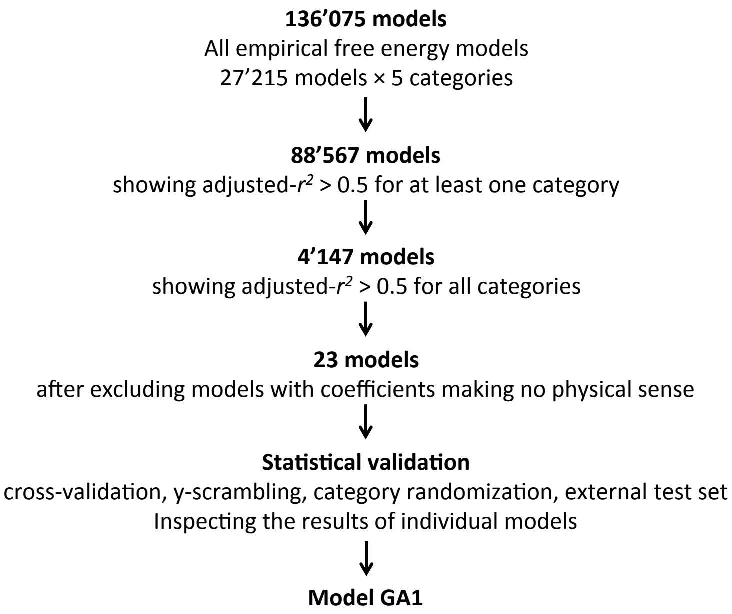


Figure I-32: Comparison of the performance of free-energy models derived from the Master Equation on the uncategorized data set (left) and after categorization according to binding-site topology (right). The vertical axis shows the fraction of all assessed models with adjusted-r^2 in the range defined in the horizontal axis. Models were assessed on the entire set (or the entire category) and with GA-guided automated exclusion of outliers at 10% and 20% thresholds.

I.4.6 Finding the best model

Although a significant number of category-specific free-energy models showed acceptable prediction accuracy in the exhaustive search depicted in Figure I-32 (at both 10% and 20% outlier thresholds), not all of them could be used as objective scoring functions for carbohydrate binding due to a number of reasons. Firstly, a scoring function showing good prediction accuracy in one category did not necessarily perform equally well in other categories. Secondly, fitting of some models in to the binding free energy in certain categories produced regression coefficients that made no physical sense for a subset of the free-energy terms, e.g. entropic penalty or ligand strain energy contributing favorably to affinity. Finally, some of the models that performed well across the five categories and had no physically senseless coefficients did not pass subsequent statistical validation tests such as cross-validation and y-scrambling. Therefore, to find a physically and statistically valid model, it was necessary to filter the pool of evaluated free-energy models (136,075 models resulting from 27,215 × 5 categories). Results of the statistical quality-based and physics-based filtering are presented in Figure I-33.



136’075 models
All empirical free energy models
27’215 models × 5 categories

88’567 models
showing adjusted-r^2 > 0.5 for at least one category

4’147 models
showing adjusted-r^2 > 0.5 for all categories

23 models
after excluding models with coefficients making no physical sense

Statistical validation
cross-validation, y-scrambling, category randomization, external test set
Inspecting the results of individual models

Model GA1

Figure I-33: Filtering the pool of free-energy models resulting from applying the combinations of free-energy terms in the Master Equation to individual binding site topological categories and employing 15% threshold for outlier exclusion.

A total of 23 models survived the statistics and physics-based screens. They were individually subjected to more thorough statistical validation. The model exhibiting the best balance between complexity and performance was designated as the GA1 model. The GA1 model has the following functional form:

Model GA1

$$-\Delta G_{bind} = c_1 E_{Coul}^{Glide} + c_2 E_{vdw}^{Glide} + c_3 SASA_{buried}^{non-polar} + c_4 SASA_{buried}^{polar} + c_5 N_{rot} + c_6 Q_{lig}$$

The model comprises Columbic and van der Waals interaction energies from the Glide scoring function, two solvent-accessible surface area terms accounting for the non-polar and polar SASA that becomes buried on binding, and two reward/penalty terms for the number of rotatable bonds (N_{rot}) and formal charge of the ligand (Q_{lig}). Statistical performance of the model is summarized in Table I-9. The GA model reproduced binding free energies within topological categories with r^2 values ranging from 0.64 to 0.71, RMSE from 1.19 to 1.57 kcal/mol and mean unsigned error of 0.99 to 1.33 kcal/mol in the predicted free energies. Results of leave-one-out and leave-k-out cross-validation confirm robustness and internal consistency of the model. In the leave-k-out cross-validation, the k is chosen such that in each cycle one *seventh* of the training set is removed then predicted using the model trained for the remaining complexes. The perturbation introduced by removing one seventh of the complexes is more significant compared to removing a single complex in leave-one-out cross-validation. The leave-k-out cross-validation, therefore, is a more stringent test for model robustness. Finally, randomization of experimental affinities across carbohydrate–protein complexes in each category resulted in a substantial drop in quality prediction.

Table I-9: Results of statistical validation for the GA1 free-energy model.

Category	N	r^2	RMSE	MUE	q^2_{LOO}	q^2_{LKO}	y-scrambling
Fully buried	62	0.64	1.29	1.08	0.55	0.54	−0.10 (−0.44,0.20)
Partially buried	44	0.67	1.57	1.33	0.55	0.54	−0.02 (−0.37,0.32)
Small mouth	37	0.68	1.19	1.00	0.58	0.57	−0.15 (−0.48,0.20)
Big mouth	54	0.71	1.44	1.09	0.61	0.60	−0.23 (−0.81,0.32)
Shallow	75	0.71	1.29	0.99	0.63	0.62	−0.33 (−0.61,0.12)
Pooled	272	0.69	1.36	1.09	0.59	0.57	n/a
Uncategorized	272	0.24	2.14	1.64	0.17	0.16	n/a

N: number of carbohydrate–protein complexes in the category after outlier exclusion, r^2: coefficient of determination, RMSE: root-mean-squared error (kcal/mol), MUE: mean unsigned error (kcal/mol), q^2: cross-validation r^2, LOO: leave-one-out cross-validation, LKO: leave-k-out cross-validation (k chosen so that the data set is divided into seven subsets), y-scrambling: r^2 values resulting from randomly assigning experimental free energy values amongst the training set complexes, average(minimum, maximum) r^2 values from 100 scrambling cycles.

Prediction errors were pooled from the five binding site topological categories to calculate the values in the *pooled* row in Table I-9. The *pooled* statistical metrics measure the overall performance of the GA1 free-energy model. The GA1 model reproduces binding free energies in the entire data set within RMSE of 1.36 kcal/mol, which corresponds to a factor of 10-off from experimental values. Prediction accuracy of GA1 model is substantially reduced when applied to the entire uncategorized data set as seen from the last row in Table I-9. Figure I-34 presents the influence of the proposed categorization scheme on the performance of the GA free-energy model. The GA1 Model does not seem to exhibit systematic over- or under-estimations in the predicted ΔG values. However, it shows a slight bias in the plot of residuals against experimental ΔG values (Figure I-35), i.e. some high affinity ligands are underestimated while some low affinity ligands are overestimated. On the other hand, in the range $3.0 \leq \Delta G_{bind} \leq 12.0$ kcal/mol, the residuals are more evenly distributed with no clear bias.

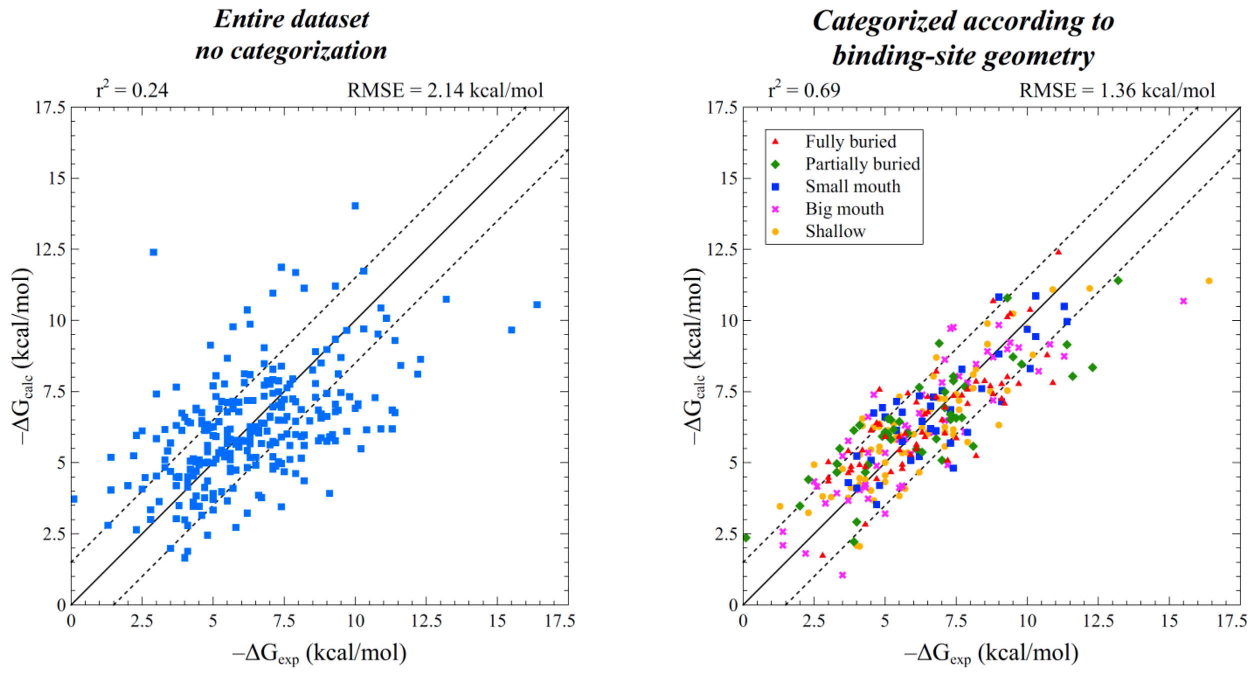


Figure I-34: Distributing the carbohydrate–protein data set into binding site topological categories according to the proposed classification scheme leads to a substantial improvement in the performance of the GA1 empirical free-energy model (N=272). Dashed lines mark 10-fold deviations from experimental binding affinity.

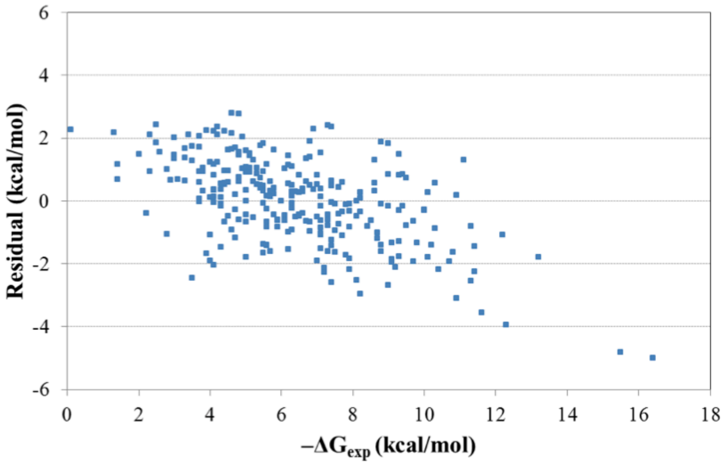


Figure I-35: Residual plot for Model GA1 (N=272), horizontal axis: experimental binding free energy, vertical axis: prediction error ($\Delta G_{calculated} - \Delta G_{experimental}$).

The improvement in the performance of the GA1 model could be a mere consequence of reducing the dimensionality of the problem from the total of 272 complexes in the complete data set to smaller subsets of 37 to 75 complexes per category. To examine this possibility, carbohydrate–protein complexes were randomly allocated to five dummy categories having the same sizes of the binding-site topological categories disregarding the actual binding-site anatomy. The GA1 model was then applied to the resultant categories and its performance was evaluated. Average performance results from 100 category-randomization runs are presented in Table I-10. The apparent deterioration of the GA1 model performance confirms that mixing complexes with differing binding site topologies in small categories is not alone sufficient to yield useful free-energy correlations. This further confirms the relevance of actual binding site topology in defining the free-energy response surface within categories and also verifies the validity of the proposed classification scheme.

Table I-10: Statistical validation for GA1 model when complexes are randomly allocated to binding site topological categories (average of 100 runs). See the remarks below Table I-9 for description of column headings.

Category	**N**	***r*^{2}**	**RMSE**	**MUE**	***q*$^2_{\text{LOO}}$**	***q*$^2_{\text{LKO}}$**
Fully buried	62	0.28	2.03	1.60	−0.03	−0.06
Partially buried	44	0.33	1.93	1.53	−0.09	−0.12
Small mouth	37	0.36	1.88	1.50	<−1	<−1
Big mouth	54	0.33	1.96	1.55	−0.01	−0.04
Shallow	75	0.33	1.99	1.57	0.11	0.09
Pooled	272	0.34	1.98	1.56	<−1	<−1

Since the GA1 free-energy model was fitted five times, once for each binding site topological category, five sets of empirical weighting coefficients were obtained. The empirical coefficients are listed in Table I-11 after multiplying each of them by the mean and standard deviation of the corresponding energy components for each category. The resulting values are the mean (± standard deviation) of the free energy contributed by each component in the GA1 model to the total binding free energy within individual categories. As seen from Table I-11, the values of the average energy contributions (and the underlying empirical weighting coefficients) show evident category-dependent variations. It is difficult, however, to provide concise physical interpretation or draw general conclusions from the observed variations for two main reasons:

1. **Unavoidable data set dependence.** It is important to remember that all empirical weighting coefficients are in fact linear-regression coefficients derived by fitting an equation to some training set, either the entire carbohydrate–protein data set or subsets thereof in our case. Any training set is essentially limited to what could have been gathered at the time of study. In absence of universal data sets, it is practically impossible to avoid some degree of training set dependence. Therefore, one should be very wary of hasty generalization of 'findings' obtained from a limited data set to the entire population it is assumed to represent. It would be inaccurate to assume that the 37 carbohydrate–protein complexes in the small-mouth category, for instance, constitute a sufficiently representative sample of the entire population of "carbohydrate binding proteins with small-mouth binding-site topology".
2. **Inherent complexity of the free-energy landscape.** Empirical free-energy models are based on the additivity assumption, i.e. the total free energy, a rather complex many-body macroscopic quantity, can be expressed as a sum of independent free-energy components. It is well-known, however, that the free-energy components

commonly used to decompose binding free energy are cooperative rather than additive (Baum, *et al.*, 2010; Dill, 1997; Williams, *et al.*, 1993; Williams, *et al.*, 2004). In our case it is nontrivial, for instance, to associate the contribution of the $SASA_{buried}^{polar}$ component in model GA1 with a specific free-energy component, since it is employed as proxy for a multitude of, sometimes opposing, binding events, e.g. desolvation costs, favorable (or unfavorable) electrostatic interactions or hydrogen bonding, release of bound water molecules, etc.

Table I-11: Average contributions of individual free-energy components in the GA1 free-energy model to the total binding free energy in different binding site topological categories. Values are given as mean ± standard deviation in kcal/mol.

Category	E_{vdw}^{Glide}	E_{Coul}^{Glide}	$SASA_{buried}^{non-polar}$	$SASA_{buried}^{polar}$	N_{rot}	Q_{lig}
Fully buried	4.58 ± 1.83	7.85 ± 2.87	−1.62 ± 0.83	−1.96 ± 0.57	−1.61 ± 0.61	−0.72 ± 1.89
Partially buried	3.60 ± 1.20	5.93 ± 2.66	0.66 ± 0.31	−2.88 ± 0.84	−0.27 ± 0.11	−0.64 ± 2.39
Small mouth	−6.58 ± 3.24	3.90 ± 2.48	10.27 ± 4.69	4.66 ± 1.54	−3.87 ± 1.89	−1.47 ± 1.52
Big mouth	7.82 ± 3.30	4.73 ± 1.59	−0.15 ± 0.07	−4.81 ± 1.35	−1.49 ± 0.75	0.05 ± 0.34
Shallow	2.22 ± 1.40	3.85 ± 1.34	1.05 ± 0.56	1.44 ± 0.46	−2.41 ± 1.32	−0.02 ± 0.12

Despite the apparent difficulty in uncovering the exact physical interpretation for the differences in empirical coefficients across different categories, a couple of interesting trends can be noted. Firstly, the contribution of electrostatic interactions to the total free energy is relatively larger in the fully buried and partially buried categories. This could be attributed to the differences in rewards for releasing the more trapped water molecules in these two categories compared the relatively more freely exchangeable waters in the remaining categories. Secondly, existence of charged groups (reflected by the formal charge of the ligand, Q_{lig}) is associated with moderate penalty in the fully buried, partially buried and small mouth categories. In the big mouth and shallow categories, however, the contribution of Q_{lig} to binding free energy is nearly negligible. This could be justified by the expected higher cost for removing charges from bulk solvent to the protein interior in the former three categories, while in the latter two categories the formal charge could interact with the solvent to some extent. It is also noteworthy that the contribution of electrostatic interactions to the binding free energy is roughly similar to those of vdW interactions, which is in agreement with the JA model reported by Hill and Reilly (2008) on an expanded carbohydrate data set, which disagrees, however, with the free-energy model reported earlier by Laederach and Reilly (2003) on the smaller dataset.

I.4.7 Final remarks

I.4.7.1 Other free-energy models

Several empirical free-energy models exhibited comparable prediction accuracy to the Glide-based GA1 free-energy model. Three exemplary models are briefly discussed here, MMFF-based GA2 model (Table I-12), OPLS-based GA3 model (Table I-13), and MM/GBSA-based GA4 model (Table I-14). Changing the potential energy function used to calculate non-bonded interactions only slightly decreased model performance in most of the cases. The MMFF-based model performed slightly better than the OPLS-based model. On the other hand, the MM/GBSA-based model (using the largest number of descriptors) exhibited the lowest quality in comparison to the former two.

Model GA2

$$-\Delta G_{bind} = c_1 E_{Coul}^{MMFF} + c_2 E_{vdw}^{MMFF} + c_3 E_{solvation}^{MMFF} + c_4 SASA_{buried}^{non-polar} + c_5 SASA_{buried}^{polar} + c_6 N_{rot} + c_7 N_{inonized\ groups}$$

Table I-12: Results of statistical validation for the GA2 free-energy model. See the remarks below Table I-9 for description of column headings.

Category	N	r^2	RMSE	MUE	q^2_{LOO}	q^2_{LKO}	y-scrambling
Fully buried	62	0.61	1.32	1.09	0.44	0.43	–0.15 (–0.48,0.18)
Partially buried	44	0.74	1.32	1.07	0.62	0.60	–0.07 (–0.47,0.34)
Small mouth	37	0.67	1.33	1.11	0.45	0.43	–0.32 (–0.80,0.11)
Big mouth	54	0.64	1.67	1.33	0.50	0.47	–0.05 (–0.43,0.34)
Shallow	73	0.79	1.12	0.88	0.73	0.70	–0.21 (–0.49,0.07)
Pooled	270	0.70	1.35	1.08	0.55	0.53	n/a
Uncategorized	270	0.21	2.23	1.72	0.10	0.10	n/a

Model GA3

$$-\Delta G_{bind} = c_1 E_{Coul}^{OPLS} + c_2 E_{vdw}^{OPLS} + c_3 E_{solvation}^{OPLS} + c_4 SASA_{buried}^{non-polar} + c_5 SASA_{buried}^{polar} + c_6 N_{rot} + c_7 N_{inonized\ groups}$$

Table I-13: Results of statistical validation for the GA3 free-energy model. See the remarks below Table I-9 for description of column headings.

Category	**N**	***r*²**	**RMSE**	**MUE**	***q*²LOO**	***q*²LKO**	**y-scrambling**
Fully buried	61	0.57	1.37	1.09	0.44	0.42	−0.18 (−0.48,0.14)
Partially buried	44	0.67	1.56	1.30	0.54	0.52	0.11 (−0.51,0.36)
Small mouth	37	0.71	1.25	0.99	0.56	0.55	−0.27 (−0.87,0.35)
Big mouth	54	0.74	1.40	1.12	0.63	0.61	−0.08 (−0.40,0.26)
Shallow	73	0.79	1.10	0.91	0.73	0.72	−0.26 (−0.57,0.15)
Pooled	269	0.70	1.33	1.07	0.58	0.56	N/A
Uncategorized	269	0.11	2.35	1.78	0.02	0.01	N/A

Model GA4

$$-\Delta G_{bind} = c_1 \Delta G_{Covalent}^{MM/GBSA} + c_2 \Delta G_{Coul}^{MM/GBSA} + c_3 \Delta G_{vdW}^{MM/GBSA} + c_4 \Delta G_{Packing}^{MM/GBSA} + c_5 \Delta G_{SelfCont}^{MM/GBSA} + c_6 \Delta G_{Lipo}^{MM/GBSA} + c_7 \Delta G_{solvation}^{MM/GBSA} + c_8 SASA_{H-bond}^{non-polar} + c_9 SASA_{buried}^{polar} + c_{10} N_{rot} + c_{11} N_{inonized\ groups}$$

Table I-14: Results of statistical validation for the GA4 free-energy model. See the remarks below Table I-9 for description of column headings.

Category	**N**	***r*²**	**RMSE**	**MUE**	***q*²LOO**	***q*²LKO**	**y-scrambling**
Fully buried	61	0.67	1.22	1.04	0.46	0.42	−0.12 (−0.47,0.19)
Partially buried	44	0.65	1.64	1.35	0.21	0.12	0.05 (−0.33,0.49)
Small mouth	37	0.74	1.11	0.94	0.48	0.41	0.13 (−0.51,0.59)
Big mouth	54	0.64	1.68	1.28	0.34	0.30	0.06 (−0.23,0.45)
Shallow	74	0.79	1.13	0.92	0.64	0.61	−0.09 (−0.46,0.24)
Pooled	270	0.69	1.36	1.09	0.43	0.37	N/A
Uncategorized	270	0.07	2.43	1.88	-0.07	-0.07	N/A

I.4.7.2 Influence of molecular dynamics simulations

In reality molecules are not static, and interesting macroscopic quantities such as binding free energies are in fact ensemble averages over a large number of microstates. In principle, efficient sampling of the phase space of the carbohydrate–protein system is necessary to build a realistic model thereof. To account for their configurational flexibility, the studied systems were subjected to 5.0 ns molecular dynamics simulations. Snapshots were taken at fixed time intervals from the MD trajectory and used to calculate dynamically averaged (4D) interactions. Fluctuations in non-bonded interaction energies from three different methods (MMFF and OPLS non-bonded interactions and the MM/GBSA Δ*G* estimates) were calculated as the coefficient of variations, i.e. standard deviation divided by mean, along the MD trajectory of each system. Figure I-36 shows the distribution of dynamic fluctuations in the values (from the three methods) in carbohydrate–protein complexes grouped according to binding site topological categories. Expectedly, ligands bound in buried binding sites (fully-buried and partially-buried categories) exhibited smaller fluctuations in the calculated interaction energies along MD simulations than ligands in relatively more exposed sites (small-, big-mouth, and shallow categories).

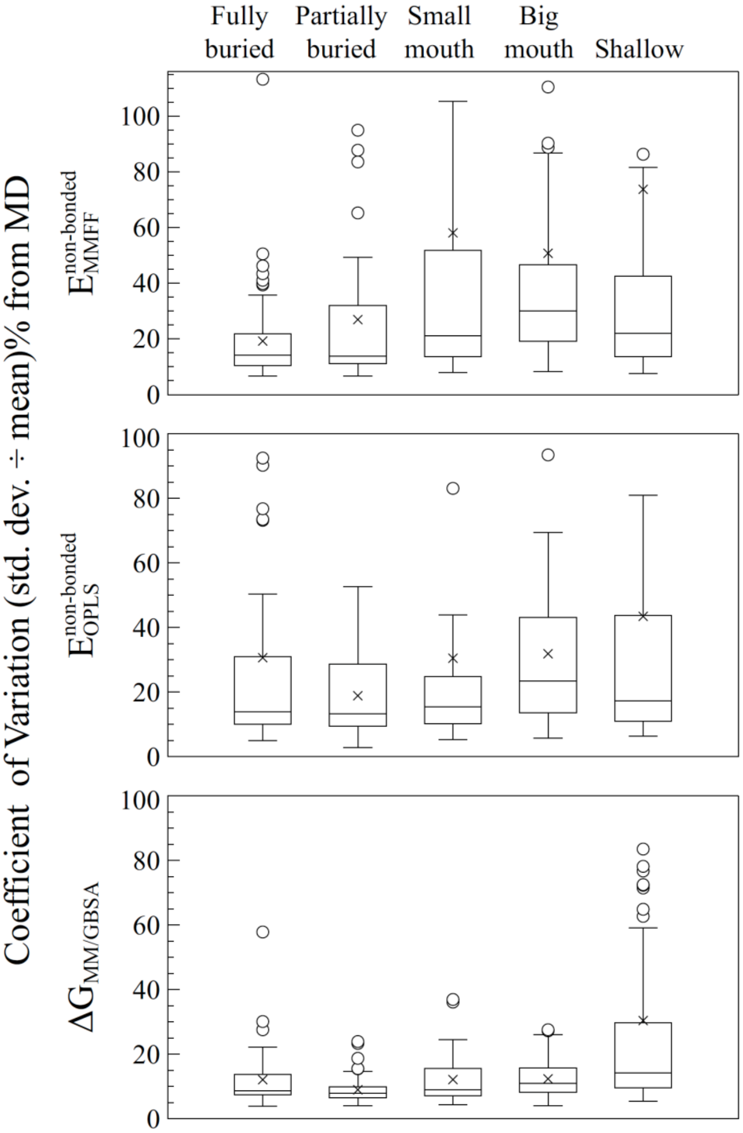


Figure I-36: Distribution of dynamic fluctuations of interaction energies calculated by three different methods (MMFF and OPLS non-bonded interactions and MM/GBSA Δ*G* estimates) within binding-site categories of the studied carbohydrate data set. Dynamic fluctuations are represented as coefficient of variation of the calculated values in the extracted MD frames for each complex. Median indicated by black bar, average indicated by the (×) marker. Boxes indicate the first (25%) and third (75%) quartiles. Whiskers plotted at 1.5 * interquartile range, roughly encompassing 99.7% of the data (mean ± 3σ). Circles represent individual outliers larger than the upper/lower whiskers.

To investigate the significance of the dynamic nature of carbohydrate–protein interactions, energy terms in GA2, GA3, and GA4 models were replaced by their corresponding molecular dynamics-derived averages. Subsequently, prediction accuracy of the resulting models GA2d, GA3d, and GA4d were evaluated and compared to the corresponding *static* models (Table I-15).

Model GA2d

$$-\Delta G_{bind} = c_1\langle E_{Coul}^{MMFF}\rangle_{MD} + c_2\langle E_{vdw}^{MMFF}\rangle_{MD} + c_3\langle E_{solvation}^{MMFF}\rangle_{MD} + c_4 SASA_{buried}^{non-polar} + c_5 SASA_{buried}^{polar} + c_6 N_{rot} + c_7 N_{inonized\ groups}$$

Model GA3d

$$-\Delta G_{bind} = c_1\langle E_{Coul}^{OPLS}\rangle_{MD} + c_2\langle E_{vdw}^{OPLS}\rangle_{MD} + c_3\langle E_{solvation}^{OPLS}\rangle_{MD} + c_4 SASA_{buried}^{non-polar} + c_5 SASA_{buried}^{polar} + c_6 N_{rot} + c_7 N_{inonized\ groups}$$

Model GA4d

$$-\Delta G_{bind} = c_1\langle \Delta G_{Covalent}^{MM/GBSA}\rangle_{MD} + c_2\langle \Delta G_{Coul}^{MM/GBSA}\rangle_{MD} + c_3\langle \Delta G_{vdW}^{MM/GBSA}\rangle_{MD} + c_4\langle \Delta G_{Packing}^{MM/GBSA}\rangle_{MD} + c_5\langle \Delta G_{SelfCont}^{MM/GBSA}\rangle_{MD} + c_6\langle \Delta G_{Lipo}^{MM/GBSA}\rangle_{MD} + c_7\langle \Delta G_{solvation}^{MM/GBSA}\rangle_{MD} + c_8 SASA_{H-bond}^{non-polar} + c_9 SASA_{buried}^{polar} + c_{10} N_{rot} + c_{11} N_{inonized\ groups}$$

Table I-15: Comparison of the performance of models using static interaction energies (GA2, 3, and 4) to models using average interaction energies from MD simulations (GA2d, 3d, and 4d). The values given here represent the pooled performance of the evaluated models in the five topological categories. Column headings are described in the remarks below Table I-9.

Model	**N**	r^2	**RMSE**	**MUE**	q^2_{LOO}	q^2_{LKO}
GA2	270	0.70	1.35	1.08	0.55	0.53
GA2d	270	0.47	1.83	1.40	0.17	0.12
GA3	269	0.70	1.33	1.07	0.58	0.56
GA3d	269	0.45	1.87	1.47	0.13	0.08
GA4	270	0.69	1.36	1.09	0.43	0.37
GA4d	270	0.48	1.80	1.37	< –1	< –1

According to the results shown in Table I-15, using dynamic averages of interaction energies instead of the values calculated from a fixed geometry had a negative impact on the prediction quality of the free-energy models. Although this result seems counter-intuitive, it is not without precedence. Studies comparing static and MD-based free energy predictions using MM/GBSA (Hou, *et al.*, 2011a) and MM/PBSA (Hou, *et al.*, 2011a; Kuhn, *et al.*, 2005) reached similar conclusions. It was even observed that in some ligand–protein systems, such as neuraminidase, the predictions based on relatively short MD simulations are slightly better than those based on longer MD simulations (Hou, *et al.*, 2011a).

Genheden and Ryde (Genheden and Ryde, 2010) suggested that the observed drop in prediction accuracy upon using MD averages could be caused by the relatively large variations in the free-energy components among snapshots from MD simulations (Gohlke and Case, 2004; Pearlman, 2005; Stoica *et al.*, 2008). According to an analysis by Genheden and Ryde, to achieve accuracy in the range of experimental results and draw statistically significant conclusions, a substantially larger number of energy calculations might be necessary (Genheden and Ryde, 2010). According to their calculations, 400-22,500 separate energy calculations are required rather than 10–200, which are traditionally employed. They also indicated that MD simulations runs of 10 ns are too short to achieve good convergence.

MD averages in this study were calculated from 25 snapshots extracted from 5 ns long MD simulations, clearly below the recommended (and rather impractical) ranges. It might be important to point out that MD simulations in explicit solvent for the 316 complexes studied in this work required a total of 56’564 CPU hours (6.5 CPU years on a single core). Fortunately, however, the MD simulations were performed on an in-house cluster comprising 96 processors. Indeed, if it were not for the availability of sufficient computational resources and the use of parallel computing, the MD simulations used in this study would not have been finished within the time frame of the study itself. From the application perspective, the use of relatively long MD simulations is quite impractical: on average, MD simulation for a single carbohydrate–protein system (5 ns) requires 179 hours to finish on a single 3.0 GHz processor. In the light of these results, MD simulations of reasonable length do not seem to offer any real advantage in binding-affinity predictions. They are more likely to just introduce counter-productive fluctuations in the calculated interaction energies, not to mention adding a long and unnecessary computational time. MD simulations lengths necessary to achieve acceptable accuracy and convergence are simply too computationally expensive for use in lead optimization or database screening applications.

I.4.7.3 Application to external test set

To further test the predictive power of the GA1 model, it was applied to an external test set comprising 106 mannose-based ligands whose binding affinities to bacterial FimH adhesin were experimentally determined in our group (Eid, *et al.*, 2013). Results of the external validation were, however, unsatisfactory (Figure I-37). Although the p^2 calculated with respect to the training set variance was 0.64, this value is not an accurate indication of the true predictive power of the GA1 model. Conversely, the value of p^2 calculated with respect to the smaller variance in the test set was 0.0 only. This disparity is a result of the larger number of complexes and the narrow range of affinities of the test set ligands in comparison to the training set. It could be argued, however, that despite the low p^2, the model is satisfactorily predictive in the higher affinity range of the test set molecules.

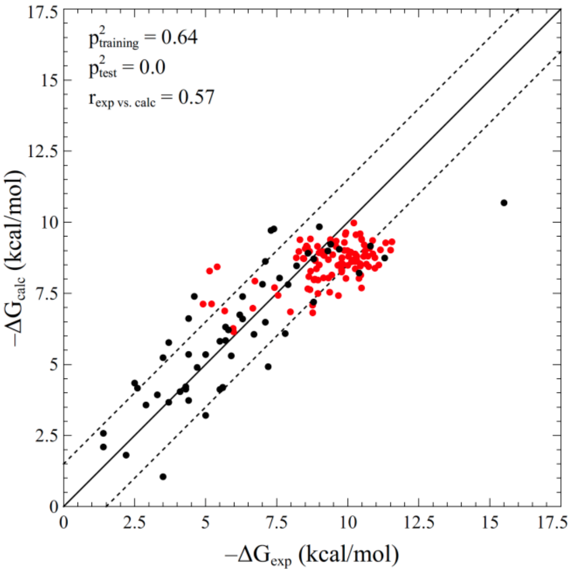


Figure I-37: Plot of experimental binding free energy (horizontal axis) vs. free energy prediction from the GA1 model using the regression coefficients of the big mouth category. Black and red points represent complexes of the training set (54 carbohydrate–protein complexes in the big mouth category) and the test set (106 FimH ligands from Eid *et al.* 2013), respectively. Dashed lines are drawn at factors of 10 from the experimental value.

It is worth noting, however, that despite the inaccuracies in the absolute values of predicted binding affinities of the training set (p^2 = 0.0), the scores from GA1 model showed good correlation with experimental values (linear correlation r = 0.57). Therefore, although the GA1 free-energy model cannot be used to predict absolute FimH binding affinities, it can provide good ranking of a set of related lead structures, which makes it useful for lead optimization and database mining purposes. As a final remark, the lower prediction quality of the GA model could be justified by the fact that it was developed and optimized for a diverse set of carbohydrate–protein complexes. It might, thus, be less capable of accurately mapping the subtle differences in a set of very similar ligands binding to a single target, which is indeed the case in the test set molecules. In cases where proper identification of the finer details of ligand–protein interactions is crucial, a specifically calibrated model, e.g. multidimensional-QSAR model (Vedani and Dobler, 2002; Vedani, *et al.*, 2005; Vedani and Zbinden, 1998) such as the one employed by Eid *et al.*, would give more accurate predictions.

I.5. Summary and conclusion

Carbohydrates are the most abundant natural products. They are involved in a wide spectrum of biological processes including energy production and storage, protein folding, cell-cell communication, and modulating immune response. Moreover, some carbohydrate-binding proteins are connected to infectious diseases both viral, e.g. DC-SIGN, and bacterial, e.g. FimH adhesin. The discovery and functional characterization of carbohydrate-related biomolecules over the past three decades highlighted their tremendous potential as drug targets in numerous disease areas. However, only a limited number of carbohydrate-based drugs have reached the market so far and carbohydrates are still considered untapped sources for new therapeutic agents. The increasing numbers of experimentally determined 3D-structures of carbohydrate-binding proteins provide the basis for structure-based design tools, e.g. virtual screening and *de novo* design, and could thereby accelerate rational design and optimization of carbohydrate leads. Nonetheless, well-established modeling methodologies for biomolecules, e.g. force fields, docking, scoring functions, etc., were optimized for peptides, proteins and nucleic acids and relatively neglecting carbohydrates. Thus, there is an urgent need for alternative methodologies or adaptation of current methodologies to improve their accuracy in modeling carbohydrate–protein interactions.

Carbohydrates are generally considered to be a molecular-modeling challenge due to certain peculiarities in their structure and the way they interact with protein targets and water molecules. Only a handful of attempts specifically dealing with quantification of carbohydrate–protein binding are reported. The earlier studies, however, would seem to have two major limitations: the small size of the employed training set and the lack of an external validation for the proposed scoring function(s). Moreover, the question of the target-dependence of scoring functions has not yet been addressed: why is it that certain scoring functions could predict binding affinities accurately in some protein families and fail in others? Towards this end we gathered and refined a large and diverse data set of 316 carbohydrate−protein complexes with experimentally determined binding affinities. We thoroughly investigated empirical formulations of structure-based descriptors with the aim of developing a reliable scheme for prediction of binding affinities of carbohydrate–protein associations. The investigated descriptors included non-bonded interaction energies from MMFF (general utility) and OPLS (carbohydrate optimized) force fields, Glide score and its components, solvation free energy from the quantum-mechanical SM8 model, the MM/GBSA free-energy model, ligand internal strain (local and global), various entropy estimates including the rigid-rotor harmonic-oscillator (RRHO) approximation, as well as several solvent-accessible surface area (SASA) values covering different aspects of the ligand–protein association process. The descriptor pool (~200 descriptors), thus, extends across a significant portion of the potential solution space. All possible permutations of relevant descriptors were exhaustively enumerated and the performance of the resultant empirical functions was evaluated. To our surprise, none of the assessed functions satisfactorily predicted binding affinities in our data set, even functions with more than 20 terms. This was rather disappointing, since the employed pool of descriptors covered a very wide scope of structural and energetic features. This could lead to the conclusion that the current repertoire of structure-based properties is not sufficient for quantifying carbohydrate–protein binding.

The investigation was extended under the assumption that the extent to which the carbohydrate-binding site is in continuity with the solvent bulk is governed by its shape and solvent accessibility. This in turn influences key parameters of the micro-environment where the intermolecular interaction takes place, e.g. dielectric properties. In principle, the complexity of the free energy landscape could be simplified by defining families of binding site topologies within which the binding micro-environments are roughly identical. Such topological classification could reduce the large and heterogeneous problem to a set of smaller more homogenous problems, for which simple free energy formulations could be applied. Therefore, a heuristic method was developed for allocating complexes in our data set into non-overlapping categories based on the geometry of the carbohydrate-binding site. The classification scheme employed in this study is based on previously reported methods for characterization of binding site geometries and refined by visual inspection of binding sites in the studied complexes. Carbohydrate–protein complexes were allocated to one of five topological categories based on shape and degree of surface exposure of the binding site: fully buried, partially buried, small-mouth groove, big-mouth groove, and shallow.

We assessed the influence of the suggested classification scheme on the performance of several empirical free energy functions by fitting the empirical function to each category separately. The results clearly indicated that the independent training of the empirical free energy functions for individual categories results in substantial improvement in prediction accuracy. A significant proportion of evaluated empirical scoring functions (nearly 20% at 10%-outlier threshold) were capable of reproducing binding affinities of the training set with acceptable accuracy (adjusted-r^2 > 0.7). This result agrees with our initial assumption: the problem we are looking at; predicting carbohydrate–protein binding affinities, is likely a collectively heterogeneous problem of smaller internally more homogeneous sub-problems. The best performing free energy model (GA1 model) exhibited an overall r^2 of 0.69 and a root-mean-squared-error (RMSE) of 1.36 kcal/mol in the predicted binding affinity (corresponding to a factor of 10 in the affinity). This represents a substantial improvement in comparison to using the same model on the entire data set, i.e. without categorization. Variation of empirical weighting coefficient between binding site geometrical classes reflects differences in the binding micro-environments, probably due to varying degrees of shielding of the ligand (and binding pocket residues) from bulk solvent depending on the shape of the binding site.

Moreover, the studied carbohydrate–protein complexes were subjected to MD simulations to investigate the influence of using dynamic averages rather than static values for interaction energy components on the accuracy of free-energy models. Models using a single carbohydrate–protein conformation performed better than models based on averages of multiple frames from MD trajectories. This result, albeit counterintuitive, is in agreement with previous free energy-modeling studies reporting analogous deterioration in prediction accuracy when MD averages were employed. This could be attributed, at least in part, to the need for much longer simulation times to achieve reasonable sampling and convergence in free-energy calculations. On the other hand, the proposed free-energy model (GA1 model) showed only modest prediction accuracy when applied to an external test set of closely related inhibitors of the bacterial FimH adhesin. This suggests that the proposed model is better at predicting free energies when the structural and geometrical differences between tested carbohydrate–protein complexes are large. When the structures of the ligands vary only slightly, however, it is better to use a specifically calibrated model capable of correctly mapping the subtle differences between the complexes, e.g. multidimensional-QSAR model.

Despite the known difficulties in calculating binding affinities for carbohydrate–protein complexes, this study have achieved three important goals. First, a high-quality binding affinity data set for a large and diverse collection carbohydrate–protein complexes has been compiled and thoroughly revised. Second, we proposed a rigorous function for predicting binding affinity from the atomic configuration of carbohydrate–protein complexes. Finally, we propose a scheme for classification of carbohydrate-binding proteins according to the topology and surface exposure of the binding site. Differences between the free-energy models individually calibrated for each topological class reflect the differences in the nature of the local binding micro-environments. Although it might be difficult to fully explain how such differences might affect the shape of the free-energy response surface, the results of this study show how these differences complicate the free-energy prediction problem and demonstrate the usefulness of calibrating free-energy functions individually according to binding-site topology and surface exposure.

I.6. Acknowledgements

This research has been made possible by financial support from the Swiss National Foundation (grant No. 315230_122478) and the Department of Pharmaceutical Sciences, University of Basel, which are both gratefully acknowledged. We would like to thank Mr. Nour Saleh for his assistance in the carbohydrate–protein data set compilation.

I.7. References

Abagyan, R., Totrov, M., & Kuznetsov, D. (1994). ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem., 15*(5), 488–506.

Abagyan, R. A., & Totrov, M. M. (1997). Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol., 268*(3), 678–685.

Agostino, M., Jene, C., Boyle, T., Ramsland, P. a., & Yuriev, E. (2009). Molecular docking of carbohydrate ligands to antibodies: structural validation against crystal structures. *J. Chem. Inf. model., 49*(12), 2749–2760.

Alexacou, K.-M., Hayes, J. M., Tiraidis, C., Zographos, S. E., Leonidas, D. D., Chrysina, E. D., *et al.* (2008). Crystallographic and computational studies on 4-phenyl-N-(beta-D-glucopyranosyl)-1H-1,2,3-triazole-1-acetamide, an inhibitor of glycogen phosphorylase: comparison with alpha-D-glucose, N-acetyl-beta-D-glucopyranosylamine and N-benzoyl-N'-beta-D-glucopyran. *Proteins, 71*(3), 1307–1323.

Allinger, N. L., Rahman, M., & Lii, J. H. (1990). A molecular mechanics force field (MM3) for alcohols and ethers. *J. Am. Chem. Soc., 112*(23), 8293–8307.

Alonso, H., Bliznyuk, A. A., & Gready, J. E. (2006). Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev., 26*(5), 531–568.

Aqvist, J., & Marelius, J. (2001). The linear interaction energy method for predicting ligand binding free energies. *Comb. Chem. High. Throughput Screen., 4*(8), 613–626.

Audie, J., & Swanson, J. (2013). Advances in the prediction of protein–peptide binding affinities: implications for peptide–based drug discovery. *Chem. Biol. Drug. Des., 81*(1), 50–60.

OpenBabel (2012). The Open Babel Package, version 2.3.1 http://openbabel.org.

Baber, J. C., Thompson, D. C., Cross, J. B., & Humblet, C. (2009). GARD: a Generally Applicable Replacement for RMSD. *J. Chem. Inf. Model., 49*(8), 1889–1900.

Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U. S. A., 98*(18), 10037–10041.

Balius, T. E., Mukherjee, S., & Rizzo, R. C. (2011). Implementation and evaluation of a docking-rescoring method using molecular footprint comparisons. *J. Comput. Chem*. *32*(10), 2273–2289

Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics, 26*(9), 1169–1175.

Basma, M., Sundara, S., Calgan, D., Vernali, T., & Woods, R. J. (2001). Solvated ensemble averaging in the calculation of partial atomic charges. *J. Comput. Chem., 22*(11), 1125–1137.

Baudner, B. C., & O'Hagan, D. T. (2010). Bioadhesive delivery systems for mucosal vaccine delivery. *J. Drug. Target., 18*(10), 752–770.

Baum, B., Muley, L., Smolinski, M., Heine, A., Hangauer, D., & Klebe, G. (2010). Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *J. Mol. Biol., 397*(4), 1042–1054.

Bender, A., & Glen, R. C. (2005). A discussion of measures of enrichment in virtual screening: comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model., 45*(5), 1369–1375.

Benson, M. L., Smith, R. D., Khazanov, N. A., Dimcheff, B., Beaver, J., Dresslar, P., *et al.* (2008). Binding MOAD, a high-quality protein-ligand database. *Nucleic Acids Res., 36*(Database issue), D674–678.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., *et al.* (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol., 112*(3), 535–542.

Bissantz, C., Kuhn, B., & Stahl, M. (2010). A medicinal chemist's guide to molecular interactions. *J. Med. Chem., 53*(14), 5061–5084.

Block, P., Sotriffer, C. a., Dramburg, I., & Klebe, G. (2006). AffinDB: a freely accessible database of affinities for protein–ligand complexes from the PDB. *Nucleic Acids Res., 34*(Database issue), D522–526.

Bohne, A., Lang, E., & von der Lieth, C.-W. (1998). W3-SWEET: Carbohydrate Modeling By Internet. *Molecular modeling annual, 4*(1), 33–43.

Boltje, T. J., Buskas, T., & Boons, G. J. (2009). Opportunities and challenges in synthetic oligosaccharide and glycoconjugate research. *Nat. Chem., 1*(8), 611–622.

Boraston, A. B., Bolam, D. N., Gilbert, H. J., & Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J., 382*(Pt 3), 769–781.

Bordner, A. J., & Huber, G. A. (2003). Boundary element solution of the linear Poisson-Boltzmann equation and a multipole method for the rapid calculation of forces on macromolecules in solution. *J. Comput. Chem., 24*(3), 353–367.

Boschitsch, A. H., & Fenley, M. O. (2004). Hybrid boundary element and finite difference method for solving the nonlinear Poisson-Boltzmann equation. *J. Comput. Chem., 25*(7), 935–955.

Bowers, K. J., Chow, E., Xu, H., Dror, R. O., Eastwood, M. P., Gregersen, B. A., *et al.* (2006). Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, Florida, November 11–17

Bradbrook, G. M., Forshaw, J. R., & Pérez, S. (2000). Structure/thermodynamics relationships of lectin-saccharide complexes: the Erythrina corallodendron case. *Eur J. Biochem., 267*(14), 4545–4555.

Brady, G. P., & Stouten, P. F. (2000). Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des., 14*(4), 383–401.

Brandl, M., Weiss, M. S., Jabs, A., Sühnel, J., & Hilgenfeld, R. (2001). C–H...pi-interactions in proteins. *J. Mol. Biol., 307*(1), 357–377.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem., 4*(2), 187–217.

Brown, S. P., Muchmore, S. W., & Hajduk, P. J. (2009). Healthy skepticism: assessing realistic model performance. *Drug Discov. Today, 14*(7–8), 420–427.

Böhm, H. J. (1994). The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des., 8*(3), 243–256.

Böhm, H. J. (1998). Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided Mol. Des., 12*(4), 309–323.

Caffarena, E. R., Grigera, J. R., & Bisch, P. M. (2002). Stochastic molecular dynamics of peanut lectin PNA complex with T-antigen disaccharide. *J. Mol. Graph. Model., 21*(3), 227–240.

Cecchini, M., Kolb, P., Majeux, N., & Caflisch, A. (2004). Automated docking of highly flexible ligands by genetic algorithms: a critical assessment. *J. Comput. Chem., 25*(3), 412–422.

Chang, C. E., Chen, W., & Gilson, M. K. (2007). Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U. S. A., 104*(5), 1534–1539.

Charifson, P. S., Corkery, J. J., Murcko, M. A., & Walters, W. P. (1999). Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem., 42*(25), 5100-5109.

Chen, X., Liu, M., & Gilson, M. K. (2001). BindingDB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screen., 4*(8), 719–725.

Chen, X., Zheng, Y., & Shen, Y. (2006). Voglibose (Basen, AO-128), one of the most important alpha-glucosidase inhibitors. *Curr. Med. Chem., 13*(1), 109–116.

Cheng, T., Li, X., Li, Y., Liu, Z., & Wang, R. (2009). Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model., 49*(4), 1079–1093.

Cho, A. E., Wendel, J. A., Vaidehi, N., Kekenes-Huskey, P. M., Floriano, W. B., Maiti, P. K., *et al.* (2005). The MPSim-Dock hierarchical docking algorithm: application to the eight trypsin inhibitor cocrystals. *J. Comput. Chem., 26*(1), 48–71.

Chothia, C. (1974). Hydrophobic bonding and accessible surface area in proteins. *Nature, 248*(446), 338–339.

Chothia, C., & Janin, J. (1975). Principles of protein-protein recognition. *Nature, 256*(5520), 705–708.

Cipolla, L., Araújo, A. C., Bini, D., Gabrielli, L., Russo, L., & Shaikh, N. (2010). Discovery and design of carbohydrate-based therapeutics. *Expert Opin. Drug Discov., 5*(8), 721–737.

Clark, A., Hirst, M. H., & Jepson, B. A. (2000). Lectin-mediated mucosal delivery of drugs and microparticles. *Adv. Drug Delivery Rev., 43*, 207–223.

Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., *et al.* (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc., 117*(3), 5179–5197.

Cross, S., & Cruciani, G. (2010). Molecular fields in drug discovery: getting old or reaching maturity? *Drug Discov. Today, 15*(1-2), 23–32.

Cummings, R. D. (2009). The repertoire of glycan determinants in the human glycome. *Mol. Biosyst., 5*(10), 1087–1104.

Damm, W., Frontera, A., Tirado–Rives, J., & Jorgensen, W. L. (1997). OPLS all-atom force field for carbohydrates. *J. Comput. Chem., 18*(16), 1955–1970.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N [center-dot] log(N) method for Ewald sums in large systems. *J. Chem. Phys., 98*(12), 10089–10092.

Davis, B. G., & Robinson, M. A. (2002). Drug delivery systems based on sugar-macromolecule conjugates. *Curr. Opin. Drug Discov. Devel., 5*(2), 279–288.

Davis, M. E., & McCammon, J. A. (1991). Dielectric boundary smoothing in finite difference solutions of the poisson equation: An approach to improve accuracy and convergence. *J. Comput. Chem., 12*(7), 909–912.

De Lucca, G. V., Erickson-Viitanen, S., & Lam, P. Y. S. (1997). Cyclic HIV protease inhibitors capable of displacing the active site structural water molecule. *Drug Discovery Today, 2*(1), 6–18.

DeMarco, M. L., & Woods, R. J. (2008). Structural glycobiology: a game of snakes and ladders. *Glycobiology, 18*(6), 426–440.

Deng, W., Breneman, C., & Embrechts, M. J. (2004). Predicting protein-ligand binding affinities using novel geometrical descriptors and machine-learning methods. *J. Chem. Inf. Comput Sci., 44*(2), 699–703.

Deng, Y., & Roux, B. (2009). Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B, 113*(8), 2234-2246.

DePristo, M. A., de Bakker, P. I., & Blundell, T. L. (2004). Heterogeneity and inaccuracy in protein structures solved by X-ray crystallography. *Structure, 12*(5), 831–838.

Desmond. (2011). Desmond Molecular Dynamics System, version 3.0, D. E. Shaw Research, New York, NY, 2011.

Dill, K. A. (1997). Additivity principles in biochemistry. *J. Biol. Chem., 272*(2), 701–704.

Drickamer, K. (1992). Engineering galactose-binding activity into a C-type mannose-binding protein. *Nature, 360*(6400), 183–186.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., & Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res., 34*(Web Server issue), W116–118.

Eid, S., Zalewski, A., Smieško, M., Ernst, B., & Vedani, A. (2013). A Molecular-Modeling Toolbox Aimed at Bridging the Gap between Medicinal Chemistry and Computational Sciences. *Int. J. Mol. Sci., 14*(1), 684–00.

Eisenberg, D., & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature, 319*(6050), 199–203.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997). Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des., 11*(5), 425–445.

Ernst, B., & Magnani, J. L. (2009). From carbohydrate leads to glycomimetic drugs. *Nat. Rev. Drug Discov., 8*(8), 661–677.

Ewing, T. J., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des., 15*(5), 411–428.

Fadda, E., & Woods, R. J. (2010). Molecular simulations of carbohydrates and protein-carbohydrate interactions: motivation, issues and prospects. *Drug Discovery Today, 15*(15-16), 596–609.

Farid, R., Day, T., Friesner, R. A., & Pearlstein, R. A. (2006). New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorg. Med. Chem., 14*(9), 3160–3173.

Fernández-Alonso, M. d. C., Cañada, F. J., Jiménez-Barbero, J., & Cuevas, G. (2005). Molecular Recognition of Saccharides by Proteins. Insights on the Origin of the Carbohydrate−Aromatic Interactions. *J. Am. Chem. Soc., 127*(20), 7379–7386.

Ferrara, P., Gohlke, H., Price, D. J., Klebe, G., & Brooks, C. L. (2004). Assessing scoring functions for protein-ligand interactions. *J. Med. Chem., 47*(12), 3032–3047.

Ferrari, A. M., Wei, B. Q., Costantino, L., & Shoichet, B. K. (2004). Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem., 47*(21), 5076–5084.

Fogolari, F., Brigo, A., & Molinari, H. (2002). The Poisson-Boltzmann equation for biomolecular electrostatics: a tool for structural biology. *J. Mol. Recognit., 15*(6), 377–392.

Foley, B. L., Tessier, M. B., & Woods, R. J. (2012). Carbohydrate force fields. *Wiley Interdisciplinary Reviews: Computational Molecular Science, 2*(4), 652–697.

Friesner, R., Murphy, R., Repasky, M., Frye, L., Greenwood, J., Halgren, T., *et al.* (2006). Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J. Med. Chem., 49*(21), 6177–6196.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., *et al.* (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem., 47*(7), 1739–1749.

Galan, M. C., Benito-Alifonso, D., & Watt, G. M. (2011). Carbohydrate chemistry in drug discovery. *Org. Biomol. Chem., 9*(10), 3598–3610.

Galzitskaya, O. V., Surin, A. K., & Nakamura, H. (2000). Optimal region of average side-chain entropy for fast protein folding. *Protein Sci., 9*(3), 580–586.

Garg, M., & Jain, N. K. (2006). Reduced hematopoietic toxicity, enhanced cellular uptake and altered pharmacokinetics of azidothymidine loaded galactosylated liposomes. *J. Drug Target., 14*(1), 1–11.

Garnier, P., Wang, X. T., Robinson, M. A., van Kasteren, S., Perkins, A. C., Frier, M., *et al.* (2010). Lectin-directed enzyme activated prodrug therapy (LEAPT): Synthesis and evaluation of rhamnose-capped prodrugs. *J. Drug Target., 18*(10), 794–802.

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J., *et al.* (1995). Molecular recognition of the inhibitor AG-1343 by HIV–1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol., 2*(5), 317-324.

Genheden, S., & Ryde, U. (2010). How to obtain statistically converged MM/GBSA results. *J. Comput. Chem., 31*(4), 837–846.

Ghosh, A., Rapp, C. S., & Friesner, R. A. (1998). Generalized Born Model Based on a Surface Integral Formulation. *J. Phys. Chem. B, 102*(52), 10983–10990.

Gilson, M. K., & Zhou, H. X. (2007). Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct., 36*, 21–42.

Glide. (2011). Glide, version 5.7, Schrödinger, LLC, New York, NY.

Gohlke, H., & Case, D. A. (2004). Converging free energy estimates: MM-PB(GB)SA studies on the protein-protein complex Ras-Raf. *J. Comput. Chem., 25*(2), 238–250.

Gohlke, H., Hendlich, M., & Klebe, G. (2000). Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol., 295*(2), 337–356.

Gohlke, H., & Klebe, G. (2001). Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol., 11*(2), 231–235.

Gohlke, H., & Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed. Engl., 41*(15), 2644–2676.

Gonzalez-Outeiriño, J., Kirschner, K. N., Thobhani, S., & Woods, R. J. (2006). Reconciling solvent effects on rotamer populations in carbohydrates — A joint MD and NMR analysis. *Can. J. Chem., 84*(4), 569–579.

Goodford, P. J. (1985). A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem., 28*(7), 849–857.

Grant, J. A., Pickup, B. T., & Nicholls, A. (2001). A smooth permittivity function for Poisson–Boltzmann solvation methods. *J. Comput. Chem., 22*(6), 608–640.

Greenidge, P. A., Kramer, C., Mozziconacci, J. C., & Wolf, R. M. (2013). MM/GBSA Binding Energy Prediction on the PDBbind Data Set: Successes, Failures, and Directions for Further Improvement. *J. Chem. Inf. Model., 53*(1), 201–209.

Grosdidier, S., & Fernández-Recio, J. (2009). Docking and scoring: applications to drug discovery in the interactomics era. *Expert Opin. Drug Discov., 4*(6), 673–686.

Grosdidier, S., Totrov, M., & Fernández-Recio, J. (2009). Computer applications for prediction of protein–protein interactions and rational drug design. *Adv. App. Bioinf. Chem., 2*, 101–123.

Guimaraes, C. R. W. (2011). Direct Comparison of the MM-GB/SA Scoring Procedure and Free-Energy Perturbation Calculations using Carbonic Anhydrase as a Test Case: Strengths and Pitfalls. *J. Chem. Theory Comput.*, 2296–2306.

Guimarães, C. R. (2012). MM-GB/SA rescoring of docking poses. *Methods Mol. Biol., 819*, 255–268.

Guimarães, C. R. W., & Cardozo, M. (2008). MM-GB/SA Rescoring of Docking Poses in Structure-Based Lead Optimization. *J. Chem. Inf. Model., 48*(5), 958–970.

Guimarães, C. R. W., & Mathiowetz, A. M. (2010). Addressing limitations with the MM-GB/SA scoring procedure using the WaterMap method and free energy perturbation calculations. *J. Chem. Inf. Model., 50*(4), 547–559.

Halgren, T. A. (1996a). Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem., 17*(5-6), 490–519.

Halgren, T. A. (1996b). Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem., 17*(5-6), 520–552.

Halgren, T. A. (1996c). Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem., 17*(5-6), 553–586.

Halgren, T. A. (1996d). Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J. Comput. Chem., 17*(5-6), 616–641.

Halgren, T. A. (1999a). MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem., 20*(7), 720–729.

Halgren, T. A. (1999b). MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem., 20*(7), 730–748.

Halgren, T. A., & Damm, W. (2001). Polarizable force fields. *Current Opin. Struct. Biol., 11*(2), 236–242.

Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., *et al.* (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem., 47*(7), 1750–1759.

Halgren, T. A., & Nachbar, R. B. (1996). Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J. Comput. Chem., 17*(5-6), 587–615.

Hansson, T., Marelius, J., & Aqvist, J. (1998). Ligand binding affinity prediction by linear interaction energy methods. *J. Comput. Aided. Mol. Des., 12*(1), 27–35.

Hartenfeller, M., & Schneider, G. (2011). De novo drug design. *Methods Mol. Biol., 672*, 299–323.

Hartshorn, M. J. (2002). AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des., 16*(12), 871–881.

Hawkins, G. D., Cramer, C. J., & Truhlar, D. G. (1995). Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett., 246*(1–2), 122–129.

Head, M. S., Given, J. A., & Gilson, M. K. (1997). "Mining Minima": Direct Computation of Conformational Free Energy. *J. Phys. Chem. A, 101*(8), 1609–1618.

Hill, A. D., & Reilly, P. J. (2008). A Gibbs free energy correlation for automated docking of carbohydrates. *J. Comput. Chem., 29*(7), 1131–1141.

Hoffmann, J., & Spengler, M. (1997). Efficacy of 24-week monotherapy with acarbose, metformin, or placebo in dietary-treated NIDDM patients: the Essen-II Study. *Am. J. Med., 103*(6), 483–490.

Hossain, M. A., & Schneider, H.-J. (1999). Supramolecular Chemistry, Part 85[+] Flexibility, Association Constants, and Salt Effects in Organic Ion Pairs: How Single Bonds Affect Molecular Recognition. *Chem. Eur. J., 5*(4), 1284–1290.

Hou, T., Wang, J., Li, Y., & Wang, W. (2011a). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model., 51*(1), 69–82.

Hou, T., Wang, J., Li, Y., & Wang, W. (2011b). Assessing the performance of the molecular mechanics/Poisson Boltzmann surface area and molecular mechanics/generalized Born surface area methods. II. The accuracy of ranking poses generated from docking. *J. Comput. Chem., 32*(5), 866–877.

Hu, L., Benson, M. L., Smith, R. D., Lerner, M. G., & Carlson, H. A. (2005). Binding MOAD (Mother Of All Databases). *Proteins, 60*(3), 333–340.

Huang, D., & Caflisch, A. (2004). Efficient Evaluation of Binding Free Energy Using Continuum Electrostatics Solvation. *J. Med. Chem., 47*(23), 5791–5797.

Huang, S.-Y., & Zou, X. (2010). Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci., 11*(8), 3016–3034.

Huang, S. Y., Grinter, S. Z., & Zou, X. (2010). Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys., 12*(40), 12899–12908.

Huang, S. Y., & Zou, X. (2006a). An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem., 27*(15), 1866–1875.

Huang, S. Y., & Zou, X. (2006b). An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem., 27*(15), 1876–1882.

Huang, S. Y., & Zou, X. (2010). Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model., 50*(2), 262–273.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem., 28*(6), 1145–1152.

Jackson, R. M., & Sternberg, M. J. (1995). A continuum model for protein-protein interactions: application to the docking problem. *J. Mol. Biol., 250*(2), 258–275.

Jaguar. (2011). Jaguar, version 7.8, Schrödinger, LLC, New York, NY.

Jain, A. N. (2000). Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J. Comput. Aided Mol. Des., 14*(2), 199–213.

Jain, A. N. (2003). Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem., 46*(4), 499–511.

Jain, A. N. (2006). Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci., 7*(5), 407–420.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997). Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol., 267*(3), 727–748.

Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science, 303*(5665), 1813–1818.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys., 79*(2), 926–935.

Jorgensen, W. L., & Madura, J. D. (1985). Temperature and size dependence for Monte Carlo simulations of TIP4P water. *Molec. Phys., 56*(6), 1381–1392.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc., 118*(45), 11225–11236.

Jorgensen, W. L., & Thomas, L. L. (2008). Perspective on Free-Energy Perturbation Calculations for Chemical Equilibria. *J. Chem. Theory Comput., 4*(6), 869–876.

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc., 110*(6), 1657–1666.

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., & Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B, 105*(28), 6474–6487.

Kaplan, J., Korty, B. D., Axelsen, P. H., & Loll, P. J. (2001). The role of sugar residues in molecular recognition by vancomycin. *J. Med. Chem., 44*(11), 1837–1840.

Kerzmann, A., Fuhrmann, J., Kohlbacher, O., & Neumann, D. (2008). BALLDock/SLICK: a new method for protein-carbohydrate docking. *J. Chem. Inf. Model., 48*(8), 1616–1625.

Kerzmann, A., Neumann, D., & Kohlbacher, O. (2006). SLICK--scoring and energy functions for protein-carbohydrate interactions. *J. Chem. Inf. Model., 46*(4), 1635–1642.

Kim, B. Y., Jeong, J. H., Park, K., & Kim, J. D. (2005). Bioadhesive interaction and hypoglycemic effect of insulin-loaded lectin-microparticle conjugates in oral insulin delivery system. *J Control Release, 102*(3), 525–538.

Kim, C. U., Lew, W., Williams, M. A., Liu, H., Zhang, L., Swaminathan, S., *et al.* (1997). Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity. *J. Am. Chem. Soc., 119*(4), 681–690.

Kirschner, K. N., & Woods, R. J. (2001). Solvent interactions determine carbohydrate conformation. *Proc. Natl. Acad. Sci. U. S. A., 98*(19), 10541–10545.

Kirschner, K. N., Yongye, A. B., Tschampel, S. M., González-Outeiriño, J., Daniels, C. R., Foley, B. L., *et al.* (2008). GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem., 29*(4), 622–655.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov., 3*(11), 935–949.

Kohlbacher, O., & Lenhof, H. P. (2000). BALL--rapid software prototyping in computational molecular biology. Biochemicals Algorithms Library. *Bioinformatics, 16*(9), 815-824.

Kony, D., Damm, W., Stoll, S., & Van Gunsteren, W. F. (2002). An improved OPLS-AA force field for carbohydrates. *J. Comput. Chem., 23*(15), 1416–1429.

Koppensteiner, W. A., & Sippl, M. J. (1998). Knowledge-based potentials--back to the roots. *Biochemistry (Mosc.), 63*(3), 247–252.

Kramer, C., Kalliokoski, T., Gedeck, P., & Vulpetti, A. (2012). The Experimental Uncertainty of Heterogeneous Public Ki Data. *J. Med. Chem., 55*(11), 5165–5173.

Kroemer, R. T., Vulpetti, A., McDonald, J. J., Rohrer, D. C., Trosset, J. Y., Giordanetto, F., *et al.* (2004). Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci., 44*(3), 871–881.

Kubik, S. (2012). Carbohydrate recognition: A minimalistic approach to binding. *Nat. Chem., 4*(9), 697–698.

Kuhn, B., Gerber, P., Schulz-Gasch, T., & Stahl, M. (2005). Validation and use of the MM-PBSA approach for drug discovery. *J. Med. Chem., 48*(12), 4040–4048.

Kuhn, B., & Kollman, P. A. (2000). Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *J. Med. Chem., 43*(20), 3786–3791.

Kumar, P. V., Asthana, A., Dutta, T., & Jain, N. K. (2006). Intracellular macrophage uptake of rifampicin loaded mannosylated dendrimers. *J. Drug Target., 14*(8), 546–556.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol., 161*(2), 269–288.

Kuntz, I. D., Chen, K., Sharp, K. A., & Kollman, P. A. (1999). The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A., 96*(18), 9997–10002.

Laederach, A., & Reilly, P. J. (2003). Specific empirical free energy function for automated docking of carbohydrates to proteins. *J. Comput. Chemistry, 24*(14), 1748–1757.

Laederach, A., & Reilly, P. J. (2005). Modeling protein recognition of carbohydrates. *Proteins, 60*(4), 591–597.

Lamb, M. L., Tirado-Rives, J., & Jorgensen, W. L. (1999). Estimation of the binding affinities of FKBP12 inhibitors using a linear response method. *Bioorg. Med. Chem., 7*(5), 851–860.

Lammerts van Bueren, A., & Boraston, A. B. (2004). Binding sub-site dissection of a carbohydrate-binding module reveals the contribution of entropy to oligosaccharide recognition at "non-primary" binding subsites. *J. Mol. Biol., 340*(4), 869–879.

Laskowski, R. A. (1995). SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph., 13*(5), 323–330.

Laughrey, Z. R., Kiehna, S. E., Riemen, A. J., & Waters, M. L. (2008). Carbohydrate-pi interactions: what are they worth? *J. Am. Chem. Soc., 130*(44), 14625–14633.

Lavine, B. K. (1996). Chemometric methods in molecular design. Han van de Waterbeemd (eds.), VCH Publishers, New York, ISBN 3-527-30044-9. *J. Chemom., 10*(3), 269–270.

Lee, M., Lloyd, P., Zhang, X., Schallhorn, J. M., Sugimoto, K., Leach, A. G., *et al.* (2006). Shapes of antibody binding sites: qualitative and quantitative analyses based on a geomorphic classification scheme. *J. Org. Chem., 71*(14), 5082–5092.

Lemieux, R. U. (1989). Rhone-Poulenc Lecture. The origin of the specificity in the recognition of oligosaccharides by proteins. *Chem. Soc. Rev., 18*(0), 347–374.

Levine, D. P. (2006). Vancomycin: a history. *Clin. Infect. Dis., 42 Suppl 1*, S5–12.

Levitt, D. G., & Banaszak, L. J. (1992). POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph., 10*(4), 229–234.

Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., & Friesner, R. A. (2011). The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling. *Proteins, 79*(10), 2794–2812.

Liang, J., Edelsbrunner, H., & Woodward, C. (1998). Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci, 7*(9), 1884–1897.

Lii, J. H., Chen, K. H., Durkin, K. A., & Allinger, N. L. (2003). Alcohols, ethers, carbohydrates, and related compounds. II. The anomeric effect. *J. Comput. Chem., 24*(12), 1473–1489.

Lill, M. A., Vedani, A., & Dobler, M. (2004). Raptor: combining dual-shell representation, induced-fit simulation, and hydrophobicity scoring in receptor modeling: application toward the simulation of structurally diverse ligand sets. *J. Med. Chem., 47*(25), 6174–6186.

Liu, C. C., & Ye, X. S. (2012). Carbohydrate-based cancer vaccines: target cancer with sugar bullets. *Glycoconj. J., 29*(5-6), 259–271.

Liu, H.-Y., Kuntz, I. D., & Zou, X. (2004). Pairwise GB/SA Scoring Function for Structure-based Drug Design. *J. Phys. Chem. B, 108*(17), 5453–5462.

Liu, H. Y., Grinter, S. Z., & Zou, X. (2009). Multiscale generalized born modeling of ligand binding energies for virtual database screening. *J. Phys. Chem. B, 113*(35), 11793–11799.

Liu, H. Y., & Zou, X. (2006). Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson-Boltzmann approach. *J. Phys. Chem. B, 110*(18), 9304–9313.

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res., 35*(Database issue), D198–201.

Lu, J., Zhu, D., Zhang, Z. R., Hai, L., Wu, Y., & Sun, X. (2010). Novel synthetic LPDs consisting of different cholesterol derivatives for gene transfer into hepatocytes. *J. Drug Target., 18*(7), 520–535.

Lyne, P. D., Lamb, M. L., & Saeh, J. C. (2006). Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem., 49*(16), 4805–4808.

MacCallum, R. M., Martin, A. C., & Thornton, J. M. (1996). Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol., 262*(5), 732–745.

Macias, A. T., & Mackerell, A. D. (2005). CH/pi interactions involving aromatic amino acids: refinement of the CHARMM tryptophan force field. *J. Comput. Chem., 26*(14), 1452–1463.

MacKerell, A. D., & Karplus, M. (1991). Importance of attractive van der Waals contribution in empirical energy function models for the heat of vaporization of polar liquids. *The J. Phys. Chem., 95*(26), 10559–10560.

MacRaild, C. A., Daranas, A. H., Bronowska, A., & Homans, S. W. (2007). Global changes in local protein dynamics reduce the entropic cost of carbohydrate binding in the arabinose-binding protein. *J. Mol. Biol., 368*(3), 822–832.

MacroModel. (2011). MacroModel, version 9.9, Schrödinger, LLC, New York, NY.

Maestro. (2011). Maestro, version 9.2, Schrödinger, LLC, New York, NY.

Magnani, J. L., & Ernst, B. (2009). Glycomimetic drugs—a new source of therapeutic opportunities. *Discov. Med., 8*(43), 247–252.

Marelius, J., Ljungberg, K. B., & Aqvist, J. (2001). Sensitivity of an empirical affinity scoring function to changes in receptor-ligand complex conformations. *Eur. J. Pharm. Sci., 14*(1), 87–95.

Marenich, A. V., Olson, R. M., Kelly, C. P., Cramer, C. J., & Truhlar, D. G. (2007). Self-Consistent Reaction Field Model for Aqueous and Nonaqueous Solutions Based on Accurate Polarized Partial Charges. *J. Chem. Theory Comp., 3*(6), 2011–2033.

Margulis, C. J. (2005). Computational study of the dynamics of mannose disaccharides free in solution and bound to the potent anti-HIV virucidal protein cyanovirin. *J. Phys. Chem. B, 109*(8), 3639–3647.

Marsden, P. M., Puvanendrampillai, D., Mitchell, J. B., & Glen, R. C. (2004). Predicting protein-ligand binding affinities: a low scoring game? *Org. Biomol. Chem., 2*(22), 3267–3273.

Martyna, G. J., Tobias, D. J., & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys., 101*(5), 4177–4189.

Maryanoff, B. E., Nortey, S. O., Gardocki, J. F., Shank, R. P., & Dodgson, S. P. (1987). Anticonvulsant O-alkyl sulfamates. 2,3:4,5-Bis-O-(1-methylethylidene)-beta-D-fructopyranose sulfamate and related compounds. *J. Med. Chem., 30*(5), 880–887.

Mehler, E. L., & Solmajer, T. (1991). Electrostatic effects in proteins: comparison of dielectric and charge models. *Protein Eng., 4*(8), 903–910.

Meng, E. C., Shoichet, B. K., & Kuntz, I. D. (1992). Automated docking with grid-based energy evaluation. *J. Comput. Chem., 13*(4), 505–524.

Mishra, N., Tiwari, S., Vaidya, B., Agrawal, G. P., & Vyas, S. P. (2011). Lectin anchored PLGA nanoparticles for oral mucosal immunization against hepatitis B. *J. Drug Target., 19*(1), 67–78.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., Forster, M. J., & Thornton, J. M. (1999). BLEEP—potential of mean force describing protein–ligand interactions: II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem., 20*(11), 1177–1185.

Mitchell, J. B. O., Laskowski, R. A., Alex, A., & Thornton, J. M. (1999). BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential. *J. Comput. Chem., 20*(11), 1165–1176.

Miyazawa, S., & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules, 18*(3), 534–552.

Mooij, W. T., & Verdonk, M. L. (2005). General and targeted statistical potentials for protein-ligand interactions. *Proteins, 61*(2), 272–287.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., *et al.* (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem., 19*(14), 1639–1662.

Moustakas, D. T., Lang, P. T., Pegg, S., Pettersen, E., Kuntz, I. D., Brooijmans, N., *et al.* (2006). Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput. Aided Mol. Des., 20*(10-11), 601–619.

Muegge, I. (2006). PMF scoring revisited. *J. Med. Chem., 49*(20), 5895–5902.

Muegge, I., & Martin, Y. C. (1999). A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem., 42*(5), 791–804.

Muraki, M., Ishimura, M., & Harata, K. (2002). Interactions of wheat-germ agglutinin with GlcNAc beta 1,6Gal sequence. *Biochimica et biophysica acta, 1569*(1-3), 10–20.

Muthana, S., Cao, H., & Chen, X. (2009). Recent progress in chemical and chemoenzymatic synthesis of carbohydrates. *Curr. Opin. Chem. Biol., 13*(5-6), 573–581.

Neumann, D., Lehr, C. M., Lenhof, H. P., & Kohlbacher, O. (2004). Computational modeling of the sugar-lectin interaction. *Adv. Drug. Deliv. Rev., 56*(4), 437–457.

Neves-Petersen, M. T., & Petersen, S. B. (2003). Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules—applications in biotechnology. *Biotechnol. Annu. Rev., 9*, 315–395.

Nicolotti, O., Convertino, M., Leonetti, F., Catto, M., Cellamare, S., & Carotti, A. (2012). Estimation of the binding free energy by Linear Interaction Energy models. *Mini Rev. Med. Chem., 12*(6), 551–561.

Nimje, N., Agarwal, A., Saraogi, G. K., Lariya, N., Rai, G., Agrawal, H., *et al.* (2009). Mannosylated nanoparticulate carriers of rifabutin for alveolar targeting. *J. Drug Target., 17*(10), 777–787.

Nosé, S. (1984). A molecular dynamics method for simulations in the canonical ensemble. *Molec. Phys., 52*(2), 255–268.

Nurisso, a., Kozmon, S., & Imberty, a. (2008). Comparison of docking methods for carbohydrate binding in calcium-dependent lectins and prediction of the carbohydrate binding mode to sea cucumber lectin CEL-III. *Molec. Simul., 34*(4), 469–479.

Núñez, S., Venhorst, J., & Kruse, C. G. (2010). Assessment of a novel scoring method based on solvent accessible surface area descriptors. *J. Chem. Inform. Model., 50*(4), 480–486.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminform., 3*, 33.

Oda, A., Tsuchida, K., Takakura, T., Yamaotsu, N., & Hirono, S. (2006). Comparison of consensus scoring strategies for evaluating computational models of protein-ligand complexes. *J. Chem. Inf. Model., 46*(1), 380–391.

Ooi, T., Oobatake, M., Némethy, G., & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U. S. A., 84*(10), 3086–3090.

Ouerfelli, O., Warren, J. D., Wilson, R. M., & Danishefsky, S. J. (2005). Synthetic carbohydrate-based antitumor vaccines: challenges and opportunities. *Expert Rev. Vaccines, 4*(5), 677–685.

Pace, C. N. (1992). Contribution of the hydrophobic effect to globular protein stability. *J. Mol. Biol., 226*(1), 29–35.

Page, M. I. (1973). The energetics of neighbouring group participation. *Chem. Soc. Rev., 2*(3), 295–323.

Page, M. I. (1977a). Entropie, Bindungsenergie und enzymatische Katalyse. *Angew. Chem., 89*(7), 456–467.

Page, M. I. (1977b). Entropy, Binding Energy, and Enzymic Catalysis. *Angew. Chem. International Edition in English, 16*(7), 449–459.

Page, M. I., & Jencks, W. P. (1971). Entropic Contributions to Rate Accelerations in Enzymic and Intramolecular Reactions and the Chelate Effect. *Proc. Natl. Acad. Sci. U. S. A., 68*(8), 1678–1683.

Parenti, M. D., & Rastelli, G. (2012). Advances and applications of binding affinity prediction methods in drug discovery. *Biotechnol. Adv., 30*(1), 244–250.

Pearlman, D. A. (2005). Evaluating the molecular mechanics poisson-boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *J. Med. Chem., 48*(24), 7796–7807.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham Iii, T. E., DeBolt, S., *et al.* (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Comm., 91*(1–3), 1–41.

Perola, E., & Charifson, P. S. (2004). Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem., 47*(10), 2499–2510.

Perola, E., Walters, W. P., & Charifson, P. S. (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins, 56*(2), 235–249.

Pickett, S. D., & Sternberg, M. J. (1993). Empirical scale of side-chain conformational entropy in protein folding. *J. Mol. Biol., 231*(3), 825–839.

Ponder, J. W., & Case, D. A. (2003). Force fields for protein simulations. *Adv. Protein Chem., 66*, 27–85.

Price, D. J., & Jorgensen, W. L. (2001). Improved convergence of binding affinities with free energy perturbation: application to nonpeptide ligands with pp60src SH2 domain. *J. Comput. Aided Mol. Des., 15*(8), 681–695.

Prime. (2011). Prime, version 3.0, Schrödinger, LLC, New York, NY.

Puhl, W., & Scharf, P. (1997). Intra-articular hyaluronan treatment for osteoarthritis. *Ann. Rheum. Dis., 56*(7), 441.

Qiu, D., Shenkin, P. S., Hollinger, F. P., & Still, W. C. (1997). The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A, 101*(16), 3005–3014.

Quiocho, F. A. (1986). Carbohydrate-binding proteins: tertiary structures and protein-sugar interactions. *Annu. Rev. Biochem., 55*, 287–315.

Ragupathi, G. (1996). Carbohydrate antigens as targets for active specific immunotherapy. *Cancer Immunol. Immunother., 43*(3), 152–157.

Raha, K., & Merz, K. M. (2005). Large-scale validation of a quantum mechanics based scoring function: predicting the binding affinity and the binding mode of a diverse set of protein-ligand complexes. *J. Med. Chem., 48*(14), 4558–4575.

Rajamani, R., & Good, A. C. (2007). Ranking poses in structure-based lead discovery and optimization: current trends in scoring function development. *Curr. Opin. Drug Discov. Devel., 10*(3), 308–315.

Ramkumar, R., & Podder, S. K. (2000). Elucidation of the mechanism of interaction of sheep spleen galectin-1 with splenocytes and its role in cell-matrix adhesion. *J. Mol. Recognit., 13*(5), 299–309.

Ramkumar, R., Surolia, A., & Podder, S. K. (1995). Energetics of carbohydrate binding by a 14 kDa S-type mammalian lectin. *Biochem. J., 308 (Pt 1)*, 237–241.

Ramteke, S., Ganesh, N., Bhattacharya, S., & Jain, N. K. (2008). Triple therapy-based targeted nanoparticles for the treatment of Helicobacter pylori. *J. Drug Target., 16*(9), 694–705.

Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol., 261*(3), 470–489.

Rastelli, G., Del Rio, A., Degliesposti, G., & Sgobba, M. (2010). Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *J. Comput. Chem., 31*(4), 797–810.

Raub, S., Steffen, A., Kämper, A., & Marian, C. M. (2008). AIScore chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. *J. Chem. Inf. Mode.l, 48*(7), 1492–1510.

Rini, J. M. (1995). Lectin structure. *Annu. Rev. Biophys. Biomol. Struct., 24*, 551–577.

Rizzo, R. C., Udier-Blagović, M., Wang, D. P., Watkins, E. K., Kroeger Smith, M. B., Smith, R. H., *et al.* (2002). Prediction of activity for nonnucleoside inhibitors with HIV-1 reverse transcriptase based on Monte Carlo simulations. *J. Med. Chem., 45*(14), 2970–2987.

Rocchia, W., Sridharan, S., Nicholls, A., Alexov, E., Chiabrera, A., & Honig, B. (2002). Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem., 23*(1), 128–137.

Rodinger, T., Howell, P. L., & Pomes, R. (2005). Absolute free energy calculations by thermodynamic integration in four spatial dimensions. *J. Chem. Phys., 123*(3), 34104–34111.

Rosenfeld, R. J., Goodsell, D. S., Musah, R. A., Morris, G. M., Goodin, D. B., & Olson, A. J. (2003). Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. *J. Comput. Aided Mol. Des., 17*(8), 525–536.

Ryckaert, J. P., Ciccotti, G., & Berendsen, H. J. C. (1977). Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys., 23*(3), 327–341.

Sadiq, S. K., Wright, D. W., Kenway, O. a., & Coveney, P. V. (2010). Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model., 50*(5), 890–905.

Salatino, M., Croci, D. O., Bianco, G. A., Ilarregui, J. M., Toscano, M. A., & Rabinovich, G. A. (2008). Galectin-1 as a potential therapeutic target in autoimmune disorders and cancer. *Expert Opin. Biol. Ther., 8*(1), 45–57.

Scheen, A. J. (1998). Clinical efficacy of acarbose in diabetes mellitus: a critical review of controlled trials. *Diabetes Metab, 24*(4), 311–320.

Schulz-Gasch, T., & Stahl, M. (2004). Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discovery Today: Technologies, 1*(3), 231–239.

Schwarz, F. P., Puri, K. D., Bhat, R. G., & Surolia, A. (1993). Thermodynamics of monosaccharide binding to concanavalin A, pea (Pisum sativum) lectin, and lentil (Lens culinaris) lectin. *J. Biol. Chem., 268*(11), 7668–7677.

Scott, L. J., & Spencer, C. M. (2000). Miglitol: a review of its therapeutic potential in type 2 diabetes mellitus. *Drugs, 59*(3), 521–549.

Searle, M., & Williams, D. (1992). The cost of conformational order: entropy changes in molecular associations. *J. Am. Chem. Soc., 114*(27), 10690–10697.

Sharma, A., & Vijayan, M. (2011). Influence of glycosidic linkage on the nature of carbohydrate binding in β-prism I fold lectins: An X-ray and molecular dynamics investigation on banana lectin–carbohydrate complexes. *Glycobiology, 21*(1), 23–33.

Shelley, J., Cholleti, A., Frye, L., Greenwood, J., Timlin, M., & Uchimaya, M. (2007). Epik: a software program for pK(a) prediction and protonation state generation for drug-like molecules. *J. Comput. Aided Mol. Des., 21*(12), 681–691.

Sherman, W., Day, T., Jacobson, M. P., Friesner, R. A., & Farid, R. (2006). Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem., 49*(2), 534–553.

Shimokhina, N., Bronowska, A., & Homans, S. W. (2006). Contribution of Ligand Desolvation to Binding Thermodynamics in a Ligand–Protein Interaction. *Angew. Chem., 118*(38), 6522–6524.

Sigurskjold, B. W., & Bundle, D. R. (1992). Thermodynamics of oligosaccharide binding to a monoclonal antibody specific for a Salmonella O-antigen point to hydrophobic interactions in the binding site. *J. Biol. Chem., 267*(12), 8371–8376.

Sims, P. A., Wong, C. F., & McCammon, J. A. (2003). A computational model of binding thermodynamics: the design of cyclin-dependent kinase 2 inhibitors. *J. Med. Chem., 46*(15), 3314–3325.

Singh, Y., Palombo, M., & Sinko, P. J. (2008). Recent trends in targeted anticancer prodrug and conjugate design. *Curr. Med. Chem., 15*(18), 1802–1826.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol., 213*(4), 859–883.

Smith, R. D., Engdahl, A. L., Dunbar, J. B., & Carlson, H. A. (2012). Biophysical limits of protein-ligand binding. *J. Chem. Inf. Model., 52*(8), 2098–2106.

Smith, R. D., Hu, L., Falkner, J. A., Benson, M. L., Nerothin, J. P., & Carlson, H. A. (2006). Exploring protein-ligand recognition with Binding MOAD. *J. Mol. Graph. Model., 24*(6), 414–425.

Smith, R. H., Jorgensen, W. L., Tirado-Rives, J., Lamb, M. L., Janssen, P. A., Michejda, C. J., *et al.* (1998). Prediction of binding affinities for TIBO inhibitors of HIV-1 reverse transcriptase using Monte Carlo simulations in a linear response method. *J. Med. Chem., 41*(26), 5272–5286.

Søndergaard, R., Garrett, A. E., Carstensen, T., Pollastri, G., Nielsen, J. E. (2009). Structural Artifacts in Protein–Ligand X-ray Structures: Implications for the Development of Docking Scoring Functions. *J. Med. Chem.*, *52*(18), 5673–5684.

Sotriffer, C. A., Sanschagrin, P., Matter, H., & Klebe, G. (2008). SFCscore: scoring functions for affinity prediction of protein-ligand complexes. *Proteins, 73*(2), 395–419.

Srivastava, H. K., & Sastry, G. N. (2012). Molecular dynamics investigation on a series of HIV protease inhibitors: assessing the performance of MM-PBSA and MM-GBSA approaches. *J. Chem. Inf. Model., 52*(11), 3088–3098.

Still, W. C., Tempczyk, A., Hawley, R. C., & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc., 112*(16), 6127–6129.

Stoica, I., Sadiq, S. K., & Coveney, P. V. (2008). Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *J. Am. Chem. Soc., 130*(8), 2639–2648.

Taft, C. A., Da Silva, V. B., & Da Silva, C. H. (2008). Current topics in computer-aided drug design. *J. Pharm. Sci., 97*(3), 1089–1098.

Tanaka, S., & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules, 9*(6), 945–950.

Taroni, C., Jones, S., & Thornton, J. M. (2000). Analysis and prediction of carbohydrate binding sites. *Protein Eng., 13*(2), 89-98.

Tempel, W., Tschampel, S., & Woods, R. J. (2002). The xenograft antigen bound to Griffonia simplicifolia lectin 1-B(4). X-ray crystal structure of the complex and molecular dynamics characterization of the binding site. *J. Biol. Chem., 277*(8), 6615–6621.

Thomas, P. D., & Dill, K. A. (1996a). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U. S. A., 93*(21), 11628–11633.

Thomas, P. D., & Dill, K. A. (1996b). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol., 257*(2), 457–469.

Thompson, D. C., Humblet, C., & Joseph-McCarthy, D. (2008). Investigation of MM-PBSA rescoring of docking poses. *J. Chem. Inf. Model., 48*(5), 1081–1091.

Tjong, H., & Zhou, H. X. (2008). On the Dielectric Boundary in Poisson-Boltzmann Calculations. *J. Chem. Theory Comput., 4*(3), 507–514.

Tvaroska, I., & Carver, J. P. (1998). The anomeric and exo-anomeric effects of a hydroxyl group and the stereochemistry of the hemiacetal linkage. *Carbohyd. Res., 309*(1), 1–9.

van den Berg, J. H., Nuijen, B., Schumacher, T. N., Haanen, J. B., Storm, G., Beijnen, J. H., *et al.* (2010). Synthetic vehicles for DNA vaccination. *J. Drug Target., 18*(1), 1–14.

Vedani, A. (1988). YETI: An interactive molecular mechanics program for small-molecule protein complexes. *J. Comput. Chem., 9*(3), 269–280.

Vedani, A., & Dobler, M. (2002). 5D-QSAR: the key for simulating induced fit? *J. Med. Chem., 45*(11), 2139–2149.

Vedani, A., Dobler, M., & Lill, M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem., 48*(11), 3700–3703.

Vedani, A., & Huhta, D. W. (1990). A new force field for modeling metalloproteins. *J. Am. Chem. Soc., 112*(12), 4759–4767.

Vedani, A., & Zbinden, P. (1998). Quasi-atomistic receptor modeling. A bridge between 3D QSAR and receptor fitting. *Pharm Acta Helv, 73*(1), 11–18.

Velec, H. F., Gohlke, H., & Klebe, G. (2005). DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem., 48*(20), 6296–6303.

Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design--a review. *Curr Top Med Chem, 10*(1), 95–115.

Verma, R. K., Kaur, J., Kumar, K., Yadav, A. B., & Misra, A. (2008). Intracellular time course, pharmacokinetics, and biodistribution of isoniazid and rifabutin following pulmonary delivery of inhalable microparticles to mice. *Antimicrob Agents Chemother, 52*(9), 3195–3201.

von Itzstein, M., Wu, W. Y., Kok, G. B., Pegg, M. S., Dyason, J. C., Jin, B., *et al.* (1993). Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature, 363*(6428), 418–423.

Wang, J., Morin, P., Wang, W., & Kollman, P. A. (2001). Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc., 123*(22), 5221–5230.

Wang, J., Wang, W., Huo, S., Lee, M., & Kollman, P. A. (2001). Solvation Model Based on Weighted Solvent Accessible Surface Area. *J. Phys. Chem. B, 105*(21), 5055–5067.

Wang, R., Fang, X., Lu, Y., & Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem., 47*(12), 2977–2980.

Wang, R., Fang, X., Lu, Y., Yang, C.-Y., & Wang, S. (2005). The PDBbind database: methodologies and updates. *J. Med. Chem., 48*(12), 4111–4119.

Warren, G. L., Andrews, C. W., Capelli, A. M., Clarke, B., LaLonde, J., Lambert, M. H., *et al.* (2006). A critical assessment of docking programs and scoring functions. *J. Med. Chem., 49*(20), 5912–5931.

Warshel, A., & Papazyan, A. (1998). Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol., 8*(2), 211–217.

Webster, D. M., Henry, A. H., & Rees, A. R. (1994). Antibody-antigen interactions. *Curr. Opin. Struct. Biol., 4*(1), 123–129.

Webster, D. M., & Rees, A. R. (1995). Molecular modeling of antibody-combining sites. *Methods Mol. Biol., 51*, 17–49.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., *et al.* (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc., 106*(3), 765–784.

Weinreb, N. J., Barranger, J. A., Charrow, J., Grabowski, G. A., Mankin, H. J., & Mistry, P. (2005). Guidance on the use of miglustat for treating patients with type 1 Gaucher disease. *Am. J. Hematol., 80*(3), 223–229.

Weitz, J. I. (1997). Low-molecular-weight heparins. *N Engl J Med, 337*(10), 688–698.

Wesolowski, S. S., & Jorgensen, W. L. (2002). Estimation of binding affinities for celecoxib analogues with COX-2 via Monte Carlo-extended linear response. *Bioorg. Med. Chem. Lett, 12*(3), 267–270.

Wildman, S. A. (2012). Approaches to Virtual Screening and Screening Library Selection. *Curr Pharm Des (In press)*.

Williams, B. A., Chervenak, M. C., & Toone, E. J. (1992). Energetics of lectin-carbohydrate binding. A microcalorimetric investigation of concanavalin A-oligomannoside complexation. *J. Biol. Chem., 267*(32), 22907–22911.

Williams, D. H., Searle, M. S., Mackay, J. P., Gerhard, U., & Maplestone, R. A. (1993). Toward an estimation of binding constants in aqueous solution: studies of associations of vancomycin group antibiotics. *Proc. Natl. Acad. Sci. U. S. A., 90*(4), 1172–1178.

Williams, D. H., Stephens, E., O'Brien, D. P., & Zhou, M. (2004). Understanding noncovalent interactions: ligand binding energy and catalytic efficiency from ligand-induced reductions in motion within receptors and enzymes. *Angew. Chem. Int. Ed. Engl., 43*(48), 6596–6616.

Woods, R. J., Dwek, R. A., Edge, C. J., & Fraser-Reid, B. (1995). Molecular Mechanical and Molecular Dynamic Simulations of Glycoproteins and Oligosaccharides. 1. GLYCAM_93 Parameter Development. *J. Phys. Chem., 99*(11), 3832–3846.

Xu, J., Deng, Q., Chen, J., Houk, K. N., Bartek, J., Hilvert, D., *et al.* (1999). Evolution of shape complementarity and catalytic efficiency from a primordial antibody template. *Science, 286*(5448), 2345–2348.

Yamazaki, N., Kojima, S., Bovin, N. V., André, S., Gabius, S., & Gabius, H. J. (2000). Endogenous lectins as targets for drug delivery. *Adv. Drug Deliv. Rev., 43*(2-3), 225–244.

Yang, C. Y., Wang, R., & Wang, S. (2006). M-score: a knowledge-based potential scoring function accounting for protein atom mobility. *J. Med. Chem., 49*(20), 5903–5911.

Yang, L. J., & Schnaar, R. L. (2008). Axon regeneration inhibitors. *Neurol. Res., 30*(10), 1047–1052.

Yin, S., Biedermannova, L., Vondrasek, J., & Dokholyan, N. V. (2008). MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J. Chem. Inf. Model., 48*(8), 1656–1662.

Yu, Z., Jacobson, M. P., & Friesner, R. A. (2006). What role do surfaces play in GB models? A new-generation of surface-generalized born model based on a novel gaussian surface for biomolecules. *J. Comput. Chem., 27*(1), 72–89.

Zhang, C., Liu, S., Zhu, Q., & Zhou, Y. (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *J. Med. Chem., 48*(7), 2325–2335.

Zhao, X., Liu, X., Wang, Y., Chen, Z., Kang, L., Zhang, H., *et al.* (2008). An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA. *J. Chem. Inf. Model., 48*(7), 1438–1447.

Zhu, K., Shirts, M. R., & Friesner, R. A. (2007). Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects. *J. Chem. Theory Comput., 3*(6), 2108–2119.

Zhu, X., & Schmidt, R. R. (2009). New principles for glycoside-bond formation. *Angew. Chem. Int. Ed. Engl., 48*(11), 1900–1934.

Zou, X., Yaxiong, & Kuntz, I. D. (1999). Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J. Am. Chem. Soc., 121*(35), 8033–8043.

Zwanzig, R. W. (1954). High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys., 22*(8), 1420–1420.

II. Simulating the binding of Lewis-type ligands to DC-SIGN

II.1. Abstract

Dendritic cells (DCs) have the function of presenting antigens to other processing cells of the immune system, particularly T-cells. DC-SIGN (DC-specific intercellular adhesion molecule-3-grabbing non-integrin) is one of the major receptors on DCs involved in the uptake of pathogens and has gained increasing interest over the last decade as it is crucially involved in infections caused by HIV-1, Ebola virus, *Mycobacterium tuberculosis*, and various other pathogens. High-mannosylated *N*-glycans or L-Fuc-containing trisaccharide motifs such as the Lewis (Le) blood group antigens Le^a and Le^x, which are surface components of these microorganisms, mediate binding to DC-SIGN.

Crystallographic data for DC-SIGN in complex with a Le^x-containing pentasaccharide suggest that the terminal sugar residues, L-Fuc and D-Gal, are predominantly involved in binding. We elucidated the interaction of DC-SIGN with Le^a and Le^x bearing two different aglycones. Binding assays together with STD NMR analysis, molecular modeling and mutagenesis studies revealed distinct binding modes dependent on the nature of the aglycone.

Introduction of phenyl aglycones at the Le trisaccharides offers the establishment of an additional hydrophobic contact with Phe313 in the binding site of DC-SIGN, which entails a switch of the binding mode. Based on this information a new series of DC-SIGN antagonists can be designed.

The work in this section was done in collaboration with *Katharina Mayer, Meike Scharenberg, Katrin Lemme, Arjan Odedra, Brian Cutting, Said Rabbani, Angelo Vedani, and Beat Ernst*.

At the time of writing this thesis, the manuscript was in the final reviewing stage and was ready for submission.

II.2. Introduction

Immature dendritic cells (DCs), found in peripheral tissues throughout the body, play an essential role in triggering the immune response as they are antigen-presenting cells (Hart *et al.*, 1997; Banchereau and Steinman, 1998). DCs recognize and capture a broad variety of pathogens including viruses (Geijtenbeek *et al.*, 2000a), bacteria (Geijtenbeek *et al.*, 2003), and yeasts (Cambi *et al.*, 2003) by pathogen recognition receptors (PRRs). Pathogen uptake by PRRs as well as inflammatory cytokines and chemokines (*e.g.* IL-4) trigger DC differentiation and migration to the lymphoid organs where the mature DCs present pathogenic peptides on the major histocompatibility complex (MHC) to resting T cells.

Dendritic cell-specific intercellular adhesion molecule-3 grabbing nonintegrin (DC-SIGN) is one of the main receptors on DCs for recognition and uptake of pathogens. Since its first discovery by Geijtenbeek *et al.* in 2000 DC-SIGN gained popularity, particularly because a variety of pathogens exploit DC-SIGN to infect their host, including HIV, Ebola virus, and SARS (Geijtenbeek *et al.*, 2000b; Alvarez, 2002; Marzi *et al.*, 2004). The fact that different pathogens have capitalized on this infection strategy makes DC-SIGN an interesting target for a new class of anti-infectives (Anderluh *et al.*, 2012). In a study on the binding and transfer of HIV in human rectal mucosa cells, DC-SIGN$^+$ cells accounted for more than 90% of bound viruses although they represented only 1–5% of the total mucosal mononuclear cells. Furthermore, anti-DC-SIGN antibodies blocked more than 90% of HIV binding (Gurney *et al.*, 2005).

DC-SIGN is a type II transmembrane protein with a C-terminal carbohydrate recognition domain (CRD). It is part of the C-type lectin family, which implies that ligand binding is Ca^{2+}-dependent. The majority of pathogens bind with *N*-linked high-mannose oligosaccharides to DC-SIGN (Feinberg *et al.*, 2001; van Kooyk and Geijtenbeek, 2003), e.g. mannan structures on the gp120 envelope protein of HIV-1 (Geijtenbeek *et al.*, 2000b; Hong *et al.*, 2002). Besides oligomannosides, L-Fuc-containing blood group antigens, such as Lewis^x (Le^x, Galβ(1-4)[Fucα(1-3)]Glc*N*Ac) and Lewis^a (Le^a, Galβ(1-3)[Fucα(1-4)]Glc*N*Ac) that are also commonly found on pathogens, are recognized by DC-SIGN (Van Die *et al.* 2002; 2003, Appelmelk *et al.*, 2003; Naarding *et al.*, 2005). Le^x and Le^a bind to DC-SIGN in the low millimolar range, with Le^a exhibiting a slightly higher binding affinity than Le^x (Timpano *et al.*, 2008; van Liempt *et al.*, 2006). Since pathogens present these rather low-affinity sugar motives in a multimeric form to the DC-SIGN tetramers, high binding avidities are observed (Mitchell *et al.*, 2001; Feinberg *et al.*, 2005).

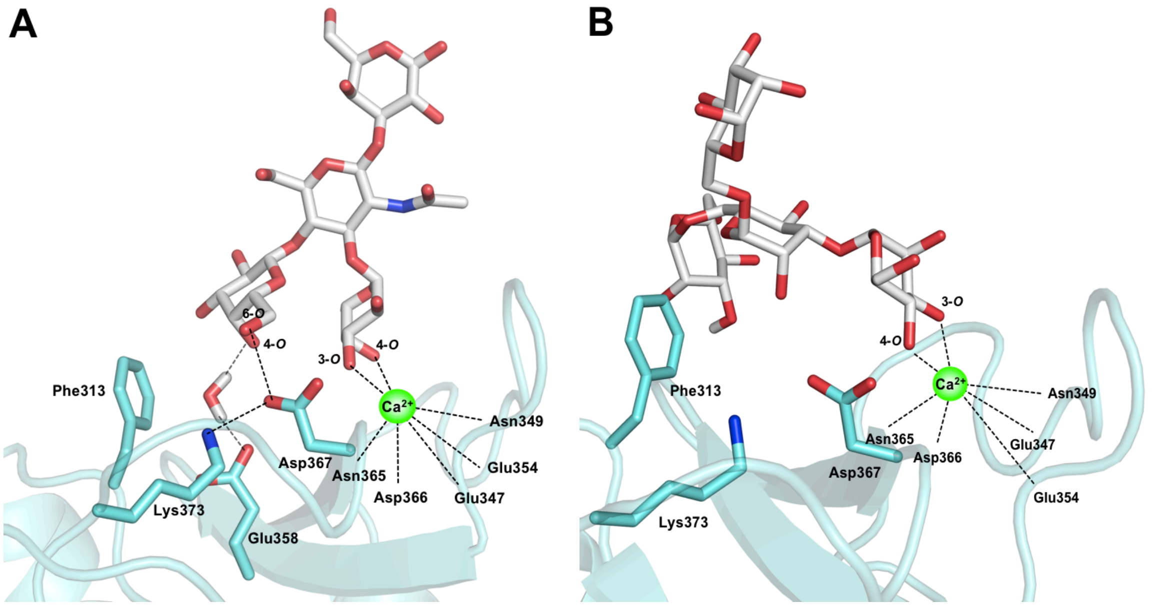


Figure II-1: A) X-ray of LNFP III/CRD of DC-SIGN (PDB 1SL5).[22] The equatorial 3-OH and the axial 4-OH of L-Fuc coordinate the calcium ion. The interaction of 4-OH with Glu358 is mediated by a water molecule. The 6-OH of D-galactose forms a H-bond with Asp367 which on its part is stabilized by an interaction with Lys373. B) X-ray of Man_4 [Manα(1-6)[Manα(1-3)]Manα(1-6)Man]/CRD of DC-SIGN (PDB 1SL4) (Feinberg *et al.*, 2001; Guo *et al.*, 2004). The calcium ion is coordinated by the equatorial 3-OH and the equatorial 4-OH of the terminal α(1-3)-linked D-Man. In addition, a hydrophobic contact of the terminal α(1-6)-linked D-Man further stabilizes the interaction.

Crystallographic data (PDB: 1SL5) (Guo *et al.*, 2004) obtained from the CRD of DC-SIGN co-crystallized with lacto-*N*-fucopentaose III (LNFP III, Galβ(1-4)[Fucα(1-3)]Glc*N*Acβ(1-3)Galβ(1-4)Glc) suggest that the equatorial 3- and axial 4-OH of the L-Fuc moiety coordinate the calcium ion (Figure II-1A). For the 4-OH of the D-Gal moiety a water-bridged H-bond with Glu358 is proposed. In addition, a H-bond of 6-OH of D-Gal to Lys373, bridged by Asp367 is assumed (Guo *et al.*, 2004). For the CRD of DC-SIGN co-crystallized with oligomannosides (Man_4 and Glc*N*Ac_2Man_3) a comparable binding mode was obtained where the equatorial 3- and 4-OH of the α(1-3)-linked D-Man moiety complex the calcium ion (Figure II-1B). In addition, Man_4 (PDB 1SL4) addresses a second binding site lined by Phe313, contributing to selectivity as well as affinity (Feinberg *et al.*, 2001; Guo *et al.*, 2004). Only recently, Bernardi *et al.* took advantage of this additional hydrophobic contact for their design of glycomimetic DC-SIGN antagonists (Obermajer *et al.*, 2010; Andreini *et al.*, 2011).

In our program directed to the identification of high-affinity DC-SIGN antagonists, a large library of carbohydrates and mimetics thereof was screened. One interesting finding was the unexpectedly improved affinity discovered for Le^x and Le^a antigens with aromatic aglycones (→**3**,**4**) compared to the corresponding methyl glycosides (→**1**,**2**). When these derivatives adopt a binding mode similar to LNFP III (Guo *et al.*, 2004), the aglycones should point to the solvent and therefore not contribute to binding. To clarify whether a modified binding mode is responsible for the increased affinity, the binding epitopes of the Le^a and Le^x derivatives **1**-**4** were analyzed by STD NMR and docking studies.

II.3. Results and discussion

II.3.1 Binding Affinities for Lewis Structures

For the determination of the affinities of methyl Le^x (**1**), methyl Le^a (**2**), phenyl Le^x (**3**) and phenyl Le^a (**4**) (Table II-1) a cell-free competitive binding was developed. It is based on the competition of a biotinylated polyacrylamide glycopolymer (Galβ(1-3)[Fucα(1-4)]Glc*N*Acβ-polyacrylamide, Le^a-PAA) and the ligand of interest for the CRD of DC-SIGN. A soluble recombinant protein consisting of the DC-SIGN CRD-Fc (amino acid residues 250-404) was expressed in CHO-K1 cells and purified by affinity chromatography (protein A- and L-Fuc-sepharose column). For the determination of IC_{50} values, a microtiter plate coated with DC-SIGN CRD-Fc was incubated with biotinylated Le^a-PAA polymer conjugated to streptavidin-horseradish peroxidase and the DC-SIGN antagonist in a serial dilution. The assay was performed in duplicates and repeated three times for each compound. To ensure comparability of different ligands, the reference compound L-Fuc was tested in parallel on each individual microtiter plate.

Table II-1: The cell-free competitive binding assay is based on the competition of a biotinylated Le^a-PAA with the antagonist of interest for the CRD of DC-SIGN. The assay was performed in duplicates and repeated three times for each compound. To ensure comparability of different ligands, the reference compound L-Fuc was tested in parallel on each individual microtiter plate. ITC experiments were performed at 25 °C. Thermodynamic parameters were calculated according to the equation $\Delta G = RT\ln K_D = \Delta H - T\Delta S$; n.d. not determined.

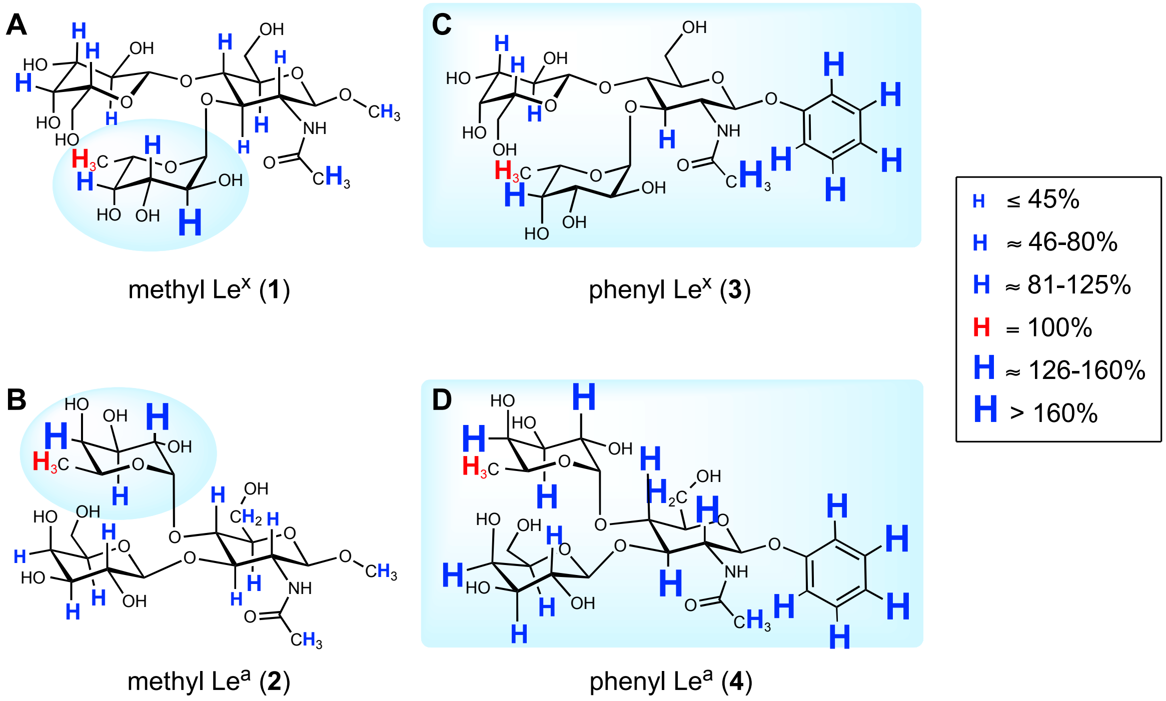
Ligand	**Competitive binding assay, IC_{50}**	**Isothermal titration calorimetry, K_D**
D-Man	9.1 ± 1.3 mM	n.d.
L-Fuc	7.6 ± 2.6 mM	n.d.
Methyl Le^x (methyl Galβ(1-4)[Fucα(1-3)]βGlc*N*Ac) (**1**)	2.3 ± 0.1 mM	n.d.
Methyl Le^a (methyl Galβ(1-3)[Fucα(1-4)]βGlc*N*Ac) (**2**)	2.9 ± 0.5 mM	n.d.
Phenyl Le^x (phenyl Galβ(1-4)[Fucα(1-3)]βGlc*N*Ac) (**3**)	1.2 ± 0.5 mM	n.d.
Phenyl Le^a (phenyl Galβ(1-3)[Fucα(1-4)]βGlc*N*Ac) (**4**)	0.9 ± 0.3 mM	582 ± 40 μM

L-Fuc and D-Man were used as reference compounds showing IC_{50} values of 7.6 mM and 9.1 mM, respectively. These affinities correlate well with published data (Mitchell *et al.* 2001). Phenyl Le^x (**3**) (IC_{50} 1.2 mM) and phenyl Le^a (**4**) (IC_{50} 0.9 mM) showed a two- to threefold increase in affinity compared to corresponding methyl derivatives [IC_{50} 2.3 mM for methyl Le^x (**1**) and 2.9 mM for methyl Le^a (**2**)]. For phenyl Le^a (**4**), the best antagonist in this series, we also performed isothermal titration calorimetry (ITC) experiments. The K_D of 582 μM for phenyl Le^a (**4**) confirms the results of the polymer binding assay with affinity in the high micromolar range. As observed for the majority of carbohydrate–lectin interactions (Toone, 1994; Ambrosi *et al.*, 2005; Dam and Brewer, 2002), the binding is enthalpy driven (ΔH = –28.0 ±2.0 kJ/mol, TΔS = – 9.5 ± 2.1 kJ/mol).

When the Le^x- and Le^a-motifs bind comparable to LNFP III (Guo *et al.*, 2004), only the L-Fuc and D-Gal moiety participate in binding, whereas the D-Glc*N*Ac moiety as well as the aglycone point to the solvent. Therefore, the observed beneficial effect of the aromatic aglycone was unexpected.

II.3.2 Saturation Transfer Difference (STD) NMR Analysis

For the interpretation of the unexpected higher affinities correlated with the phenyl aglycone of antagonists **3** and **4**, the binding epitopes of the Le^a and Le^x derivatives were characterized by STD NMR (Figure II-2A-D), which is particularly suited for epitope mapping of ligand receptor couples with weak interactions (Mayer and Meyer, 1999; Meyer and Peters, 2003; Mayer and Meyer, 2001; Haselhorst *et al.*, 2009). STD NMR experiments are based on spin magnetization transfer from a macromolecule, the protein, to a smaller binding molecule, the ligand. The saturation transfer proceeds through space via dipolar coupling and is therewith dependent on the distance (r^{-6}) of ligand hydrogens to the protein surface.



methyl Le^{x} (**1**)

phenyl Le^{x} (**3**)

methyl Le^{a} (**2**)

phenyl Le^{a} (**4**)

Figure II-2: Binding epitopes of the Lewis antigens **1**-**4** interacting with DC-SIGN CRD-Fc determined by STD NMR. The contribution of each hydrogen to the STD epitope is quantified by forming the ratio of the signal intensities in the STD to those in the reference spectrum. These values are normalized to H-6 of L-Fuc (in red, 100%) to give the percentage epitope. STD values greater than 100% represent proximity to DC-SIGN CRD-Fc closer than that of the H-6 of L-Fuc. The letter size used for the hydrogens expresses the proximity to the protein, *i.e.* the relative amount of saturation transfer. The STD epitope for methyl Le^{x} (**1**) is consistent with recently published data with respect to experimental accuracy (Guzzi *et al.* 2011). Further details regarding the percentage epitope, sample preparation and parameters for the STD NMR measurement are available in the experimental section.

In the STD NMR analysis significantly higher STD values for the aromatic hydrogens (**3** and **4**, Figure II-2C&D) compared to the methyl groups (in **1** and **2**, Figure II-2A&B) were found. This clearly indicates spatial proximity of the aromatic aglycones to DC-SIGN. However, a comparison of the binding epitopes reveals further differences going beyond aglycones. For the D-Glc*N*Ac moieties of methyl Le^{x} (**1**) and Le^{a} (**2**) the maximal STD values for ring hydrogens are smaller than for H-6 of L-Fuc (up to 75%), whereas for the phenyl derivatives **3** and **4** the values reach up to 165%. Especially for phenyl Le^{a} (**4**), and to a lesser extend for phenyl Le^{x} (**3**), high STD values (80-220%) are equally distributed over the entire structure. In contrast, for methyl Le^{x} (**1**), methyl Le^{a} (**2**) high STD values are predominantly located on the L-Fuc moiety. The latter finding corresponds with X-ray data when the Le^{x}-containing LNFP III is co-crystallized with DC-SIGN (Guo *et al.*, 2004), indicating the dominant role of the L-Fuc moiety in these binding epitopes.

II.3.3 Molecular Modeling Studies

Overall, the correlation of increased affinity with the presence of aromatic aglycones as well as the STD NMR data suggest a spatial proximity of the phenyl substituent to DC-SIGN. This is in contrast to the structural information deduced from the co-crystallization of LNFP III with the CRD of DC-SIGN (Guo *et al.*, 2004). For a possible solution of this riddle, docking studies were initiated. The crystal structure 1SL5 (Guo *et al.*, 2004) was used as starting point for the docking studies. The replacement of the internal D-Gal moiety in LNFP III by a methyl aglycone [LNFP III → methyl Le^x (**1**)] is not expected to have a significant influence on its binding mode as indicated by the small STD value of the aglycone in **1** (Figure II-2A). In addition, the proximity of the *N*-acetyl of the D-Glc*N*Ac moiety to Val351 as proposed by the crystal structure (inter-proton distance of 2.5 Å) (Guo *et al.*, 2004) is reflected by the increased STD value.

Automated docking of methyl Le^x (**1**) positions the Le^x subunit in close agreement (RMSD 0.7Å) with its orientation in the crystal structure (Guo *et al.*, 2004) as shown in Figure II-3A. In the docking pose of methyl Le^a (**2**), on the other hand, the D-Glc*N*Ac residue is flipped along its C1-O5 axis, thereby positioning L-Fuc moiety similar to the LNFP III crystal structure. Calcium coordination and H-bond network to L-Fuc are thus maintained (Figure 3B). In this new orientation D-Gal can establish the same characteristic H-bond to Asp367 as well. However, *N*-acetyl group of D-Glc*N*Ac no longer forms a hydrophobic contact with Val351 but with Phe313 instead, with a much longer inter-proton distance of ~ 4.5 Å. This is in good agreement with the lower intensity of the STD NMR signal of the *N*-acetyl group of Glc*N*Ac in methyl Le^a (**2**, Figure II-2B).

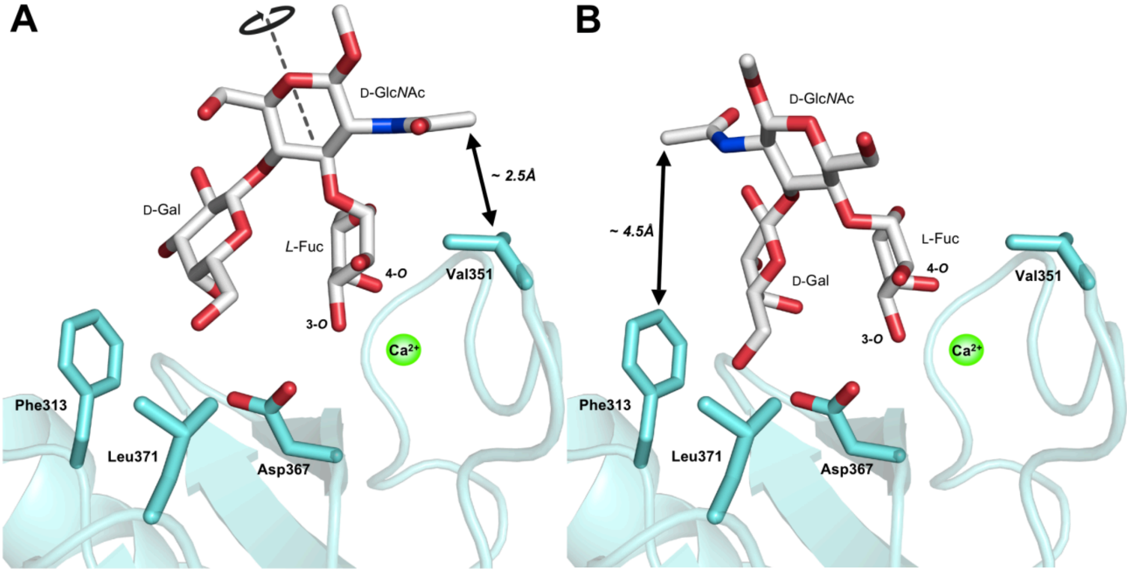


Figure II-3: A) Docking modes of methyl Le^x (**1**) and B) methyl Le^a (**2**). Contacts between the *N*-acetyl groups and closest protein residues are highlighted with double-headed arrows.

A binding mode for phenyl Le^x (**3**) where the Le^x subunit adopts an analogous orientation to LNFP III (Figure II-1A) is inconsistent with the significant saturation transfer observed for the aromatic protons, since the aglycone would point to the solvent with no close contacts to the protein (Figure II-4A). The top-ranked pose from Glide XP induced-fit docking (Glide, version 5.7, Schrödinger, LLC, New York, NY, 2011) presents an alternative pose where the ligand lies "flat" on the receptor and the phenyl aglycone makes a close contact with a hydrophobic cavity formed by the side chains of Phe313 and Leu371 (Figure II-4B). This docking pose perfectly explains the large STD values of the aromatic protons of phenyl Le^x (**3**, Figure II-2C), indicating a close proximity to DC-SIGN.

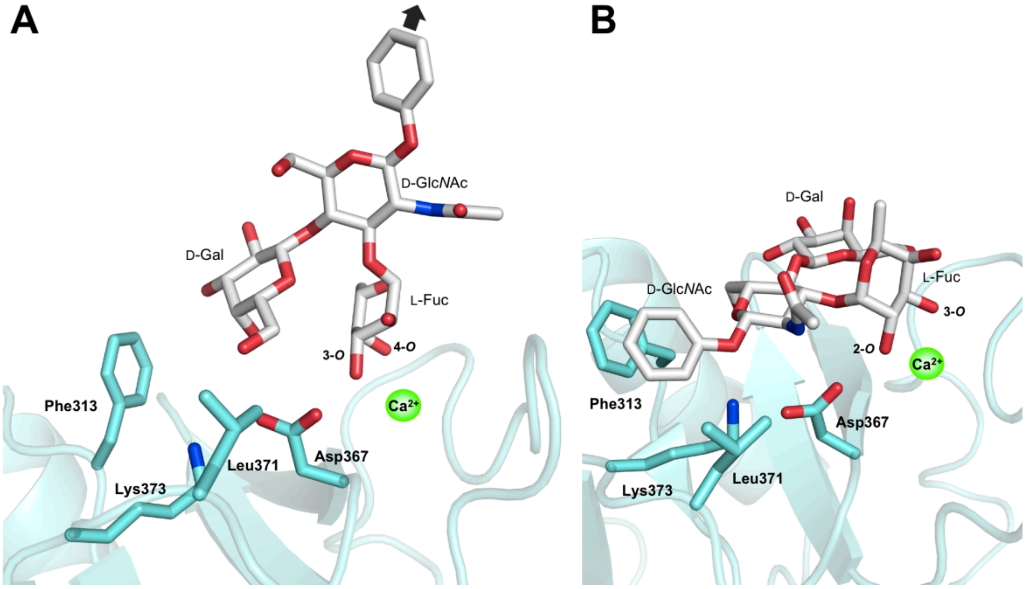


Figure II-4: A) When phenyl Le^x (**3**) binds to DC-SIGN in a manner comparable to methyl Le^x (**2**) (Figure II-3A) and LNFP III (Figure II-1A), the phenyl aglycone points to the solvent (black arrow), not exhibiting an apparent protein contact. B) The induced-fit docking pose for phenyl Le^x (**3**) shows an interaction of the phenyl aglycone with the hydrophobic cleft formed by Phe313 and Leu371, rationalizing the strong aromatic proton signals in STD NMR.

Because of smaller overlaps of the resonances in the ^{1}H-NMR spectrum of phenyl Le^a (**4**), its STD NMR analysis is more detailed. The automated docking pose of phenyl Le^a (Figure II-5) is similar to phenyl Le^x (**3**) where L-Fuc coordinates to Ca^{2+} via the two equatorial hydroxyl groups at the 2- and 3-position. In addition, H-bonds from 2-OH to both Glu354 and Asn365 and between 3-OH and Glu347 are formed. The D-Gal moiety lies close to the primary binding site forming two H-bonds from 6-OH to Glu347 and from 2-OH to Ser360 (not shown). The phenyl aglycone occupies the same hydrophobic pocket (Phe313 and Leu371) as phenyl Le^x (**3**) (Figure II-4B), rationalizing the large STD values for the aromatic protons (Figure II-2D). Moreover, D-Glc*N*Ac also interacts via a H-bond between its 6-OH and Asp367, which in turn bridges this H-bond to Lys373. In the proposed orientation, the D-Glc*N*Ac moiety of phenyl Le^a (**4**) is in closer contact with the receptor compared to methyl

Le^a (2) (Figure II-3B), which explains the observed larger STD values for the D-Glc*N*Ac protons.

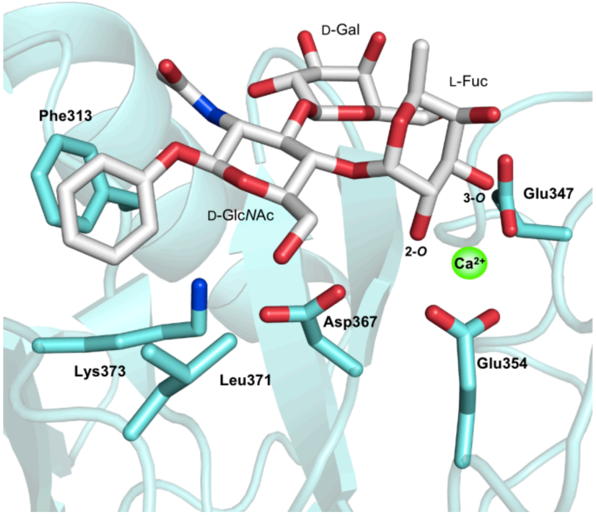


Figure II-5: Binding mode of phenyl Le^a (**4**) to DC-SIGN. Binding of the phenyl aglycone in the hydrophobic cleft formed by Phe313 and Leu371 and proximity of the D-Glc*N*Ac moiety to protein surface coincides with the measured STD NMR values (Figure 2D).

Dynamic stability of this novel binding mode was confirmed by molecular dynamics (MD) simulation. Analysis of MD trajectories revealed that the interactions of phenyl Le^a (**4**) with key residues in the DC-SIGN binding site were maintained throughout the simulation (Figure II-6A). Particularly, the favorable interaction of phenyl Le^a with Phe313 was stable during the simulated time span (Figure II-6B). Despite the alteration in binding mode in comparison to the crystal structure of LNFP III, the Ca^{2+} coordination via Fuc-O2 and Fuc-O3 of phenyl Le^a (**4**) is of comparable stability as reflected by the variation in the distance between Ca^{2+} and its two coordinating oxygens (Figure II-6C&6D). Additionally, throughout the MD simulation all protons of the phenyl in phenyl Le^a exhibited one or more contacts with a proton from a nearby protein residue, consistent with the observed STD signals (Figure II-7).

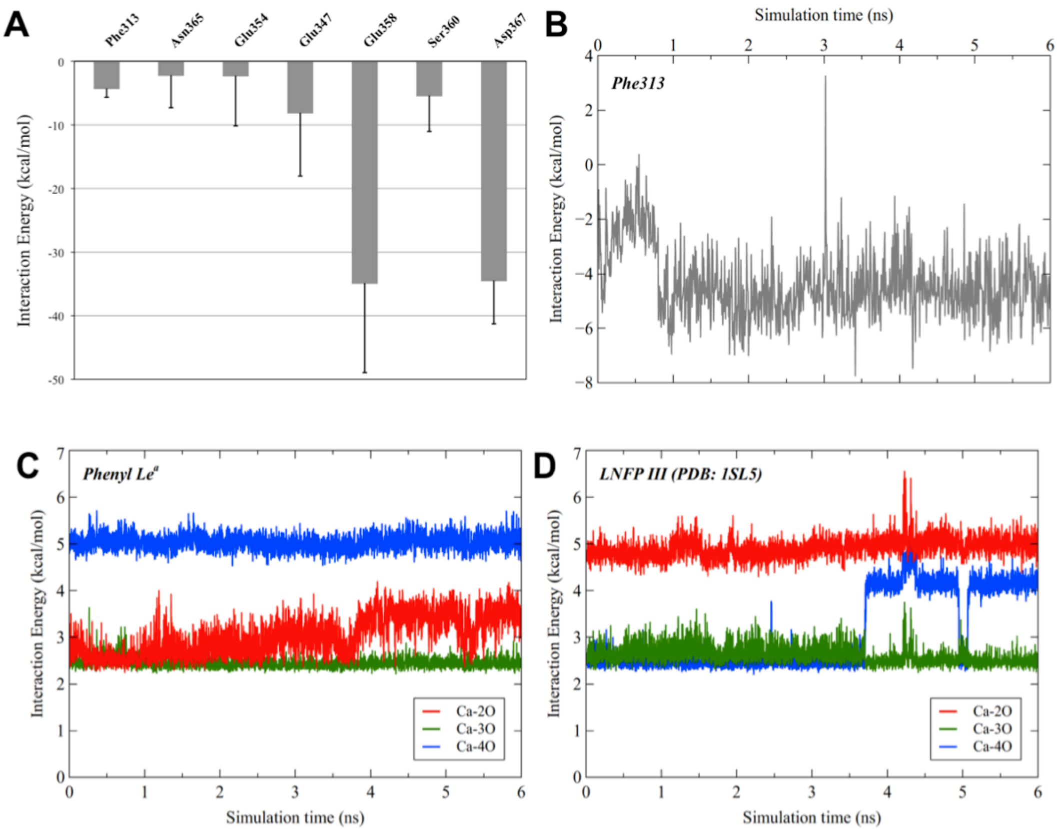


Figure II-6: Results of molecular dynamics simulations. A) average interaction energies between phenyl Le^a (**4**) and some binding site residues during a 6 ns MD simulation, standard deviations are indicated by error bars. B) time evolution of interaction energy between phenyl Le^a and Phe313 residue throughout the MD simulation. C) and D) time evolution of the distances between Ca^{2+} and L-Fuc oxygens (2O, 3O, 4O) along MD simulations starting from LNFP III (in 1SL5 crystal structure) and the docking mode of phenyl Le^a, respectively. The third (non-Ca^{2+}-coordinating) oxygen is shown for comparison.

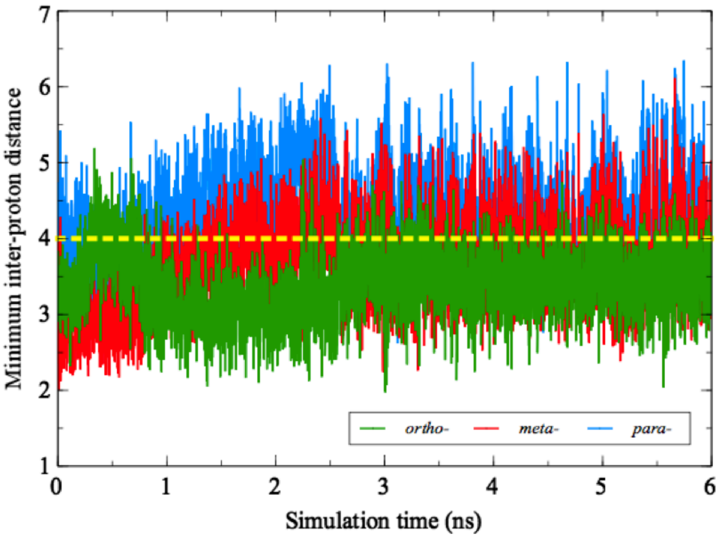


Figure II-7: Time evolution of inter-proton distance between each of the phenyl protons of phenyl Le^a (**4**) and the closest proton of neighboring protein side chains in a 6 ns MD simulation.

II.3.4 Mutagenesis Studies

To further confirm the proposed hydrophobic interaction between Phe313 and the aromatic aglycone, the known DC-SIGN CRD F313A (Guo *et al.*, 2004) was expressed and the binding affinities for methyl Le^a (**2**) and phenyl Le^a (**4**) were determined. Table II-2 summarizes the results of the competitive binding assay with wild type and mutant DC-SIGN CRD. L-Fuc was included as reference compound. The F313A mutation should not have an impact on binding affinity of L-Fuc since the monosaccharide is assumed to bind exclusively in the primary binding site (Guo *et al.*, 2004). However, L-Fuc showed a lower IC$_{50}$ value for the mutant protein (IC$_{50}$ 3.9 mM) than for the wild type (IC$_{50}$ 7.6 mM). This can be explained by the lower affinity of Le^a-PAA for the F313A mutant, reflected by the EC$_{50}$ value (Table II-2). For a better comparison we state relative IC$_{50}$ values (rIC$_{50}$) with L-Fuc as reference (Table II-2).

Table II-2: Results of the competitive binding assay for L-Fuc, methyl Le^a (**2**), and phenyl Le^a (**4**) with wild type and mutant DC-SIGN. The observed differences in the absolute inhibitory potencies between wild type and mutant are due to different binding affinities to Le^a-PAA reflected by a higher EC$_{50}$ value (half maximal effective concentration) in case of the mutant protein. The rIC$_{50}$ values of methyl Le^a (**2**) and phenyl Le^a (**4)** with L-Fuc as reference were determined by dividing the respective IC$_{50}$ values by the IC$_{50}$ of L-Fuc; a value below 1 resembles higher affinity than L-Fuc. Detailed information on protein expression and competitive binding assay is given in the experimental section.

Ligand	**DC-SIGN wild type**	**DC-SIGN F313A mutant**
EC$_{50}$ Le^a-PAA	66.9 ± 0.3 ng/ml	111.2 ± 0.2 ng/ml
rIC$_{50}$ L-Fuc	1	1
rIC$_{50}$ methyl Le^a (**2**)	0.38	0.46
rIC$_{50}$ phenyl Le^a (**4**)	0.12	0.43
Factor of **2** to **4**	3.2	1.1

In Figure II-8, inhibition curves for methyl Le^a (**2**) and phenyl Le^a (**4**) with wild type and mutant DC-SIGN CRD-Fc are shown. Graph A visualizes the aforementioned difference in binding affinity of methyl Le^a (**2**) and phenyl Le^a (**4**) to wild type DC-SIGN (factor 3.2). In contrast, both compounds exhibited near identical binding affinities (factor 1.1, Figure II-8B) for the F131A mutant, which indicates the omission of the beneficial hydrophobic contact of Phe313 with the phenyl aglycone.

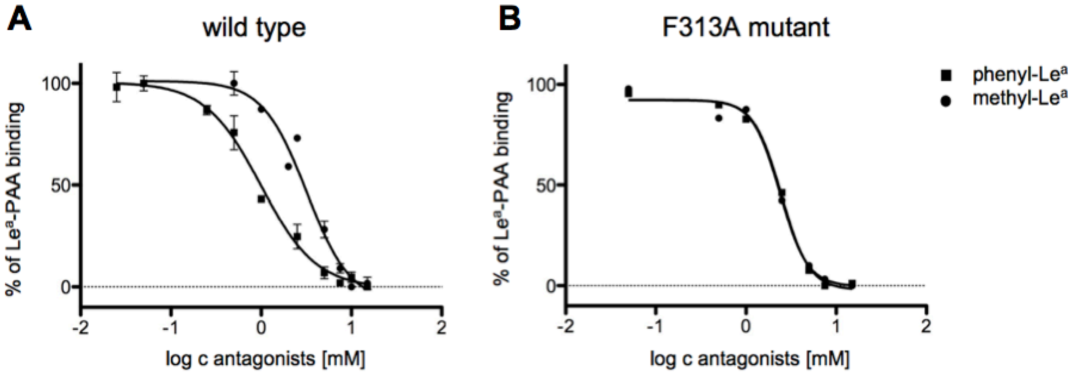


Figure II-8: Inhibition curves for methyl Le^a (**2**) and phenyl Le^a (**4**) obtained from the competitive binding assay, with (A) wild type DC-SIGN and (B) F313A mutant.

II.4. Conclusions

STD NMR spectroscopy and molecular modeling supplemented with a protein mutation study were used to rationalize diverging binding modes of Le^a and Le^x antigens to DC-SIGN induced by the nature of the aglycone. The originally found improved binding affinity of phenyl Le^x (**3**) and phenyl Le^a (**4**) indicated a contribution of the phenyl aglycone to binding, presumably by a hydrophobic contact with the protein. Strong STD NMR values further confirmed this assumption. Docking and MD studies finally revealed a favorable interaction of the phenyl aglycone with a hydrophobic pocket formed by Phe313 and Leu371. With a single-point mutation of the DC-SIGN CRD the proposed interactions of the phenyl aglycone of **4** with Phe313 could be verified.

Here, we report an interesting example, illustrating how flexible binding modes on shallow protein surfaces can be, especially when the starting affinity is low, a situation often present in carbohydrate-lectin interactions. Therefore, improved affinities induced by structural modifications should be carefully analyzed regarding possible reorientations of binding modes. STD NMR experiments (Mayer and Meyer, 1999; Meyer and Peters, 2003) represent an excellent tool for this endeavor.

Based on the new binding mode of phenyl Le^x (**3**) and phenyl Le^a (**4**), the interaction within the hydrophobic pocket formed by Phe313 and Leu371 provides a promising rational for the design of more potent DC-SIGN antagonists. Therewith, our findings support recent approaches from other researches with the objection of using this interaction for the design of glycomimetic DC-SIGN ligands (Andreini, 2011).

Our findings that introduction of a hydrophobic moiety at Lewis trisaccharides induces a switch in the binding mode in order to establish an additional contact with the protein demonstrates the value of this interaction. In fact, recently Bernardi *et al.* made use of this interaction for the design of glycomimetic DC-SIGN antagonists (Andreini, 2011).

II.5. Experimental section

II.5.1 Ligands

Methyl Le^x (**1**) and methyl Le^a (**2**) were purchased from Toronto Research Chemicals Inc. Phenyl Le^x (**3**) was prepared according to Su *et al.* (2009) and phenyl Le^a (**4**) was prepared using standard procedures.

II.5.2 Cloning of DC-SIGN CRD-IgG(Fc)

Plasmids containing the full-length cDNA of DC-SIGN were kindly provided by Daniel A. Mitchell (Glycobiology Institute, Department of Biochemistry, University of Oxford). Standard molecular techniques (Sambrook *et al.*, 1989) were used for the cloning of the carbohydrate recognition domain of DC-SIGN (DC-SIGN CRD; aa residues 250-404, GenBank accession no. M98457). The DC-SIGN CRD encoding insert was amplified by PCR using specific forward and reverse primers containing the restriction sites *EcoRI* and *NcoI* (New England BioLabs, Allschwil, Switzerland), respectively. The insert was ligated into the corresponding cloning site of the pFUSE-hIgG2-Fc2 expression vector (Invivogen, Toulouse, France). The construct was amplified in chemocompetent DH5α *E. coli* (Novagen, Lucerne, Switzerland). After plasmid minipreparation and restriction control, the construct correctness was confirmed by DNA sequencing.

II.5.3 Expression and purification of DC-SIGN CRD-Fc

CHO-K1 cells (American Type Culture Collection No. CCL-61™) were cultivated in Ham's Nutrient Mixture F-12 (Invitrogen, Paisley, UK) supplemented with 2 mM L-glutamate, 10% fetal calf serum (FCS, Invitrogen, Paisley, UK), 100 U/mL penicillin, and 100 µg/mL streptomycin (Sigma-Aldrich, Basel, Switzerland). The cells were cultivated as monolayers in tissue culture flasks (Nunc, Roskilde, Denmark). The CHO-K1 cells were transfected with the DC-SIGN CRD expression vector using the FuGENE® HD transfection reagent (Roche Applied Science, Rotkreuz, Switzerland) following to the instructions of the supplier. Stably transfected CHO-K1 cells were selected by treatment with Zeocin™ (0.5 µg/ml, Invitrogen, Paisley, UK) and single clones were obtained by limiting dilution. For DC-SIGN CRD-Fc production the cells were cultivated as described above and the culture medium, containing the secreted DC-SIGN CRD-Fc chimera was harvested weekly, adjusted to pH 7.6 and sterile filtrated (conditioned medium).

The purification of the recombinant protein was achieved by applying conditioned medium on a protein A-sepharose column (BioVision, Mountain View, CA, USA) attached to a fast protein liquid chromatography apparatus (BioLogic (FPLC) system, BioRad, Reinach BL, Switzerland), which was previously equilibrated with loading buffer I (20 mM Tris/HCl, pH 7.6, 150 mM NaCl, 0.05% (v/v) Tween-20™). The protein was eluted with elution buffer I (0.5 M acetic acid/ammonium acetate, pH 3.4). The collected protein was further purified on a L-Fuc-sepharose column (prepared in house) using loading buffer II (20 mM Tris/HCl, pH 7.8, 0.5 M NaCl, 25 mM $CaCl_2$) and elution buffer II (20 mM Tris/HCl, pH 7.8, 0.5 M NaCl, 2 mM EDTA). For long-term storage, the protein was frozen at –80 °C.

II.5.4 Cloning of the F313A DC-SIGN CRD mutant

The PCR overlap extension method (Ho *et al.* 1989) was used for the substitution of the codon TTC against CGC at cDNA bp 968-970, resulting in the mutation of phenyl alanine 313 to an alanine. In a first step, two overlapping DNA fragments were generated separately, both using wild type DC-SIGN cDNA as template (PCR 1: primer fw: 5` g gaa ttc cat atg gaa cgc ctg tgc cac ccc 3` and primer F313A rv: 5`tcc aga agt aac cgc **gcg** acc tgg atg gga 3`; PCR 2: primer F313A fw: 5`aag tcc cat cca ggt **cgc** gcg gtt act tct 3` and primer rv: 5` cgc gga tcc tta cta cgc agg agg ggg gtt tgg g 3`). The two internal primers contained a mismatch for the site-directed base substitution (bold). In a second step, both overlapping DNA fragments were elongated to the full-length gene, containing the single point mutation. The *NdeI* and *BamHI* (New England BioLabs, Allschwil, Switzerland) treated insert was ligated into the corresponding cloning site of the expression vector pET-3a. After *E.coli* DH5α transformation, plasmid minipreparation, the mutation was confirmed by DNA sequencing. Finally, for protein expression the construct was transformed into BL21 *E.coli* (Novagen, Lucerne, Switzerland).

II.5.5 Expression and purification of F313A DC-SIGN CRD mutant

Protein expression was carried out in TB medium (terrific broth) containing 100 µg/mL ampicillin (Applichem, Darmstadt, Germany). The bacteria were cultured at 37 °C until an OD_{600} of 1.0 was reached. The expression was induced by the addition of isopropyl-β-D-thiogalactoside (IPTG, Applichem, Darmstadt, Germany) at the final concentration of 0.4 mM. The cells were further cultivated for 12 h, prior to harvesting by centrifugation at 4000 rpm for 20 min at 4 °C. For bacterial lysis, the pellet was dissolved in 20 mM Tris-HCl buffer, pH 7.8, 0.5 M NaCl, containing 1 mg/mL lysozym (Sigma, Buchs, Switzerland) and incubated for 30 min at 4 °C under shaking. The inclusion bodies were solubilized by addition of β-mercaptoethanol (0.01 % v/v), urea (8 M), and brief sonication followed by gentle rotation for 30 min at 4 °C. The mixture was centrifuged at 22000 rpm for 1 h at 4°C and the supernatant was diluted by slow addition of the fivefold volume loading buffer II. The mixture was dialyzed against 6 volumes of loading buffer II with 6 buffer exchanges. After dialysis, insoluble precipitate was removed by centrifugation at 22000 rpm for 1 h at 4°C. The protein was purified using a L-Fuc-sepharose column as described above.

Protein purity was confirmed by standard SDS-PAGE analysis (Laemmli 1970) followed by Coomassie Brilliant Blue G-250 staining (Bio-Rad laboratories, Hercules, CA, USA). Protein concentration was determined either by the Bradford method (Bio-Rad laboratories, Hercules, CA, USA) or with HPLC (Bitsch *et al.*, 2003).

II.5.6 Competitive binding assay

Biotinylated Le^a-PAA polymer (20 µL, 1 mg/mL, GlycoTech, Gaithersburg, MD, USA) was mixed with 80 µL assay buffer (20 mM HEPES, 150 mM NaCl, 10 mM $CaCl_{2,}$ pH 7.4), 20 µL FCS and 80 µL streptavidin-horseradish peroxidase-conjugate (500 U/mL, Roche, Mannheim, Germany) and incubated for 2 h at 37 °C. The complex was stable for several weeks when stored at 4 °C.

Flat-bottom 96-well microtiter plates (F96 MaxiSorp, Nunc) were coated with 100 µL/well of a 2.5 µg/mL solution of DC-SIGN CRD-Fc protein in assay buffer overnight at 4 °C in a humidified chamber. The coating solution was discarded and the wells were blocked with 200 µL/well of 3% bovine serum albumin (BSA, Sigma-Aldrich, Buchs, Germany) in assay buffer for 2 h at 4 °C. After three washing steps with assay buffer (150 µL/well), a serial dilution of the test compound (25 µL/well) in assay buffer and streptavidin-peroxidase coupled Le^a-PAA (25 µL/well, 0.25 µg/mL final concentration) were added. Subsequent to an incubation of 3 h at room temperature and 350 rpm the plate was carefully washed four times with 200 µL/well assay buffer. Le^a-PAA binding was detected by addition of 100 µL/well of ABTS-substrate (2,2'-azino-bis-(3-ethylbenzthiazoline-6-sulfonic acid, Invitrogen, Paisley, UK). The colorimetric reaction was allowed to develop for 2 min, then stopped by the addition of 2% aqueous oxalic acid before the optical density (OD) was measured at 415 nm on a microplate-reader (Spectramax 190, Molecular Devices, Ca, USA). The IC_{50}-values were calculated using the Prism software (GraphPad Software, Inc, La Jolla, USA). The IC_{50} (half maximal inhibitory concentration) defines the molar concentration of the test compound that reduces the maximal specific binding of carbohydrate-polymer to DC-SIGN-CRD-Fc by 50%.

For EC_{50} determination (half maximal effective concentration) of the Le^a-PAA, the assay was performed as described above with a serial dilution of Le^a-PAA (0-3 µg/mL) in absence of antagonist.

II.5.7 Isothermal titration Calorimetry

ITC experiments were performed at 298 K and a reference power of 10 µcal/sec under constant stirring speed of 307 rpm using a MicroCal VP-ITC instrument (MicroCal, Northampton, MA). The concentration of DC-SIGN CRD-Fc was determined by HPLC-UV against a standard curve of BSA at 210 nm (Bitsch *et al.*, 2003) after extensive dialysis against 10 mM HEPES, 150 mM NaCl, 10 mM $CaCl_2$, pH 7.4. The ligand was diluted in the dialysat. Injections of 3-5 µl ligand solutions were added from a syringe at an interval of 5 min into the sample cell solution containing DC-SIGN CRD-Fc (cell volume 1.4523 ml). Control experiments were performed, where identical ligand solutions were injected into buffer without protein, and showed insignificant heat of dilution. The experimental data were fitted to a theoretical titration curve (one site binding model) using Origin software (version 7, MicroCal). The quantity $c=Mt(0)/K_D$ with $Mt(0)$ as initial macromolecule concentration, is of importance in titration microcalorimetry (Wiseman *et al.*, 1989). The experiments were performed with c values below 1. The stoichiometry was fixed to 1 (concentration expressed in terms of binding site) to allow reliable determination of K_D and ΔH (Turnbull and Daranas, 2003; Tellinghuisen, 2008). Thermodynamic parameters were calculated from the following equation,

$$\Delta G^o = \Delta H^o - T\Delta S^o = -RTlnK_A = RTlnK_D$$

where ΔG°, ΔH°, and ΔS° are the changes in free energy, enthalpy, and entropy of binding, respectively. T is the absolute temperature, and R = 8.314 J/mol/K.

II.5.8 STD NMR

Experiments were performed on a Bruker 11.7 T spectrometer with an Avance III console at a temperature of 298 K. Shigemi NMR tubes with a sample volume of 250 µL were used for the measurements. Each sample contained 20-30 µM DC-SIGN CRD-Fc (dimer) and 1-2 mM ligand. A d-Tris buffer was used as solvent containing 20 mM d-TRIS (98% Cambridge Isotope Libraries), 4 mM $CaCl_2$ and 150 mM NaCl in D_2O (99.8% Sigma-Aldrich) adjusted to a pH of 8.1 with HCl.

Using a pulse sequence modified from Mayer and Meyer (1999) allows simultaneous saturation of the protein at two frequencies, which leads to a more intense STD epitope. The cosine modulated E-Burp-1 pulse (Geen *et al.* 1989) for the on-resonance spectrum was centered at 1555 Hz and resulted in two sidebands at 0 and 3110 Hz with a power of 53 dB (Cutting *et al.*, 2007). The duration of each of the 40 E-Burp-1 pulses used to saturate the protein was 50 ms with a 1 ms recovery between the pulses.

Off-resonance excitation was set to 26000 Hz. STD NMR experiments were performed applying a Watergate solvent suppression. Specific parameters were determined via preliminary experiments including negative control experiments with only ligand-containing sample to avoid artifacts from direct excitation. Scaling each STD signal on an off-resonance reference spectrum resulted in a relative binding epitope (Mayer and Meyer, 2001). Ligand resonances were assigned by using 2D NMR and 1D selective TOCSY experiments. Not all protons could be assigned doubtlessly, due to solvent suppression and partial signal overlap.

Detailed conditions: STD NMR of methyl-bearing compounds: 2 mM ligand with 20 µM DC-SIGN CRD-Fc, STD NMR of phenyl Le^a: 1 mM ligand with 20 µM DC-SIGN CRD-Fc, STD NMR of phenyl Le^x: 1 mM with 30 μM DC-SIGN CRD-Fc; number of scans was typically 14k for on-resonance spectra and 512 for off-resonance spectra.

Experiments with different saturation times were performed for phenyl Le^a. These data indicate an overall consistent epitope at either saturation times of 0.7, 1, 2, and 3 s and exclude misinterpretation due to T1 bias for different proton species.

II.5.9 Molecular Modeling

All ligands were manually built using Maestro (Maestro, version 9.2, Schrödinger, LLC, New York, NY, 2011), and optimized using standard procedures. Model for DC-SIGN in complex with LNFP III was downloaded from the Protein Data Bank (code: 1SL5). Hydrogens were added and water molecules were removed using Maestro Protein Preparation Wizard. Partial charges were calculated from OPLS2005 force field while protonation states and oxidation states for metals were assigned by Epik (Shelley *et al.*, 2007). Orientation of added hydrogens was sampled for optimal H-bond formation and the model was then refined by minimization within RMSD of 0.3Å.

GlideXP (version 5.7, Schrödinger, LLC, New York, NY, 2011) was used for docking of novel ligands to DC-SIGN. To account for the possibility of side chain re-organization upon ligand binding the Induced-Fit Docking (IFD) methodology was employed (Sherman *et al.*, 2006). The binding site was defined to include residues within 5 Å radius around the co-crystallized ligand LNFP III in the prepared complex. In the initial stages of IFD protocol amino acids within 5 Å radius around any found pose were considered as flexible, and their side chain conformations were optimized. Up to 50 poses were retained for each calculation within an energy window of 40 kcal/mol to allow for larger diversity in output poses. Prioritization was done by Standard Precision (SP) (Friesner *et al.*, 2004) scoring function in the initial soft-docking stage followed by more rigorous Extra Precision (XP) (Friesner *et al.*, 2006) scoring in the redocking stage. Output poses were then visually inspected for agreement with STD NMR experiment, and those showing considerable discrepancy were disregarded.

Stability of the proposed modes was assessed using molecular dynamics. Docking poses and crystal structure (PDB 1SL5) were used as a starting point for 6 ns MD simulations using Desmond package from D. E. Shaw Research lab (Bowers *et al.*, 2006). The protein–ligand complex was soaked in an orthorhombic TIP3P water box extending 10 Å away from the complex. Counter-ions were added to make it neutral and 0.15 M sodium and chloride ions were added to approximate physiological conditions. The complex was then minimized to a convergence threshold of 1.0 kcal/mol/Å. MD experiments were carried out using the OPLS2005 force field and the NPT ensemble (constant number of particles, pressure and temperature) at 300 K with periodic boundary conditions. Default parameters were used and snapshots recorded every 1.2 ps. Output files were analyzed using component-interactions script in Maestro to compute interaction energies between the ligand and individual amino acids defining the binding site as well as the conserved calcium along the MD simulations. Interaction energies were computed as the sum of OPLS2005 Van der Waals and electrostatic terms.

II.6. Acknowledgements

The authors gratefully acknowledge the financial support by the Swiss National Science Foundation (grant no.) and by GlycoMimetics Inc., Gaithersburg, MD, USA.

II.7. References

Alvarez, C. P. (2002). *J. Virol.*, *76*, 6841–6844.

Ambrosi, M.; Cameron, N. R.; Davis, B. G. (2005). *Org. Biomol. Chem.*, *3*, 1593–1608.

Anderluh, M.; Jug, G.; Svajger, U.; Obermajer, N. (2012). *Curr. Med. Chem.*, *19*, 992–1007.

Andreini, M.; Doknic, D.; Sutkeviciute, I.; Reina, J. J.; Duan, J.; Chabrol, E.; Thepaut, M.; Moroni, E.; Doro, F.; Belvisi, L.; Weiser, J.; Rojo, J.; Fieschi, F.; Bernardi, A. (2011). *Org. Biomol. Chem.*, *9*, 5778–5786.

Appelmelk, B. J.; van Die, I.; van Vliet, S.; Vandenbroucke-Grauls, C.; Geijtenbeek, T.; van Kooyk, Y. (2003). *J. Immunol.*, *170*, 1635–1639.

Banchereau, J.; Steinman, R. M. (1998). *Nature*, *392*, 245–252.

Bitsch, F.; Aichholz, R.; Kallen, J.; Geisse, S.; Fournier, B.; Schlaeppi, J.-M. (2003). *Anal. Biochem.*, *323*, 139–149.

Cambi, A.; Gijzen, K.; de Vries, J. M.; Torensma, R.; Joosten, B.; Adema, G. J.; Netea, M. G.; Kullberg, B. J.; Romani, L.; Figdor, C. G. (2003). *Eur. J. Immunol.*, *33*, 532–538.

Cutting, B.; Shelke, S. V.; Dragic, Z.; Wagner, B.; Gathje, H.; Kelm, S.; Ernst, B. (2007). *Magn. Reson. Chem.*, *45*, 720–724.

Dam, T. K.; Brewer, C. F. (2002). *Chem. Rev.*, *102*, 387–429.

Feinberg, H.; Mitchell, D. A.; Drickamer, K.; Weis, W. I. (2001). *Science*, *294*, 2163–2166.

Feinberg, H.; Guo, Y.; Mitchell, D. A.; Drickamer, K.; Weis, W. I. (2005). *J. Biol. Chem.*, *280*, 1327–1335.

Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. (2004). *J. Med. Chem.*, *47*, 1739–1749.

Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. (2006). *J. Med. Chem.*, *49*, 6177–6196.

Geen, H.; Wimperis, S.; Freeman, R. (1989). *J. of Magn. Reson.*, *85*, 620–627.

Geijtenbeek, T. B. H.; Torensma, R.; van Vliet, S. J.; van Duijnhoven, G. C. F.; Adema, G. J.; van Kooyk, Y.; Figdor, C. G. (2000a). *Cell*, *100*, 575–585.

Geijtenbeek, T. B. H.; Kwon, D. S.; Torensma, R. ; van Vliet, S. J.; van Duijnhoven, G. C. F.; Middel, J.; Cornelissen, I.; Nottet, H.; KewalRamani, V. N.; Littman, D. R.; Figdor, C. G.; van Kooyk, Y. (2000b). *Cell*, *100*, 587–597.

Geijtenbeek, T. B. H.; van Vliet, S. J.; Koppel, E. A.; Sanchez-Hernandez, M.; Vandenbroucke-Grauls, C.; Appelmelk, B.; van Kooyk, Y. (2003). *J. Exp. Med.*, *197*, 7–17.

Guo, Y.; Feinberg, H.; Conroy, E.; Mitchell, D.; Alvarez, R.; Blixt, O.; Taylor, M.; Weis, W. I.; Drickamer, K. (2004). *Nat. Struct. Mol. Biol.*, *11*, 591–598.

Gurney, K. B. ; Elliott, J.; Nassanian, H.; Song, C.; Soilleux, E.; McGowan, I.; Anton, P. A.; Lee, B. (2005). *J. Virol.*, *79*, 5762–5773.

Guzzi, C.; Angulo, J.; Doro, F.; Reina, J. J.; Thepaut, M.; Fieschi, F.; Bernardi, A.; Rojo, J.; Nieto, P. M. (2011). *Org. Biomol. Chem.*, *9*, 7705–7712.

Hart, D. N. J.; Clark, G. J.; Dekker, J. W.; Fearnley, D. B.; Kato, M.; Hock, B. D., McLellan, A. D.; Neil, T.; Sorg, R. V., Sorg, U.; Summers, K. L.; Vuckovic, S. (1997). *Dendritic Cells in Fundamental and Clinical Immunology, Vol 3*, *417*, 439–442.

Haselhorst, T.; Lamerz, A.-C.; von Itzstein, M. (2009). *Methods Mol Biol*, *534*, 375–387.

Ho, S. N.; Hunt, H. D.; Horton, R. M.; Pullen, J. K.; Bell, M. P.; McKean, D. J.; Pease, L. R. (1989). *FASEB J.*, *3*, A519.

Hong, P. W.; Flummerfelt, K. B.; de Parseval, A.; Gurney, K.; Elder, J. H.; Lee, B. (2002). *J. Virol.*, *76*, 12855–12865.

Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters, in Proceedings of the 2006 ACM/IEEE conference on Supercomputing, ACM, Tampa, Florida, p. 84.

Laemmli, U. K. (1970). *Nature*, *227*, 680–685.

Marzi, A.; Gramberg, T.; Simmons, G.; Moller, P.; Rennekamp, A. J.; Krumbiegel, M.; Geier, M.; Eisemann, J.; Turza, N.; Saunier, B.; Steinkasserer, A.; Becker, S.; Bates, P.; Hofmann, H.; Pohlmann, S. (2004). *J. Virol.*, *78*, 12090–12095.

Mayer, M. ; Meyer, B. (1999). *Angew. Chem., Int. Ed.*, *38*, 1784–1788.

Mayer, M.; Meyer, B. (2001). *J. Am. Chem. Soc.*, *123*(25), 6108–6117.

Meyer, B.; Peters, T. (2003). *Angew. Chem., Int. Ed.*, *42*, 864–890.

Mitchell, D. A.; Fadden, A. J.; Drickamer, K. (2001). *J. Biol. Chem.*, *276*, 28939–28945.

Naarding, M. A.; Ludwig, I. S.; Groot, F.; Berkhout, B.; Geijtenbeek, T. B. H.; Pollakis, G.; Paxton, W. A. (2005). *J. Clin. Invest.*, *115*, 3256–3264.

Obermajer, N.; Sattin, S.; Colombo, C.; Bruno, M.; Švajger, U.; Anderluh, M.; Bernardi, A. (2010). *Mol. Diversity*, 1–14.

Sambrook, J.; Fritsch, E. F.; Maniatis, T. (1989). Cold Spring Harbor Laboratory Press, NY.

Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. (2007). *J. Comput.-Aided Mol. Des.*, *21*, 681–691.

Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. (2006). *J. Med. Chem.*, *49*, 534–553.

Su, Z.; Wagner, B.; Cocinero, E. J.; Ernst, B.; Simons, J. P. (2009). *Chem. Phys. Lett.*, *477*, 365–368.

Tellinghuisen, J. (2008). *Anal. Biochem.*, *373*, 395–397.

Timpano, G.; Tabarani, G.; Anderluh, M.; Invernizzi, D.; Vasile, F.; Potenza, D.; Nieto, P. M.; Rojo, J.; Fieschi, F.; Bernardi, A. (2008). *ChemBioChem*, *9*, 1921–1930.

Toone, E. J. (1994). *Curr. Opin. Struct. Biol.*, *4*, 719–728.

Turnbull, W. B.; Daranas, A. H. (2003). *J. Am. Chem. Soc.*, *125*, 14859–14866.

Van Die, I.; Van Vliet, S. J.; Schiphorst, W.; Bank, C. M. C.; Appelmelk, B.; Nyame, A. K.; Cummings, R. D.; Geijtenbeek, T. B. H.; Van Kooyk, Y. (2002). *Glycobiology*, *12*, 3.

Van Die, I.; van Vliet, S. J.; Nyame, A. K.; Cummings, R. D.; Bank, C. M. C.; Appelmelk, B.; Geijtenbeek, T. B. H.; van Kooyk, Y. (2003). *Glycobiology*, *13*, 471–478.

van Kooyk, Y.; Geijtenbeek, T. B. H. (2003). *Nat. Rev. Immunol.*, *3*, 697–709.

van Liempt, E.; Bank, C. M. C.; Mehta, P.; Garcia-Vallejo, J. J.; Kawar, Z. S.; Geyer, R.; Alvarez, R. A.; Cummings, R. D.; van Kooyk, Y.; van Die, I. (2006). *FEBS Lett.*, *580*, 6123–6131.

Wiseman, T.; Williston, S.; Brandts, J. F.; Lin, L. N. (1989). *Anal. Biochem.*, *179*, 131–137.

III. Developing a molecular modeling toolbox for medicinal chemists

Abstract

In the current era of high-throughput drug discovery and development, molecular modeling has become an indispensable tool for identifying, optimizing and prioritizing small-molecule drug candidates. The required background in computational chemistry and the knowledge of how to handle the complex underlying protocols, however, might keep medicinal chemists from routinely using *in silico* technologies. Our objective is to encourage those researchers to exploit existing modeling technologies more frequently through easy-to-use graphical user interfaces. In this account, we present two innovative tools (which we are prepared to share with academic institutions) facilitating computational tasks commonly utilized in drug discovery and development: (1) the *VirtualDesignLab* estimates the binding affinity of small molecules by simulating and quantifying their binding to the three-dimensional structure of a target protein; and (2) the *MD Client* launches molecular dynamics simulations aimed at exploring the time-dependent stability of ligand–protein complexes and provides residue-based interaction energies. This allows medicinal chemists to identify sites of potential improvement in their candidate molecule. As a case study, we present the application of our tools towards the design of novel antagonists for the FimH adhesin.

This work has been published: Eid S., Zalewski A., Smieško M., Ernst B., Vedani A. *Int. J. Mol. Sci.* **2013**; *14*: 684-700.

A copy of the article is attached in Appendix 3.

A copy of the *MD Client Quick Start Guide* is attached in Appendix 4.

Appendices

Appendix 1: List of carbohydrate–protein complexes

Flags

α Overlapping anomers in crystal structure; the α-anomer was used

β Overlapping anomers in crystal structure; the β-anomer was used

X Complex was excluded from the study; reason indicated by one of the flags (M, O, R, T) or explained in the comments column

M Multiple ligand copies in crystal structure; either the protein is polyvalent or the ligand is several copies of the ligand were resolved which differ significantly in orientation and/or conformation

O High molecular weight ligand (> 1000 Dalton)

T Missing important atoms

R Redundant entry

H The ligand HET ID used by Protein Data Bank was overwritten; either because it was incorrect or because the ligand consists of several residue that were identified individually in the PDB (Refer to the Comments column for the HET ID's of the constituent residues)

P The ligand contains phosphate group

W Warning; check the Comments Column

Fields

PDB ID 4-letter unique PDB accession code for the complex

HET ID 3-letter ligand unique identifier, codes starting with an underscore, e.g. _NAL, are non-standard codes

Affinity Type of experimental binding affinity (K_a, K_d, K_i, or IC_{50})

$-\Delta G_{exp}$ Experimental Gibb's free energy of binding in kcal/mol, calculated from binding affinity using the thermodynamic master equation: $\Delta G_{exp} = RT\mathit{ln}K$

Res. Resolution of the crystal structure in Å

Mol. Wt. Molecular weight of the ligand

Formula Molecular formula of the ligand

References Original publication(s) reporting the experimental binding affinity

Comments Notes on the structure preparation of the ligand–protein complex, reasons for rejecting the complex, warnings; etc.

Flags	PDB ID	HET ID	Affinity	– ΔG_{exp}	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
XM	1A0T	SUC	Kd	1.8	2.4	Sucrose-Specific Porin	Sucrose	342.3	$C_{12}H_{22}O_{11}$	Nat Struct Biol. 1998, 5(1): 37-46 Mol. Microbiol. 1995, 17: 757−767	Two non-overlapping sucrose binding sites, and two bound sucrose moieties with 0.7 and 0.8 occupanices	P
	1A8I	GLS	Kd	7.5	1.78	Glycogen Phosphorylase B	Glucopyranose Spirohydantoin	248.19	$C_8H_{12}N_2O_7$	Bioorg Med Chem Lett. 2008, 18: 5680-3 Bioorg Med Chem Lett. 1999, 9: 1385-90 Protein Science 1998, 7: 915-927 Tetrahedron Letters 1995, 36(12): 2145-2148	PDB: 1GGN has the same ligand–protein combination, discarded due to lower resolution (2.39 Å)	A
α	1ABF	FCA	Kd	7.4	1.9	L-Arabinose-Binding Protein	Alpha-D-Fucose	164.16	$C_6H_{12}O_5$	Nature. 1989, 340(6232): 404-7 J Biol Chem 1983, 258 (22): 13665-72		A
	1ADD	1DA	Ki	9.2	2.4	Adenosine Deaminase	1-Deaza-Adenosine	266.26	$C_{11}H_{14}N_4O_4$	Biochemistry 1992, 31(1): 39-48 Biochemistry 1993, 32(7):1689-94 J Med Chem 1997, 40: 3336-3345		A
H	1AF6	SUC	Kd	2.5	2.4	Maltoporin	Sucrose	342.3	$C_{12}H_{22}O_{11}$	J Memb Biol 1987, 100: 21-29 J Mol Biol 1997, 272(1): 56-63	GLC+FRU	A
	1AJ6	NOV	Kd	8.1	2.3	Gyrase	Novobiocin	612.63	$C_{31}H_{36}N_2O_{11}$	Biochemistry 1997, 36(32): 9663-73		
H	1ANF	MAL	Kd	7.4	1.67	Maltodextrin-Binding Protein	Maltose	342.3	$C_{12}H_{22}O_{11}$	Structure 1997, 5(8): 997-1015	GLC+GLC	
β	1APB	FCA	Kd	7.9	1.76	L-Arabinose-Binding Protein	Alpha-D-Fucose	164.16	$C_6H_{12}O_5$	J Biol Chem 1990, 265(27):16592-16603		
	1AX0	A2G	Ka	4.3	1.9	Lectin	N-Acetyl-2-Deoxy-2-Amino-Galactose	221.21	$C_8H_{15}NO_6$	J Biol Chem 1996, 271: 17697-17703 J Mol Biol 1998, 277: 917-932		
H	1AX1	LAT	Ka	4.5	1.95	Lectin	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	J Biol Chem 1996, 271: 17697-17703 J Mol Biol 1998, 277: 917-932	GAL+BCG	
H	1AX2	_NAL	Ka	5.4	1.95	Lectin	N-Acetyl-Lactosamine	383.35	$C_{14}H_{25}NO_{11}$	J Biol Chem 1996, 271: 17697-17703 J Mol Biol 1998, 277: 917-932	GAL+NDG	
	1AXR	HTP	Ki	4.5	2.3	Glycogen Phosphorylase	4,5,6-Trihydroxy-7-Hydroxymethyl-4,5,6,7-Tetrahydro-1H-[1,2,3]Triazolo[1,5-A]Pyridin-8-Ylium	202.19	$C_7H_{12}N_3O_{41}$	Helv Chim Acta 1998, 81: 853-864		
B	1AXZ	GAL	Ka	4.4	1.95	Lectin	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Biol Chem 1996, 271: 17697-17703 J Mol Biol 1998, 277: 917-932		
	1B4D	CRA	Ki	6.5	2	Protein (Glycogen Phosphorylase B)	1-Deoxy-1-Methoxycarbamido-Beta-D-Gluco-2-Heptulopyranosonamide	280.23	$C_9H_{16}N_2O_8$	Bioorg Med Chem. 2010, 18: 1171-80		
α	1BAP	ARA	Kd	9.3	1.75	L-Arabinose-Binding Protein	Alpha-L-Arabinose	150.13	$C_5H_{10}O_5$	J Biol Chem 1990, 265(27):16592-16603		
	1BB6	UMG	Ka	7.8	2	Lysozyme	Methyl-Umbellifertl-N-Acetyl-Chitotriose	785.76	$C_{34}H_{47}N_3O_{18}$	Acta Crystallogr D Biol Crystallogr. 1999, 55(Pt 1): 60-6 J Biochem 1980, 87 (4): 1003-1014		
	1BB7	GUM	Ka	5.7	2	Lysozyme	4-Methyl-Umbelliferyl-N-Acetyl-Chitobiose	582.56	$C_{26}H_{34}N_2O_{13}$	Acta Crystallogr D Biol Crystallogr. 1999, 55(Pt 1): 60-6 J Biochem 1980, 87 (4): 1003-1014		
	1BCH	NGA	Ki	5.0	2	Mannose-Binding Protein-A	N-Acetyl-D-Galactosamine	221.21	$C_8H_{15}NO_6$	J Biol Chem. 1998, 273(31):19502-8.		2

Flags	PDB ID	HET ID	Affinity	– ΔG_{exp}	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	1BYK	T6P	Kd	6.8	2.5	Protein (Trehalose Operon Repressor)	Trehalose-6-Phosphate	422.28	$C_{12}H_{23}O_{14}P$	Protein Sci. 1998, 7(12): 2511-21 J Biol Chem 1997, 272(20): 13026-32		A
	1CTT	DHZ	Ki	6.2	2.2	Cytidine Deaminase	3,4-Dihydro-1H-Pyrimidin-2-One Nucleoside	230.22	$C_9H_{14}N_2O_5$	Biochemistry 1995, 34: 4516-4523 Biochemistry 1989, 28(24): 9423-30		
	1CTU	ZEB	Ki	16.3	2.3	Cytidine Deaminase	4-Hydroxy-3,4-Dihydro-Zebularine	246.22	$C_9H_{14}N_2O_6$	Biochemistry 1995, 34: 4516-4523 Biochemistry 1989, 28(24): 9423-30		
XO	1DMB	BCD	Kd	7.8	1.8	D-Maltodextrin Binding Protein	Beta-Cyclodextrin	1134.99	$C_{42}H_{70}O_{35}$	Biochemistry. 1993 Oct 12, 32(40):10553-9 J Biol Chem 1983, 258 (22): 13665-72		
	1DMT	RDF	IC50	10.3	2.1	Neutral Endopeptidase	N-Alpha-L-Rhamnopyranosyloxy(Hydroxyphosphinyl)-L-Leucyl-L-Tryptophan	543.51	$C_{23}H_{34}N_3O_{10}P$	J Med Chem. 2010, 53: 208-20 J Mol Biol 2000, 296: 341-349	Two different IC50 values available: 2.4 and 27 nM from the 2000 and 2010 articles, respectively; the most recent was used	
M	1DOG	NOJ	Ki	5.5	2.3	Glucoamylase-471	1-Deoxynojirimycin	163.17	$C_6H_{13}NO_4$	Biochemistry 1993, 32: 161 Angew Chem Intel Ed Eng 1981, 20(9): 744–761 Natunvissenschaften 1979, 66(11): 584-5	The neutral form is used as reported by authors., although Epik suggests the doubly protonated species to be dominant We deleted a second copy of the ligand in what's described by authors to be a "low affinity subsite"	
	1DRJ	RIP	Kd	10.1	2.5	D-Ribose-Binding Protein	Ribose	150.13	$C_5H_{10}O_5$	J Biol Chem 1994, 269(48): 30206-11		
	1DRK	RIP	Kd	9.3	2	D-Ribose-Binding Protein	Ribose	150.13	$C_5H_{10}O_5$	J Biol Chem 1994, 269(48): 30206-11		
H	1E55	_DHR	Ki	5.6	2	Beta-Glucosidase	Dhurrin	311.29	$C_{14}H_{17}NO_7$	Proc Natl Acad Sci USA 2000, 97: 13555 J Biol Chem 2000. 275: 20002–20011	DHR+GLC	B
	1E6Q	NTZ	Ki	4.3	1.35	Myrosinase	Nojirimycine Tetrazole	202.17	$C_6H_{10}N_4O_4$	J Biol Chem 2000, 275: 39385		M
	1E6S	GOX	Ki	4.4	1.35	Myrosinase Ma1	(2S,3S,4R,5R)-6-(Hydroxyamino)-2-(Hydroxymethyl)-2,3,4,5-Tetrahydropyridine-3,4,5-Triol	192.17	$C_6H_{12}N_2O_5$	J Biol Chem 2000, 275: 39385		M
XR	1E72	GOX	Ki	4.4	1.6	Myrosinase	(2S,3S,4R,5R)-6-(Hydroxyamino)-2-(Hydroxymethyl)-2,3,4,5-Tetrahydropyridine-3,4,5-Triol	192.17	$C_6H_{12}N_2O_5$	J Biol Chem 2000, 275: 39385	redundant, 1E6S	
H	1EEF	_PEPG	IC50	4.0	1.8	Heat-Labile Enterotoxin B Chain	2-Phenethyl- 7-(2,3-Dihydrophthalazine-1,4-Dione)-Alpha-D-Galactoside	444.43	$C_{22}H_{24}N_2O_8$	Acta Crystallogr D Biol Crystallogr 2001, 57(Pt. 2): 201-212	GLA+I06	G
	1EEI	GAA	IC50	4.3	2	Cholera Toxin B	Metanitrophenyl-Alpha-D-Galactoside	301.25	$C_{12}H_{15}NO_8$	Acta Crystallogr D Biol Crystallogr. 2001, 57(Pt 2):201-12.		D
	1EFI	GAT	IC50	2.6	1.6	Heat-Labile Enterotoxin B Chain	4'-Aminophenyl-Alpha-D-Galactopyranoside	271.27	$C_{12}H_{17}NO_6$	Acta Crystallogr D Biol Crystallogr 2001, 57(Pt. 2): 201-212		D
	1EOU	SMS	Kd	10.9	2.1	Carbonic Anhydrase II	Sulfamic Acid 2,3-O-(1-Methylethylidene)-4,5-O-Sulfonyl-Beta-Fructopyranose Ester	361.34	$C_9H_{15}NO_{10}S_2$	J Med Chem 2006, 49: 3496-500		
X	1EXV	700	IC50	10.0	2.4	Liver Glycogen Phosphorylase	[5-Chloro-1H-Indol-2-Carbonyl-Phenylalaninyl]-Azetidine-3-Carboxylic Acid	425.87	$C_{22}H_{20}ClN_3O_4$	Chem Biol 2000, 7: 677-682	non-carbohydrate ligand	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	1F4X	MGS	Ka	4.6	2.3	Antibody S-20-4, Fab Fragment, Heavy Chain	1,2-O-Dimethyl-4-[2,4-Dihydroxy-Butyramido]-4,6-Dideoxy-Alpha-D-Mannopyranoside	293.32	$C_{12}H_{23}NO_7$	Proc Natl Acad Sci USA, 2000, 97(15):8433-8		H
	1F8B	DAN	Ki	7.4	1.8	Neuraminidase	2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	291.26	$C_{11}H_{17}NO_8$	Protein Sci 2001, 10: 689 Nature 1993, 363: 418		
	1F8C	4AM	Ki	10.1	1.7	Neuraminidase	4-Amino-2-Deoxy-2,3-Dehydro-N-Neuraminic Acid	290.27	$C_{11}H_{18}N_2O_7$	Protein Sci 2001, 10: 689 Nature 1993, 363: 418	Protonated state used (Epik)	
	1F8D	9AM	Ki	4.6	1.4	Neuraminidase	9-Amino-2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	290.27	$C_{11}H_{18}N_2O_7$	Protein Sci 2001, 10: 689 Nature 1993, 363: 418	Protonated state used (Epik)	
	1F8E	49A	Ki	6.6	1.4	Neuraminidase	4,9-Amino-2,4-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	289.29	$C_{11}H_{19}N_3O_6$	Protein Sci 2001, 10: 689 Nature 1993, 363: 418	Protonated state used (Epik)	
	1FD7	AI1	IC50	2.5	1.8	Heat-Labile Enterotoxin B Chain	N-Benzyl-3-(Alpha-D-Galactos-1-Yl)-Benzamide	389.4	$C_{20}H_{23}NO_7$	Acta Crystallogr D Biol Crystallogr. 200, 57(Pt 2):201-12.		D
H	1FH7	_XDNJ	Ki	7.1	1.82	Beta-1,4-Xylanase	Xylobiodeoxynojirimycin	265.26	$C_{10}H_{19}NO_7$	Biochemistry 2000, 39: 11553-63 J Am Chem Soc 2000, 122: 2223-2235	XYP+XDN Protonated state used (Epik + article)	
H	1FH8	_XIFG	Ki	9.4	1.95	Beta-1,4-Xylanase	1,5-Imino-1,4,5-Trideoxy-3-O-(Beta-D-Xylopyran- Osyl)-D-Threo-Pentitol	249.26	$C_{10}H_{19}NO_6$	Biochemistry 2000, 39: 11553-63 J Am Chem Soc 2000, 122: 2223-2235	XYP+XIF Protonated state used (Epik + article)	
H	1FH9	_XLOX	Ki	8.8	1.72	Beta-1,4-Xylanase	D-Xylobiono-(Z)-Hydroximo-1,5-Lactam	294.26	$C_{10}H_{18}N_2O_8$	Biochemistry 2000, 39: 11553-63 J Am Chem Soc 2000, 122: 2223-2235	XYP+LOX Protonated state used (Epik + article)	
H	1FHD	_XXIM	Ki	9.3	1.9	Beta-1,4-Xylanase	(6R,7S,8S)-7,8-Dihydroxy-6-(Beta-D-Xylopyranosyloxy)-5,6,7,8-Tet-Rahydroimidazo[1,2-A]Pyridine	302.28	$C_{12}H_{18}N_2O_7$	Biochemistry 2000, 39: 11553-63 J Am Chem Soc 2000, 122: 2223-2235	XYP+XIM Protonated state used (Epik + article)	
	1FU8	CR6	Ki	4.8	2.35	Glycogen Phosphorylase	1-Deoxy-1-Acetylamino-Beta-D-Gluco-2-Heptulopyranosonamide	264.23	$C_9H_{16}N_2O_7$	Proteins 2005, 61: 966-983 Bioorg Med Chem. 2010, 18: 1171-80		
M	1GA8	DEL	Ki	2.4	2	Galactosyl Transferase Lgtc	4-Deoxylactose	326.3	$C_{12}H_{22}O_{10}$	Nat Struct Biol 2001, 8(2):166-75	Two carbohydrate ligands with known binding affinities under this PDB code	
M	1GA8	UPF	Ki	7.8	2	Galactosyl Transferase Lgtc	Uridine-5'-Diphosphate-2-Deoxy-2-Fluorogalactose	568.3	$C_{15}H_{23}FN_2O_{16}P_2$	Nat Struct Biol 2001, 8(2):166-75	Two carbohydrate ligands with known binding affinities under this PDB code	
	1GAH	ACR	Kd	16.4	2	Glucoamylase-471	Alpha-Acarbose	645.61	$C_{25}H_{43}NO_{18}$	Biochemistry. 1996, 35(25): 8319-28 Carbohyd Res 1992, 227: 29-44 J Biol Chem 1994, 269: 15631-9	Three complexes of acarbose with glucoamylase are available; 1AGM (2.3Å), 1LF9 (2.2Å), and 1GAH (2.0Å) Protonated state of acarbose (NH2+) is used as suggested by Epik and described in article	
	1GAI	GAC	Ki	10.9	1.7	Glucoamylase-471	Dihydro-Acarbose	647.63	$C_{25}H_{45}NO_{18}$	Biochemistry 1996, 35(25): 8319-28	Protonated state (NH2+) suggested by Epik	
	1GCA	GAL	Ki	9.1	1.7	Glucose/Galactose-Binding Protein	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Mol Biol 1993, 233(4): 739-52		
	1GG8	GLG	Ki	4.7	2.31	Protein (Glycogen Phosphorylase)	Alpha-D-Glucopyranosyl-2-Carboxylic Acid Amide	207.18	$C_7H_{13}NO_6$	Biochemistry 1994, 33(19): 5745-58		
	1GPY	G6P	Kd	6.3	2.4	Glycogen Phosphorylase B	Alpha-D-Glucose-6-Phosphate	260.14	$C_6H_{13}O_9P$	J Mol Biol 1993, 232(1): 253-67		

Flags	PDB ID	HET ID	Affinity	– ΔG_{exp}	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
H	1GX4	_NAL	Kd	4.4	1.48	N-Acetyllactosaminide Alpha-1,3-Galactosyltransferase	N-Acetyl-Lactosamine	383.35	$C_{14}H_{25}NO_{11}$	J Biol Chem 2002, 277: 28310-28318	GAL+NAG	A
	1GYM	MYG	IC50	3.7	2.2	Phosphatidylinositol-Specific Phospholipase C	Glucosaminyl-(Alpha-6)-D-Myo-Inositol	341.31	$C_{12}H_{23}NO_{10}$	Biochemistry 1996, 35(29):9496-504	Protonated form (NH3+) suggested by Epik	
H	1GZ9	_FLAC	Ka	4.8	1.7	Erythrina Crista-Galli Lectin	2'-Alpha-L-Fucosyllactose	488.44	$C_{18}H_{32}O_{15}$	J Mol Biol 2002, 321(1): 69-83	FUC+LAT	
	1GZC	LAT	Ka	4.8	1.58	Erythrina Crista-Galli Lectin	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	J Mol Biol 2002, 321(1): 69-83		
	1GZT	FUC	Ka	7.1	1.3	Fucose-Specific Lectin	Alpha-L-Fucose	164.16	$C_6H_{12}O_5$	Proteins 2005, 58: 735-746		A
H	1HEW	_NAG3	Ka	6.8	1.75	Hen Egg White Lysozyme	Tri-N-Acetylchitotriose	627.59	$C_{26}H_{45}N_3O_{16}$	J Mol Biol 1992, 224(3): 613-28	NAG+NAG+NAG	
	1HLF	GL4	Ki	7.7	2.26	Glycogen Phosphorylase	8,9,10-Trihydroxy-7-Hydroxymethyl-2-Thioxo-6-Oxa-1,3-Diaza-Spiro[4.5]Decan-4-One	264.25	$C_8H_{12}N_2O_6S$	Bioorg Med Chem 2002, 10(2): 261-8 Bioorg Med Chem Lett. 1999, 9: 1385-90		
H	1I3H	_2MAN	Ka	6.3	1.2	Concanavalin A	Alpha(1,2)-Dimannose	342.3	$C_{12}H_{22}O_{11}$	J Mol Biol 2001, 310: 875-84 Biochemistry 1994, 33: 1149-56	MAN+MAN	
H	1I82	CBI	Ka	8.7	1.9	Endo-1,4-Beta-Xylanase A	Cellobiose	342.3	$C_{12}H_{22}O_{11}$	Biochemistry 2001, 40, 6248 Biochemistry 2001, 40, 6240	BGC+BGC	
	1I8A	BGC	Ka	5.6	1.9	Endo-1,4-Beta-Xylanase A	Beta-D-Glucose	180.16	$C_6H_{12}O_6$	Biochemistry 2001, 40, 6248 Biochemistry 2001, 40, 6240		AB
XTO	1IGJ	DGX	Kd	13.6	2.5	Igg2A-Kappa 26-10 Fab (Heavy Chain)	Digoxin	780.95	$C_{41}H_{64}O_{14}$	Proc Natl Acad Sci USA 1993, 90(21): 10310-4	Although the missing atoms are solvent exposed and not interacting with binding site residues, they all belong to the carbohydrate part of the glycoside digoxin	AB
	1J01	XIL	Ki	8.8	2	Beta-1,4-Xylanase	3-Hydroxy-4-(3,4,5-Trihydroxy-Tetrahydro-Pyran-2-Yloxy)-Piperidin-2-One	263.25	$C_{10}H_{17}NO_7$	J Am Chem Soc 2000, 122: 4229		
WT	1J8V	LAM	Ki	4.9	2.4	Beta-D-Glucan Glucohydrolase Isoenzyme Exo1	4'-Nitrophenyl-S-(Beta-D-Glucopyranosyl)-(1-3)-(3-Thio-Beta-D-Glucopyranosyl)-(1-3)-Beta-D-Glucopyranoside	641.6	$C_{24}H_{35}NO_{17}S$	Plant Cell. 2002, 14(5):1033-52		
	1JAC	AMG	Ka	6.3	2.43	Jacalin	Alpha-Methyl-D-Galactoside	194.18	$C_7H_{14}O_6$	J Mol Biol 2003, 332: 217-228		AB
	1JAK	IFG	Ki	7.6	1.75	Beta-N-Acetylhexosaminidase	(2R,3R,4S,5R)-2-Acetamido-3,4-Dihydroxy-5-Hydroxymethyl-Piperidine	204.23	$C_8H_{16}N_2O_4$	J Biol Chem 2001, 276: 42131-7	Protonated state (NH2+) is used as suggested by Epik and described in article	
	1JII	383	IC50	12.3	3.2	Tyrosyl-Trna Synthetase	[2-Amino-3-(4-Hydroxy-Phenyl)-Propionylamino]- (2,4,5,8-Tetrahydroxy-7-Oxa-2-Aza-Bicyclo[3.2.1]Oct-3-Yl)- Acetic Acid	413.38	$C_{17}H_{23}N_3O_9$	Protein Sci 2001, 10: 2008-16	Protonated state (NH3+) is used as suggested by Epik and described in article	
	1JIJ	629	IC50	11.6	3.2	Tyrosyl-Trna Synthetase	[2-Amino-3-(4-Hydroxy-Phenyl)-Propionylamino]-(1,3,4,5-Tetrahydroxy-4-Hydroxymethyl-Piperidin-2-Yl)- Acetic Acid	415.4	$C_{17}H_{25}N_3O_9$	Protein Sci 2001, 10: 2008-16	Protonated state (NH3+) is used as suggested by Epik and described in article	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	1JIK	545	IC50	13.2	2.8	Tyrosyl-Trna Synthetase	[2-Amino-3-(4-Hydroxy-Phenyl)-Propionylamino]-(1,3,4,5-Tetrahydroxy-4-Hydroxymethyl-Piperidin-2-Yl)- Acetic Acid Butyl Ester	471.5	$C_{21}H_{33}N_3O_9$	Protein Sci 2001, 10: 2008-16	Protonated state (NH3+) is used as suggested by Epik and described in article	
	1JIL	485	IC50	11.4	2.2	Tyrosyl-Trna Synthetase	[2-Amino-3-(4-Hydroxy-Phenyl)-Propionylamino]- (3,4,5-Trihydroxy-6-Methyl-Tetrahydro-Pyran-2-Yl)-Acetic Acid	384.39	$C_{17}H_{24}N_2O_8$	Protein Sci 2001, 10: 2008-16	Protonated state (NH3+) is used as suggested by Epik and described in article	
H	1JLX	_TDSC	IC50	7.9	2.2	Agglutinin	Benzyl T-Antigen Disaccharide	473.47	$C_{21}H_{31}NO_{11}$	Nat Struct Biol 1997, 4(10): 779-783 J Biol Chem 19889, 264: 16123-16131	GAL+A2G+MBN Ligand interacts with residues from the two chains, ligand's copy in chain B was deleted	AB
	1JQY	A32	Kd	6.7	2.14	Heat-Labile Enterotoxin B Chain	(3-Nitro-5-(3-Morpholin-4-Yl-Propylaminocarbonyl)Phenyl)-Galactopyranoside	471.46	$C_{20}H_{29}N_3O_{10}$	Chem Biol 2002, 9(2): 215-24	Ligand interacts with residues from the two chains, ligand's copy in chain D was deleted	DE
	1JR0	A24	Kd	6.7	1.3	Cholera Toxin B Subunit	(3-Nitro-5-(2-Morpholin-4-Yl-Ethylaminocarbonyl)Phenyl)-Galactopyranoside	457.44	$C_{19}H_{27}N_3O_{10}$	Chem Biol 2002, 9(2): 215-24	Ligand interacts with residues from the two chains, ligand's copy in chain D was deleted	DE
XM	1JZ7	GAL	Ki	0.5	1.5	Beta-Galactosidase	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	Biochemistry 2001, 40(49): 14781-94	Multiple ligand copies, substantially different in binding poses Complex of intermediates along interaction coordinates	
H	1JZN	LAT	Ki	5.6	2.2	Galactose-Specific Lectin	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	Biochemistry 2004, 43: 3783-92	BGC+GAL	A
	1JZS	MRC	Ki	9.0	2.5	Isoleucyl-Trna Synthetase	Mupirocin	500.63	$C_{26}H_{44}O_9$	J Biol Chem 2001, 276(50): 47387-93		
M	1K06	BZD	Ki	7.3	1.8	Glycogen Phosphorylase	N-Benzoyl-N'-Beta-D-Glucopyranosyl Urea	326.3	$C_{14}H_{18}N_2O_7$	Bioorg Med Chem 2009, 17: 4773-85 Eur J Biochem 2002, 269 :1684-96	Two ligand copies, one in the catalytic site and the other in a "new" allosteric site (cf. 1KTI) Affinity measurement pertains to the allosteric inhibition, thus the former copy was deleted	
	1K1Y	ACR	Ki	4.4	2.4	4-Alpha-Glucanotransferase	Alpha-Acarbose	645.61	$C_{25}H_{43}NO_{18}$	J Biol Chem 2003, 278(21): 19378-86		A
HM	1K7T	_NGGA	Ka	5.4	2.4	Agglutinin Isolectin 3	Glcnac-Beta-1,6-Gal	383.35	$C_{14}H_{25}NO_{11}$	Biochim Biophys Acta 2002, 1569: 10-20	NAG+GAL Ligand copy in chain B deleted	AB
HM	1K7U	_2NAG	Ka	5.3	2.2	Agglutinin Isolectin 3	Glcnac-Beta-1,4-Glcnac	424.4	$C_{16}H_{28}N_2O_{11}$	Biochim Biophys Acta 2002, 1569: 10-20	NAG+NAG Ligand copy in chain B deleted	AB
	1KTI	AZC	Ki	4.7	1.97	Glycogen Phosphorylase, Muscle Form	N-Acetyl-N'-Beta-D-Glucopyranosyl Urea	264.23	$C_9H_{16}N_2O_7$	Bioorg Med Chem 2009, 17: 4773-85 Eur J Biochem 2002, 269 :1684-96	Unlike 1K06, only one resolved ligand in the "new" allosteric site, to which the affinity measurement is ascribed	
	1KZN	CBN	Ka	12.2	2.3	Dna Gyrase Subunit B	Clorobiocin	697.14	$C_{35}H_{37}CLN_2O_{11}$	Biochemistry 2002, 41(23): 7217-23	Neutral form used as indicated in article, although Epik suggested the deportonated form	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
HWM	1LAX	MAL	Kd	7.8	1.85	Maltose-Binding Protein Mutant Male31	Maltose	342.3	$C_{12}H_{22}O_{11}$	Protein Sci 2003, 12(3): 577-85	GLC+GLC	A
H	1LZB	_NAGT	Kd	6.7	1.5	Hen Egg White Lysozyme	Tri-N-Acetylchitotriose	627.59	$C_{24}H_{41}N_3O_{16}$	J Mol Biol. 1995 Mar 24, 247(2):281-93.	NAG+NAG+NAG	
	1M01	NAG	Kd	3.4	2.1	Beta-N-Acetylhexosaminidase	N-Acetyl-D-Glucosamine	221.21	$C_8H_{15}NO_6$	J Biol Chem 2002, 277(42):40055-65		
H	1M26	_TANT	Ka	6.9	1.62	Jacalin	(6R,7S,8S)-7,8-Dihydroxy-6-(Beta-D-Xylopyranosyloxy)-5,6,7,8-Tet-Rahydroimidazo[1,2-A]Pyridine	383.35	$C_{14}H_{25}NO_{11}$	J Mol Biol 2003, 332: 217-228	GAL+A2G	AB
	1M6P	M6P	Kd	7.0	1.8	Cation-Dependent Mannose-6-Phosphate Receptor	Alpha-D-Mannose-6-Phosphate	260.14	$C_6H_{13}O_9P$	Cell 1998, 93(4): 639-48 J Biol Chem 1989, 264(14): 7962-9		A
	1MOQ	GLP	Ki	4.7	1.57	Glucosamine 6-Phosphate Synthase	Glucosamine 6-Phosphate	259.15	$C_6H_{14}NO_8P$	Structure 1998, 6(8): 1047-55 Biochemistry 1988, 27(7): 2282-7	Protonated amino group suggested by Epik making an extra H-bond, used	
	1N3W	MAL	Kd	9.5	2.6	Maltose-Binding Periplasmic Protein	Maltose	342.3	$C_{12}H_{22}O_{11}$	J Biol Chem 2003, 278(36): 34555-67		
M	1NAA	ABL	Ki	4.9	1.8	Cellobiose Dehydrogenase	(2R,3R,4R,5R)-4,5-Dihydroxy-2-(Hydroxymethyl)-6-Oxopiperidin-3-Yl Beta-D-Glucopyranoside	339.3	$C_{12}H_{21}NO_{10}$	J Biol Chem 2003, 278(9): 7160-6	Ligand copy in chain B was not deleted	AB
	1NF3	GNP	Kd	10.0	2.1	G25K Gtp-Binding Protein, Placental Isoform	Phosphoaminophosphonic Acid-Guanylate Ester	522.2	$C_{10}H_{17}N_6O_{13}P_3$	EMBO J 2003, 22: 1125-33		A
M	1NJJ	GET	Ki	2.9	2.45	Ornithine Decarboxylase	Geneticin	496.56	$C_{20}H_{40}N_4O_{10}$	J Biol Chem 2003, 278: 22037-43	Copies of the two ligands (GET and ORX) in chain B were deleted Epik suggested a state with a total charge of +3, used	AB
XM	1NJJ	ORX	Kd	4.9	2.45	Ornithine Decarboxylase	N~2~-((3-Hydroxy-2-Methyl-5-[(Phosphonooxy)Methyl]Pyridin-4-Yl)Methyl)-D-Ornithine	363.3	$C_{13}H_{22}N_3O_7P$	J Biol Chem 2003, 278: 22037-43	non-carbohydrate ligand	
	1NOI	NTZ	Ki	5.8	2.5	Glycogen Phosphorylase	Nojirimycine Tetrazole	202.17	$C_6H_{10}N_4O_4$	Biochemistry 1996, 35: 7341-55		A
	1NOJ	NTZ	Ki	5.8	2.4	Glycogen Phosphorylase	Nojirimycine Tetrazole	202.17	$C_6H_{10}N_4O_4$	Biochemistry 1996, 35: 7341-55		
	1NOK	NTZ	Ki	4.3	2.4	Glycogen Phosphorylase	Nojirimycine Tetrazole	202.17	$C_6H_{10}N_4O_4$	Biochemistry 1996, 35: 7341-55		
XMH	1NPL	_MAN2	Ka	3.7	2	Protein (Agglutinin)	Alpha-1,3 Mannobiose	356.32	$C_{13}H_{24}O_{11}$	J Mol Biol 1999, 290(1): 185-99	MAN+MAN Multiple ligand binding sites	
	1O7O	LAT	Kd	3.5	1.97	N-Acetyllactosaminide Alpha-1,3-Galactosyltransferase	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	Biochemistry 2003, 42(46): 13512-21		A
X	1O9V	SNG	Ka	5.0	1.75	F17-Ag Lectin	Methyl 2-Acetamido-1,2-Dideoxy-1-Seleno-Beta-D-Glucopyranoside	298.2	$C_9H_{17}NO_5Se$	Mol Microbiol 2003, 49: 705-15	Atom-type problem: ligand is a seleno-glycoside	
	1O9W	NAG	Ka	4.0	1.65	F17-Ag Lectin	N-Acetyl-D-Glucosamine	221.21	$C_8H_{15}NO_6$	Mol Microbiol 2003, 49: 705-15		
H	1OCQ	_GIFG	Ki	7.1	1.08	Endoglucanase 5A	Glc-Beta(1,4)-Ifg	309.31	$C_{12}H_{23}NO_8$	J Am Chem Soc 2003, 125: 7496-7	IFM+BGC Protonated state (NH2+) used as suggested by Epik and described in article	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
XMH	1OD8	_XIF2	Ki	6.9	1.05	Endo-1,4-Beta-Xylanase A	(3S,4R)-3-Hydroxy-4-(((2S,3R,4S,5R)-3,4,5-Trihydroxytetrahydro-2H-Pyran-2-Yl)Oxy)Piperidin-2-One	263.24	$C_{10}H_{17}NO_7$	Chem Commun (Camb). 2003 (8): 944-5	XYP+XDL Two ligand molecules bound in two adjacent subsites	
XM	1OFZ	FUL	Kd	6.3	1.5	Fucose-Specific Lectin	Beta-L-Fucose	164.16	$C_6H_{12}O_5$	J Biol Chem. 2003, 278(29): 27059-67		
	1OGD	RIP	Kd	4.1	1.95	High Affinity Ribose Transport Protein Rbsd	Ribose(Pyranose Form)	150.13	$C_5H_{10}O_5$	J Biol Chem 2003, 278(30): 28173-80		A
	1OIF	IFM	Kd	9.8	2.12	Beta-Glucosidase	5-Hydroxymethyl-3,4-Dihydroxypiperidine	147.17	$C_6H_{13}NO_3$	J Am Chem Soc 2003, 125: 14313-23	Protonated state (NH2+) is used as suggested by Epik and described in article	A
	1OIM	NOJ	Kd	6.9	2.15	Beta-Glucosidase A	1-Deoxynojirimycin	163.17	$C_6H_{13}NO_4$	J Am Chem Soc 2003, 125: 14313-23	Protonated state (NH2+) is used as suggested by Epik and described in article	A
β	1OKO	GAL	Ka	6.2	1.6	PA-I Galactophilic Lectin	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	FEBS Lett 2003, 555(2): 297-301		A
α	1OXC	FUC	Ka	7.1	1.2	PA-LII (LecB)	Alpha-L-Fucose	164.16	$C_6H_{12}O_5$	Proteins 2005, 58: 735-746		A
	1P4G	CGF	Ki	3.7	2.1	Glycogen Phosphorylase, Muscle Form	C-(1-Azido-Alpha-D-Glucopyranosyl) Formamide	250.21	$C_7H_{14}N_4O_6$	Biocatal Biotransfor 2003, 21: 233-242		
	1P4H	CR6	Ki	4.8	2.06	Glycogen Phosphorylase, Muscle Form	1-Deoxy-1-Acetylamino-Beta-D-Gluco-2-Heptulopyranosonamide	264.23	$C_9H_{16}N_2O_7$	Biocatal Biotransfor 2003, 21: 233-242		
	1P4J	CBF	Ki	4.2	2	Glycogen Phosphorylase, Muscle Form	C-(1-Hydrogyl-Beta-D-Glucopyranosyl) Formamide	223.18	$C_7H_{13}NO_7$	Biocatal Biotransfor 2003, 21: 233-242		
	1PX4	IPT	Ki	4.4	1.6	Beta-Galactosidase	Isopropyl-1-Beta-D-Thiogalactoside	238.3	$C_9H_{18}O_5S$	Biochemistry 2003, 42: 13505-13511		AD
WT	1PZI	1DM	Kd	5.8	1.99	Heat-Labile Enterotoxin B Subunit	N-(2-Morpholin-4-Yl-1-Morpholin-4-Ylmethyl-Ethyl)-3-Nitro-5-(3,4,5-Trihydroxy-6-Hydroxymethyl-Tetrahydro-Pyran-2-Yloxy)-Benzamide	556.57	$C_{24}H_{36}N_4O_{11}$	Bioorg Med Chem 2004, 12(5):907-20	Ligand copy in chain E was deleted	DE
XM	1PZJ	15B	IC50	4.8	1.46	Cholera Toxin B Subunit	N-(3-[4-(3-Amino-Propyl)-Piperazin-1-Yl]-Propyl)-3-Nitro-5-(Galactopyranosyl)-Beta-Benzamide	527.57	$C_{23}H_{37}N_5O_9$	Bioorg Med Chem 2004, 12(5):907-20	Two very differently positioned anomers in crystal structure	
XM	1PZJ	J15	IC50	4.8	1.46	Cholera Toxin B Subunit	N-(3-[4-(3-Amino-Propyl)-Piperazin-1-Yl]-Propyl)-3-Nitro-5-(Galactopyranosyl)-Alpha-Benzamide	527.57	$C_{23}H_{37}N_5O_9$	Bioorg Med Chem 2004, 12(5):907-20	Two very differently positioned anomers in crystal structure	
WT	1PZK	J12	IC50	5.0	1.35	Cholera Toxin B Subunit	N-(3-[4-(3-Amino-Propyl)-Piperazin-1-Yl]-Propyl)-3-(2-Thiophen-2-Yl-Acetylamino)-5-(3,4,5-Trihydroxy-6-Hydroxymethyl-Tetrahydro-Pyran-2-Yloxy)-Benzamide	621.75	$C_{29}H_{43}N_5O_8S$	Bioorg Med Chem 2004, 12(5):907-20	Ligand copy in chain E was deleted	DE

Flags	PDB ID	HET ID	Affinity	– ΔG_{exp}	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
XTO	1RCV	BV1	IC50	6.2	1.6	Cholera Toxin B Protein (Ctb)	[3-(4-(3-[3-Nitro-5-(Galactopyranosyloxy)-Benzoylamino]-Propyl}-Piperazin-1-Yl)-Propylamino] -2-(3-{4-[3-(3-Nitro-5-[Galactopyranosyloxy]-Benzoylamino)-Propyl]-Piperazin-1-Yl) -Propyl-Amino)-3,4-Dioxo-Cyclobutene	1133.17	$C_{50}H_{72}N_{10}O_{20}$	Chem Biol 2004, 11(9):1205-15	Bivalent ligand with a very long flexible linker whose atoms are not fully resolved	
XTO	1RD9	BV2	IC50	6.7	1.44	Cholera Toxin B Protein (Ctb)	1,3-Bis-([3-(4-(3-[3-Nitro-5-(Galactopyranosyloxy)-Benzoylamino]-Propyl)-Piperazin-1-Yl)-Propyl-Amino]-Carbonyloxy)-2-Amino-Propane	1198.25	$C_{51}H_{79}N_{11}O_{22}$	Chem Biol 2004, 11(9):1205-15	Bivalent ligand with a very long flexible linker whose atoms are not fully resolved	
	1RDI	MFU	Ki	2.8	1.8	Mannose-Binding Protein-C	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	J Biol Chem 1996, 271(2):663-74		1
	1RDJ	MFB	Ki	2.3	1.8	Mannose-Binding Protein-C	Beta-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	J Biol Chem. 1996 Jan 12, 271(2):663-74.	missing the glycosidic methyl group, manually added in preparation step	1
	1RDK	GAL	Ki	1.3	1.8	Mannose-Binding Protein-C	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Biol Chem. 1996 Jan 12, 271(2):663-74.		1
	1RDL	MMA	Ki	3.1	1.7	Mannose-Binding Protein-C	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	J Biol Chem 1996, 271(2):663-74	missing the glycosidic methyl group, manually added in preparation step	1
H	1RDN	_MNDG	Ki	2.5	1.8	Mannose-Binding Protein-C	Alpha-Methyl-D-N-Acetylglucosaminide	235.23	$C_9H_{17}NO_6$	J Biol Chem 1996, 271(2):663-74	missing the glycosidic methyl group, manually added in preparation step	1
XTO	1RDP	BV3	IC50	6.9	1.35	Cholera Toxin B Protein (Ctb)	1,3-Bis-([[3-(4-(3-[3-Nitro-5-(Galactopyranosyloxy)-Benzoylamino]-Propyl)-Piperazin-1-Yl)-Propylamino-3,4-Dioxo-Cyclobutenyl]-Amino-Ethyl]-Amino-Carbonyloxy)-2-Amino-Propane	1474.5	$C_{63}H_{91}N_{15}O_{26}$	Chem Biol. 2004 Sep, 11(9):1205-15.	Bivalent ligand with a very long flexible linker whose atoms are not fully resolved	
XTO	1RF2	BV4	IC50	6.5	1.35	Cholera Toxin B Protein (Ctb)	1,3-Bis-([3-[3-[3-(4-(3-[3-Nitro-5-(Galactopyranosyloxy)-Benzoylamino]-Propyl)-Piperazin-1-Yl)-Propylamino-3,4-Dioxo-Cyclobutenyl]-Amino-Propoxy-Ethoxy-Ethoxy]-Propyl-]Amino-Carbonyloxy)-2-Amino-Propane	1794.92	$C_{79}H_{123}N_{15}O_{32}$	Chem Biol. 2004 Sep, 11(9):1205-15.	Bivalent ligand with a very long flexible linker whose atoms are not fully resolved	
	1RO7	CSF	Ki	4.3	1.8	Alpha-2,3/8-Sialyltransferase	Cytidine-5'-Monophosphate-3-(A)-Fluoro-N-Acetyl-Neuraminic Acid	632.45	$C_{20}H_{30}FN_4O_{16}P$	Nat Struct Mol Biol 2004, 11: 163-170	Second suggested protonaion state from Epik used to match chemical strucutre reported in article	A
	1RPJ	ALL	Kd	8.8	1.8	Protein (Precursor Of Periplasmic Sugar Receptor)	D-Allopyranose	180.16	$C_6H_{12}O_6$	J Mol Biol 1999, 286(5): 1519-31		
	1S14	NOV	Ki	9.5	2	Topoisomerase IV Subunit B	Novobiocin	612.63	$C_{31}H_{36}N_2O_{11}$	Antimicrob Agents Chemother 2004, 48: 1856-1864 J Med Chem. 2008, 51: 5243-63	Neutral state used (Epik)	A

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	1SEU	SA3	IC50	10.3	3	DNA Topoisomerase I	2,10-Dihydroxy-12-(Beta-D-Glucopyranosyl)-6,7,12,13-Tetrahydroindolo[2,3-A]Pyrrolo[3,4-C]Carbazole-5,7-Dione	519.47	$C_{26}H_{21}N_3O_9$	J Med Chem 2005, 48: 2336-45		
XM	1SZ0	M6P	Kd	12.3	2.1	Cation-Independent Mannose 6-Phosphate Receptor	Alpha-D-Mannose-6-Phosphate	260.14	$C_6H_{13}O_9P$	J Biol Chem 2004, 279(32): 34000-9	Binding site not well-defined, two ligand molecules close to each other	
	1TLG	GAL	Ki	4.6	2.2	Polyandrocarpa Lectin	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Mol Biol. 1999, 290(4): 867-79		A
	1U33	LM2	Ki	6.3	1.95	Alpha-Amylase, Pancreatic	4'-O-Methyl-Maltosyl-Alpha (1,4)-(Z, 3S,4S,5R,6R)-3,4,5-Trihydroxy-6-Hydroxymethyl-Piperidin-2-One	530.48	$C_{19}H_{34}N_2O_{15}$	J Biol Chem 2004, 279(46): 48282-91		
	1UDA	UFG	Kd	4.0	1.8	UDP-Galactose-4-Epimerase	Uridine-5'-Diphosphate-4-Deoxy-4-Fluoro-Alpha-D-Galactose	568.3	$C_{15}H_{23}FN_2O_{16}P_2$	Biochemistry 1997, 36: 6294-6304 J Org Chem 1994, 59: 6994- 6998		
H	1UDB	_UFGC	Kd	3.9	1.65	UDP-Galactose-4-Epimerase	Uridine-5'-Diphosphate-4-Deoxy-4-Fluoro-Alpha-D-Glucose	568.3	$C_{15}H_{23}FN_2O_{16}P_2$	Biochemistry 1997, 36: 6294-6304 J Org Chem 1994, 59: 6994- 6998	UFG (galactose in PDB) corrected to glucose	
	1UGW	GAL	Ka	4.2	1.7	Jacalin	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Mol Biol 2003, 332: 217-228		AB
H	1UGX	_MTNT	Ka	8.1	1.6	Jacalin	Galactose-Beta(1-3)-N-Acetyl-D-Galactosamine-Alpha-O-Me	397.37	$C_{15}H_{27}NO_{11}$	J Mol Biol 2003, 332: 217-228	GAL+MGC	AB
H	1UGY	_GAGC	Ka	5.1	2.4	Jacalin	Galacose-Alpha(1-6)-Glucose	342.3	$C_{12}H_{22}O_{11}$	J Mol Biol 2003, 332: 217-228	GLA+GLC	AB
	1UH0	MGC	Ka	6.6	2.8	Jacalin	Alpha-Methyl-N-Acetyl-D-Galactosamine	235.23	$C_9H_{17}NO_6$	J Mol Biol 2003, 332: 217-228		AB
H	1UH1	_NGMG	Ka	7.5	2.8	Jacalin	N-Acetyl-Galactosamine-Beta(1-3)-Galacose-Alpha-O-Me	397.37	$C_{15}H_{27}NO_{11}$	J Mol Biol 2003, 332: 217-228	AMG+NGA	CD
H	1ULC	LAT	Kd	5.5	2.6	Galectin-2	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	Structure 2004, 12: 689-702	GAL+BGC	A
H	1ULD	_BLDH	Kd	7.3	2.1	Galectin-2	Blood Group H Type II	529.49	$C_{20}H_{35}NO_{15}$	Structure 2004, 12: 689-702	GAL+NAG+FUC	A
H	1ULE	_LNB2	Kd	8.1	2.15	Galectin-2	Linear B2 Trisaccharide	545.49	$C_{20}H_{35}NO_{16}$	Structure 2004, 12: 689-702	GLA+GAL+NAG	A
H	1ULG	_TFAN	Kd	5.7	2.2	Galectin-2	Thomsen-Friedenreich Antigen	383.35	$C_{14}H_{25}NO_{11}$	Structure 2004, 12: 689-702	GAL+NAG	A
	1URG	MAL	Kd	7.9	1.8	Maltose-Binding Protein	Maltose	342.3	$C_{12}H_{22}O_{11}$	J Mol Biol 2004, 335(1): 261-74 J Bacteriol 2000, 182(22): 6292-301		
	1UWF	DEG	Kd	9.3	1.69	Fimh Protein	Butyl Alpha-D-Mannopyranoside	236.26	$C_{10}H_{20}O_6$	Mol Microbiol. 2005, 55(2): 441-55		
	1UWT	GTL	Ki	8.1	1.95	Beta-Galactosidase	D-Galactohydroximo-1,5-Lactam	192.17	$C_6H_{12}N_2O_5$	Biochemistry 2004, 43: 6101-9		A
	1UWU	GOX	Ki	8.2	1.95	Beta-Galactosidase	(2S,3S,4R,5R)-6-(Hydroxyamino)-2-(Hydroxymethyl)-2,3,4,5-Tetrahydropyridine-3,4,5-Triol	192.17	$C_6H_{12}N_2O_5$	Biochemistry 2004, 43: 6101-9		A
	1UZV	FUC	Ka	7.1	1	Pseudomonas Aeruginosa Lectin Ii	Alpha-L-Fucose	164.16	$C_6H_{12}O_5$	Proteins 2005, 58: 735-746		AB
H	1V0K	_XDNJ	Ki	6.0	1.03	Endo-1,4-Beta-Xylanase A	Xylobiodeoxynojirimycin	265.26	$C_{10}H_{19}NO_7$	Chem Commun 2004, 3(16): 1794-5	XYP+XDN Protonated state (NH2+) used (Epik)	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
H	1V0L	_XIFG	Ki	8.6	0.98	Endo-1,4-Beta-Xylanase A	1,5-Imino-1,4,5-Trideoxy-3-O-(Beta-D-Xylopyran- Osyl)-D-Threo-Pentitol	249.26	$C_{10}H_{19}NO_6$	Chem Commun 2004, 3(16): 1794-5	XYP+XIF Protonated state (NH2+) used (Epik)	
	1VZT	UDP	Kd	6.1	2	N-Acetyllactosaminide Alpha-1,3-Galactosyltransferase	Uridine-5'-Diphosphate	404.16	$C_9H_{14}N_2O_{12}P_2$	Biochemistry 2003, 42(46): 13512-21		A
	1W3J	OXZ	Kd	8.6	2	Beta-Glucosidase	Tetrahydrooxazine	149.15	$C_5H_{11}NO_4$	J Biol Chem 2004, 279: 49236-42		A
H	1W3K	_CELB	Ki	5.9	1.2	Endoglucanase 5A	Cellobio-Tetrahydrooxazine	311.29	$C_{11}H_{21}NO_9$	J Biol Chem 2004, 279: 49236-42	BGC+OXZ	
H	1W3L	_CELT	Ki	8.6	1.04	Endoglucanase 5A	Cellotrio-Tetrahydrooxazine	473.43	$C_{17}H_{31}NO_{14}$	J Biol Chem 2004, 279: 49236-42	BGC+BGC+OXZ Water molecules overlapping with ligand were removed	
	1W6O	LAT	Ka	4.6	1.9	Galectin-1	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	J Mol Biol 2004, 343(4): 957-70		A
H	1W6P	_GAND	Ka	5.5	1.8	Galectin-1	N-Acetyl-Lactosamine	383.35	$C_{14}H_{25}NO_{11}$	J Mol Biol 2004, 343(4): 957-70	NDG+GAL	A
WTH	1W8F	_LNPV	Ka	8.4	1.05	Pseudomonas Aeruginosa Lectin II	Lacto-N-Neo-Fucopentaose V	853.77	$C_{32}H_{55}NO_{25}$	Biochem J 2005, 389 (2): 325-32	FUC+BGC+GAL+GAL+NAG Ligand copy in chain C deleted	AC
XMH	1W8H	_LEWA	Ka	9.1	1.75	Pseudomonas Aeruginosa Lectin II	Lewis A Trisaccharide	529.49	$C_{20}H_{35}NO_{15}$	Biochem J 2005, 389 (2): 325-32	NDG+FUC+GAL or NAG+FUC+GAL Ligand copies positioned differently in binding site	
XMH	1W9T	_XYLB	Ka	4.7	1.62	Bh0236 Protein	Beta-1,4-Xylobiose	282.24	$C_{10}H_{18}O_9$	J Biol Chem 2005, 280: 530-7	XYP+XYP or XYP+XYS Multiple ligand binding sites	
H	1W9W	_LMHX	Ka	7.4	2.1	Bh0236 Protein	Laminarihexaose	990.86	$C_{36}H_{62}O_{31}$	J Biol Chem 2005, 280: 530-7	BGC+BGC+BGC+GLC+BGC+BGC	
	1WS4	GYP	Ka	4.1	1.9	Agglutinin Alpha Chain	Methyl-Alpha-D-Glucopyranoside	194.18	$C_7H_{14}O_6$	J Mol Biol 2005, 347(1): 181-8		AB
	1WS5	MMA	Ka	4.1	1.9	Agglutinin Alpha Chain	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	J Mol Biol 2005, 347(1): 181-8		AB
	1WW2	NBG	Ki	6.1	1.9	Glycogen Phosphorylase, Muscle Form	1-N-Acetyl-Beta-D-Glucosamine	221.21	$C_8H_{15}NO_6$	Bioorg Med Chem 2006, 14: 181-189		
	1WW3	NTF	Ki	5.6	1.8	Glycogen Phosphorylase, Muscle Form	N-Trifluro-Acetyl-Beta-D-Glucopyranosylamine	275.18	$C_8H_{12}F_3NO_6$	Bioorg Med Chem 2006, 14: 181-189		
	1X9D	SMD	Kd	5.4	1.41	Endoplasmic Reticulum Mannosyl-Oligosaccharide 1,2-Alpha-Mannosidase	Methyl-2-S-(Alpha-D-Mannopyranosyl)-2-Thio-Alpha-D-Mannopyranoside	372.39	$C_{13}H_{24}O_{10}S$	J Biol Chem 2005, 280(16): 16197-207		
	1XC7	GL6	Ki	3.0	1.83	Glycogen Phosphorylase, Muscle Form	(3,4,5-Trihydroxy-6-Hydroxymethyl-Tetrahydro-Pyran-2-Yl)-Phosphoramidic Acid Dimethyl Ester	287.21	$C_8H_{18}NO_8P$	Bioorg Med Chem 2005, 13(3): 765-72		
	1XD0	ARE	Ki	9.7	2	Alpha-Amylase	Acarbose Derived Pentasaccharide	807.75	$C_{31}H_{53}NO_{23}$	Biochemistry 2005, 44(9): 3347-57	Neutral state used (Epik + article)	
	1XD1	6SA	Ki	10.8	2.2	Alpha-Amylase	Acarbose Derived Hexasaccharide	969.9	$C_{37}H_{63}NO_{28}$	Biochemistry 2005, 44(9): 3347-57	Neutral state used (Epik + article)	
XM	1XKX	IMK	Ki	6.9	1.93	Glycogen Phosphorylase, Muscle Form	2-(Beta-D-Glucopyranosyl)-5-Methyl-1-Benzimidazole	294.31	$C_{14}H_{18}N_2O_5$	Protein Sci 2005, 14(4): 873-88	Multiple binding sites, and multiple ligand molecules bound	
	1XL0	OX2	Ki	5.2	1.92	Glycogen Phosphorylase, Muscle Form	2-(Beta-D-Glucopyranosyl)-5-Methyl-1,3,4-Oxadiazole	246.22	$C_9H_{14}N_2O_6$	Protein Sci 2005, 14(4): 873-88		

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
H	1XL1	_TH1	Ki	5.6	2.1	Glycogen Phosphorylase, Muscle Form	2-(Beta-D-Glucopyranosyl)-Benzothiazole	297.33	$C_{13}H_{15}NO_5S$	Protein Sci 2005, 14(4): 873-88	Incorrect HET record in PDB, the ligand has no 5-methyl group (article)	
	1XLI	GLT	Ki	2.0	2.5	D-Xylose Isomerase	5-Deoxy-5-Thio-Alpha-D-Glucose	196.22	$C_6H_{12}O_5S$	J Mol Biol 1990, 212: 211-35		A
WT	1XNK	XS2	Ka	5.5	1.55	Endoxylanase 11A	Methyl4,4II,4III,4IV-Tetrathio-Beta-D-Xylopentoside	756.86	$C_{26}H_{44}O_{17}S_4$	FEBS J 2005, 272(9): 2317-33	Missing two solvent-exposed sugar units, which do not have direct interactions with binding site residues	A
XM	1XOI	288	IC50	9.9	2.1	Glycogen Phosphorylase, Liver Form	5-Chloro-1H-Indole-2-Carboxylic Acid([Cyclopentyl-(2-Hydroxy-Ethyl)-Carbamoyl]-Methyl)-Amide	365.85	$C_{18}H_{24}ClN_3O_3$	Bioorg Med Chem Lett 15: 459-465	Two ligand molecules binding at the dimer interface	
	1YFZ	IMP	Ki	5.9	2.2	Hypoxanthine-Guanine Phosphoribosyltransferase	Inosinic Acid	348.21	$C_{10}H_{13}N_4O_8P$	J Mol Biol 2005, 348: 1199-210		A
	1Z3T	CBI	Kd	5.4	1.7	Cellulase	Cellobiose	342.3	$C_{12}H_{22}O_{11}$	FEBS J 2005, 272(8): 1952-64		
W	1Z3V	LAT	Kd	5.6	1.61	Cellulase	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	FEBS J 2005, 272(8): 1952-64	Incorrect glucose geometry in crystal structure (very flat), corrected manually	
	1Z4O	GL1	Ki	6.2	1.9	Beta-Phosphoglucomutase	1-O-Phosphono-Alpha-D-Galactopyranose	260.14	$C_6H_{13}O_9P$	J Am Chem Soc. 2005 Apr 20, 127(15):5298-9		A
XO	2A4W	BLM	Kd	11.3	1.5	Mitomycin-Binding Protein	Bleomycin A2	1416.56	$C_{55}H_{85}N_{17}O_{21}S_3$	J Mol Biol 2006, 360(2): 398-408	Very bing ligand, with a bound copper atom	
	2AAC	FCB	Ki	3.0	1.6	Arac	Beta-D-Fucose	164.16	$C_6H_{12}O_5$	J Mol Biol 1997, 273(1): 226-37		A
	2ADD	SUC	Ki	3.3	2.5	Fructan 1-Exohydrolase Iia	Sucrose	342.3	$C_{12}H_{22}O_{11}$	New Phytol 2007, 174: 90-100		
	2AM4	U2F	Ki	5.1	1.7	Alpha-1,3-Mannosyl-Glycoprotein 2-Beta-N-Acetylglucosaminyltransferase	Uridine-5'-Diphosphate-2-Deoxy-2-Fluoro-Alpha-D-Glucose	568.29	$C_{15}H_{23}FN_2O_{16}P_2$	J Mol Biol 2006, 360: 67-79		
	2APC	UDM	Ki	6.2	1.5	Alpha-1,3-Mannosyl-Glycoprotein 2-Beta-N-Acetylglucosaminyltransferase	Uridine-Diphosphate-Methylene-N-Acetyl-Glucosamine	605.39	$C_{18}H_{29}N_3O_{16}P_2$	J Mol Biol 2006, 360: 67-79		
	2ARC	ARA	Kd	4.1	1.5	Arabinose Operon Regulatory Protein	Alpha-L-Arabinose	150.13	$C_5H_{10}O_5$	Science 1997 , 276(5311):421-5		A
	2ARE	MAN	Ka	4.5	1.8	Lectin	Alpha-D-Mannose	180.16	$C_6H_{12}O_6$	FEBS J 2006, 273 : 2407-2420		A
	2B1Q	TRE	Ki	2.2	2.2	Hypothetical Protein Slr0953	Trehalose	342.3	$C_{12}H_{22}O_{11}$	Proteins 2007, 68: 796-801		
	2B1R	CBI	Ki	1.4	2.2	Hypothetical Protein Slr0953	Cellobiose	342.3	$C_{12}H_{22}O_{11}$	Proteins 2007, 68: 796-801		
A	2B3B	GLC	Kd	9.7	1.95	Glucose-Binding Protein	Alpha-D-Glucose	180.16	$C_6H_{12}O_6$	J Mol Biol 2006, 362: 259-270		A
	2B3F	GAL	Kd	8.2	1.56	Glucose-Binding Protein	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Mol Biol 2006, 362: 259-270		A
XM	2BMZ	XLM	Kd	5.0	2.4	Ripening-Associated Protein	Methyl 3-O-Beta-D-Xylopyranosyl-Alpha-D-Mannopyranoside	326.3	$C_{12}H_{22}O_{10}$	Glycobiology 2005, 15(10): 1033-42 Glycobiology 2005, 15 (10) 1043–50	Four ligand molecules in the "binding interface" between two protein chains	
XMH	2BN0	_LAMI	Kd	4.1	2.8	Ripening-Associated Protein	Laminaribiose	342.3	$C_{12}H_{22}O_{11}$	Glycobiology 2005, 15(10): 1033-42 Glycobiology 2005, 15 (10) 1043–50	GLC+GLC Four ligand molecules in the "binding interface" between two protein chains	
	2BOI	MFU	Ka	6.5	1.1	CV-III Lectin	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	Biochemistry 2006, 45: 7501-10		A

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2BOJ	ARW	Ka	7.9	1.8	Pseudomonas Aeruginosa Lectin Ii	Methyl Beta-D-Arabinopyranoside	164.16	$C_6H_{12}O_5$	FEBS Lett 2006, 580: 982-7		A
XMH	2BS5	_FLAC	Kd	8.7	2.1	Lectin	2'-Alpha-L-Fucosyllactose	488.44	$C_{18}H_{32}O_{15}$	J Biol Chem 2005, 280(30): 27839-49	FUC+LAT Two ligand molecules per protein in different binding sites	
XMH	2BS6	_XXFG	Kd	7.6	1.8	Lectin	Xyloglucan Fragment	458.41	$C_{17}H_{30}O_{14}$	J Biol Chem 2005, 280(30): 27839-49	FUC+GAL+XYS Six ligand molecules per protein in different binding sites	
XM	2BT9	MFU	Kd	8.4	0.94	Lectin	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	J Biol Chem 2005, 280(30): 27839-49	Six ligand molecules per protein in different binding sites	
	2BV4	MMA	Ka	5.1	1	Lectin Cv-Iil	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	Biochemistry 2006, 45: 7501-10		A
	2BVD	ISX	Ki	8.2	1.6	Endoglucanase H	Glucose Beta-1,3-Isofagamine	309.31	$C_{12}H_{23}NO_8$	J Biol Chem 2005, 280(38): 32761-7	Protonated state (NH2+) is used as suggested by Epik and described in article	
	2BVE	PH5	IC50	5.0	2.2	Sialoadhesin	2-Phenyl-Prop5Ac	413.42	$C_{19}H_{27}NO_9$	J Mol Biol 2007, 365: 1469-79		A
	2BZD	GAL	Kd	4.5	2	Bacterial Sialidase	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	Acta Crystallogr Sect. D 2005, 61: 1483-91 J Mol Biol 2003, 327: 659–669.		A
	2CB3	MLD	Kd	10.3	2.4	Peptidoglycan-Recognition Protein-Le	Glcnac(Beta1-4)-Murnac(1,6-Anhydro)-L-Ala-Gamma-D-Glu-Meso-A2Pm-D-Ala	921.91	$C_{37}H_{59}N_7O_{20}$	J Biol Chem 2006, 281(12): 8286-95	Ligand copies in chains B and D deleted	ABD
	2CBJ	OAN	Ki	11.3	2.35	Hyaluronidase	O-(2-Acetamido-2-Deoxy D-Glucopyranosylidene) Amino-N-Phenylcarbamate	353.33	$C_{15}H_{19}N_3O_7$	Embo J 2006, 25 :1569-78		A
	2CCV	A2G	Kd	5.3	1.3	Helix Pomatia Agglutinin	N-Acetyl-2-Deoxy-2-Amino-Galactose	221.21	$C_8H_{15}NO_6$	J Biol Chem 2006, 281: 20171-80		
	2CEX	DAN	Kd	6.4	2.2	Protein Hi0146	2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	291.26	$C_{11}H_{17}NO_8$	J Biol Chem 2006, 281(31): 22212-22		B
	2CHN	NGT	Ki	9.1	1.95	Glucosaminidase	3Ar,5R,6S,7R,7Ar-5-Hydroxymethyl-2-Methyl-5,6,7,7A-Tetrahydro-3Ah-Pyrano[3,2-D]Thiazole-6,7-Diol	219.26	$C_8H_{13}NO_4S$	Nat Struct Mol Biol 2006, 13(4): 365-71		A
	2D2V	MAL	Ki	1.4	2.5	Hypothetical Protein Slr0953	Maltose	342.3	$C_{12}H_{22}O_{11}$	Proteins 2007, 68: 796-801		
XM	2D7F	MMA	IC50	4.7	2.31	Concanavalin A	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	Bmc Struct Biol 2007, 7: 52-52 Bioorg Med Chem Lett 2008, 18: 6573-5	Multiple ligand copies and binding sites	
	2DRI	RIP	Kd	9.4	1.6	D-Ribose-Binding Protein	Ribose(Pyranose Form)	150.13	$C_5H_{10}O_5$	J Biol Chem 1994, 269(48): 30206-11		
W	2E22	MAN	Ki	0.1	2.4	Xanthan Lyase	Alpha-D-Mannose	180.16	$C_6H_{12}O_6$	Biochemistry 2007, 46(3): 781-91	Extremely low affinity (confirmed from article, though)	
	2F2H	XTG	Ki	7.8	1.95	Putative Family 31 Glucosidase Yici	4-Nitrophenyl 6-Thio-6-S-Alpha-D-Xylopyranosyl-Beta-D-Glucopyranoside	449.43	$C_{17}H_{23}NO_{11}S$	J Am Chem Soc 2006, 128(7): 2202-3		A
	2F3P	4GP	Ki	4.3	1.94	Glycogen Phosphorylase, Muscle Form	N-(Beta-D-Glucopyranosyl)Oxamic Acid	251.19	$C_8H_{13}NO_8$	Bioorg Med Chem 2006, 14(11): 3872-82		

Flags	PDB ID	HET ID	Affinity	– ΔG$_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2F3Q	6GP	Ki	5.0	1.96	Glycogen Phosphorylase, Muscle Form	Methyl-N-(Beta-D-Glucopyranosyl)Oxamate	265.22	$C_9H_{15}NO_8$	Bioorg Med Chem 2006, 14(11): 3872-82		
	2F3S	7GP	Ki	4.1	1.96	Glycogen Phosphorylase, Muscle Form	Ethyl-N-(Beta-D-Glucopyranosyl)Oxamate	279.25	$C_{10}H_{17}NO_8$	Bioorg Med Chem 2006, 14(11): 3872-82		
	2F3U	8GP	Ki	3.9	1.93	Glycogen Phosphorylase, Muscle Form	N-(Beta-D-Glucopyranosyl)-N'-Cyclopropyl Oxalamide	290.27	$C_{11}H_{18}N_2O_7$	Bioorg Med Chem 2006, 14(11): 3872-82		
	2F5T	MAL	Kd	7.0	1.45	Archaeal Transcriptional Regulator Trmb	Maltose	342.3	$C_{12}H_{22}O_{11}$	J Biol Chem 2006, 281(16): 10976-82		X
	2FKF	G16	Kd	4.0	2	Phosphomannomutase/Phosphoglucomutase	Alpha-D-Glucose 1,6-Bisphosphate	339.11	$C_6H_{13}O_{12}P_{2-1}$	J Biol Chem 2006, 281(22): 15564-71 Biochemistry 2003, 42: 9946–51	Prime side chain addition didn't complete succesfully Missing side chains far from binding site	
	2GGU	MLR	IC50	4.1	1.9	Pulmonary Surfactant-Associated Protein D	Maltotriose	504.44	$C_{18}H_{32}O_{16}$	J Biol Chem 2006, 281(26): 18008-14		A
	2GGX	NPJ	IC50	4.8	1.9	Pulmonary Surfactant-Associated Protein D	4-Nitrophenyl 4-O-Alpha-D-Glucopyranosyl-Alpha-D-Galactopyranoside	463.39	$C_{18}H_{25}NO_{13}$	J Biol Chem 2006, 281(26): 18008-14		A
	2GPB	GLC	Ki	3.7	2.3	Glycogen Phosphorylase B	Alpha-D-Glucose	180.16	$C_6H_{12}O_6$	Biochemistry 1990, 29(48): 10745-57 Biochemistry 1982, 21: 5364-71		
	2H15	B19	Kd	6.3	1.9	Carbonic Anhydrase 2	N-([(3As,5Ar,8Ar,8Bs)-2,2,7,7-Tetramethyltetrahydro-3Ah-Bis[1,3]Dioxolo[4,5-B:4',5'-D]Pyran-3A-Yl]Methyl)Sulfamide	338.38	$C_{12}H_{22}N_2O_7S$	J Med Chem 2006, 49: 3496-500	A mercury atom (residue 263) far away from the binding site was deleted	
	2H1H	AFH	IC50	6.2	2.4	Lipopolysaccharide Heptosyltransferase 1	Adenosine-5'-Diphosphate-2-Deoxy-2-Fluoro Heptose	621.36	$C_{17}H_{26}FN_5O_{15}P_2$	J Mol Biol 2006, 363: 383-394	Ionized phosphate used (Epik)	A
	2H44	7CA	IC50	7.9	1.8	Cgmp-Specific 3',5'-Cyclic Phosphodiesterase	5,7-Dihydroxy-2-(4-Methoxyphenyl)-8-(3-Methylbutyl)-4-Oxo-4H-Chromen-3-Yl 6-Deoxy-Alpha-L-Mannopyranoside	516.54	$C_{27}H_{32}O_{10}$	J Biol Chem 2006, 281: 21469-79		
	2HL4	BO1	Ki	10.4	1.55	Carbonic Anhydrase 2	N-[4-(Aminosulfonyl)Phenyl]-Beta-D-Glucopyranosylamine	334.34	$C_{12}H_{18}N_2O_7S$	Bioorg Med Chem Lett 2007, 17: 1726-1731 Bioorg Med Chem Lett 2010, 20: 2178-82	A mercury atom (residue 266) far away from the binding site was deleted	
	2IHJ	CSF	Ki	6.3	2	Alpha-2,3/2,6-Sialyltransferase/Sialidase	Cytidine-5'-Monophosphate-3-(A)-Fluoro-N-Acetyl-Neuraminic Acid	632.45	$C_{20}H_{30}FN_4O_{16}P$	Biochemistry 2007, 46: 6288-6298	Tautomric state chosen to match the one reported in the article	
H	2IHK	_CSFE	Ki	6.8	1.9	Alpha-2,3/2,6-Sialyltransferase/Sialidase	Cytidine-5'-Monophosphate-3-(E)-Fluoro-N-Acetyl-Neuraminic Acid	632.45	$C_{20}H_{30}FN_4O_{16}P$	Biochemistry 2007, 46: 6288-6298	the e-(F) isomer of CSF	
XR	2IHZ	CSF	Ki	6.3	2	Alpha-2,3/2,6-Sialyltransferase/Sialidase	Cytidine-5'-Monophosphate-3-(A)-Fluoro-N-Acetyl-Neuraminic Acid	632.45	$C_{20}H_{30}FN_4O_{16}P$	Biochemistry 2007, 46: 6288-6298	redundant, 2IHJ	
	2IXH	TRH	Ka	5.5	2	Dtdp-4-Dehydrorhamnose 3,5-Epimerase	2'-Deoxy-Thymidine-Beta-L-Rhamnose	548.33	$C_{16}H_{26}N_2O_{15}P_2$	J Mol Biol 2007, 365: 146-59		A
	2J0D	ERY	IC50	9.2	2.75	Cytochrome P450 3A4	Erythromycin A	733.93	$C_{37}H_{67}NO_{13}$	J Med Chem 2009, 52: 1180-9		A
	2J1A	GAL	Ka	4.1	1.49	Hyaluronidase	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Biol Chem 2006, 281(49): 37748-57		
H	2J1E	_GAND	Ka	5.5	2.4	Hyaluronidase	N-Acetyl-Lactosamine	383.35	$C_{14}H_{25}NO_{11}$	J Biol Chem 2006, 281(49): 37748-57	NDG+GAL	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2J47	GDV	Ki	7.4	1.98	Glucosaminidase	(5R,6R,7R,8S)-8-(Acetylamino)-6,7-Dihydroxy-5-(Hydroxymethyl)-N-Phenyl-1,5,6,7,8,8A-Hexahydroimidazo[1,2-A]Pyridine-2-Carboxamide	361.37	$C_{17}H_{21}N_4O_5$	Chem Commun 2006, 42: 4372-4		
	2J4G	NB1	Ki	9.0	2.25	Hyaluronoglucosaminidase	(3Ar,5R,6S,7R,7Ar)-5-(Hydroxymethyl)-2-Propyl-5,6,7,7A-Tetrahydro-3Ah-Pyrano[3,2-D][1,3]Thiazole-6,7-Diol	247.31	$C_{10}H_{17}NO_4S$	J Am Chem Soc 2007, 129(3): 635-44		A
	2J62	GSZ	Ki	15.5	2.26	Hyaluronidase	N-[(5R,6R,7R,8S)-6,7-Dihydroxy-5-(Hydroxymethyl)-2-(2-Phenylethyl)-1,5,6,7,8,8A-Hexahydroimidazo[1,2-A]Pyridin-8-Yl]-2-Methylpropanamide	374.45	$C_{20}H_{28}N_3O_4$	J Am Chem Soc 2006, 128(51): 16484-5	Protonated imidazole used (Epik + article)	A
H	2J7M	_BLDH	Ka	4.2	2.3	Hyaluronidase	Blood Group H Type II	529.49	$C_{20}H_{35}NO_{15}$	J Biol Chem 2006, 281(49): 37748-57	NDG+GAL+FUC	
α	2JDM	MFU	Kd	7.4	1.7	Fucose-Binding Lectin PA-IIL	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	BMC Struct Biol 2007, 7: 36	beta-Fucose in chain A deleted	AB
	2JDN	MMA	Kd	7.6	1.3	Fucose-Binding Lectin PA-IIL	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	BMC Struct Biol 2007, 7: 36	Ligand copy in chain A deleted	AB
	2JDP	MFU	Kd	9.0	1.3	Fucose-Binding Lectin PA-IIL	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	BMC Struct Biol 2007, 7: 36	Ligand copy in chain A deleted	AB
	2JDU	MFU	Kd	9.2	1.5	Fucose-Binding Lectin PA-IIL	Alpha-L-Methyl-Fucose	178.18	$C_7H_{14}O_5$	BMC Struct Biol 2007, 7: 36		CD
	2JDY	MMA	Kd	6.0	1.7	Fucose-Binding Lectin PA-IIL	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	BMC Struct Biol 2007, 7: 36	Ligand copy in chain A deleted	AB
	2JF4	VDM	Kd	11.1	2.2	Periplasmic Trehalase	Validoxylamine	335.35	$C_{14}H_{25}NO_8$	Angew Chem Int Ed Engl 2007, 46(22): 4115-9	Neutral state (NH) used (article)	
	2JG0	TTZ	Kd	10.7	1.5	Periplasmic Trehalase	N-[(3As,4R,5S,6S,6As)-4,5,6-Trihydroxy-4-(Hydroxymethyl)-4,5,6,6A-Tetrahydro-3Ah-Cyclopenta[D][1,3]Thiazol-2-Yl]-Alpha-D-Glucopyranosylamine	382.39	$C_{13}H_{22}N_2O_9S$	Angew Chem Int Ed Engl 2007, 46(22): 4115-9	Neutral state (NH) used (article)	
	2JIW	BEU	Ki	6.3	1.95	O-Glcnacase Bt_4395	N-[(1S,2R,5R,6R)-2-Amino-5,6-Dihydroxy-4-(Hydroxymethyl)Cyclohex-3-En-1-Yl]Acetamide	216.23	$C_9H_{16}N_2O_4$	Org Biomol Chem 2007, 5: 3013-19	Neutral state (NH2) used (article)	A
	2JJO	EY5	Kd	7.4	1.99	Cytochrome P450 113A1	(3R,4S,5S,6R,7R,9R,11R,12S,13R,14R)-4-([(2R,4R,5S,6S)-4,5-Dihydroxy-4,6-Dimethyltetrahydro-2H-Pyran-2-Yl]Oxy)-6-([(2S,3R,4S,6R)-4-(Dimethylamino)-3-Hydroxy-6-Methyltetrahydro-2H-Pyran-2-Yl]Oxy)-14-Ethyl-7,12-Dihydroxy-3,5,7,9,11,13-Hexamethyloxacyclotetradecane-2,10-Dione	703.9	$C_{36}H_{65}NO_{12}$	J Biol Chem 2009, 284(42): 29170-9	Neutral state (tertiary-N) used (article)	
	2JLB	UDM	Kd	7.3	2.5	Xcc0866	Uridine-Diphosphate-Methylene-N-Acetyl-Glucosamine	605.39	$C_{18}H_{29}N_3O_{16}P_2$	Embo J 2008, 27: 2780-8	Ligand copy in chain B deleted	AB

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
H	2NMO	LAT	Kd	5.0	1.35	Galectin-3	Beta-Lactose	342.3	$C_{12}H_{22}O_{11}$	Acta Crystallogr, Sect.D 2007, 63: 415-419 Bioorg Med Chem Lett 2005, 15: 2343-5	GAL+BGC Glycerol molecule (GOL) overlapping with ligand, deleted	
	2O9R	TCB	Ki	2.3	2.3	Beta-Glucosidase B	Thiocellobiose	358.36	$C_{12}H_{22}O_{10}S$	J Mol Biol 2007, 371(5): 1204-18		
	2OYK	9MR	Ki	7.2	1.5	Endoglycoceramidase II	(3R,4R,5R)-3-Hydroxy-5-(Hydroxymethyl)Piperidin-4-Yl Beta-D-Glucopyranoside	309.32	$C_{12}H_{23}NO_8$	Angew Chem Int Ed Engl 2007, 46(24): 4474-6	Neutral state used (article)	A
	2OYL	IDC	Ki	8.6	1.8	Endoglycoceramidase II	5-Hydroxymethyl-5,6,7,8-Tetrahydro-Imidazo[1,2-A]Pyridin-6Yl-7,8-Diol-Glucopyranoside	362.34	$C_{14}H_{22}N_2O_9$	Angew Chem Int Ed Engl 2007, 46(24): 4474-6	Neutral state used (article)	A
	2OYM	MNI	Ki	6.8	1.86	Endoglycoceramidase II	1-(4-Dimethylamino)Benzoylamino-1,2,5-Trideoxy-2,5-Imino-D -Mannitol	309.36	$C_{15}H_{23}N_3O_4$	Angew Chem Int Ed Engl 2007, 46(24): 4474-6	Neutral state used (article)	A
	2PRI	D6G	Ki	4.0	2.3	Glycogen Phosphorylase B	2-Deoxy-Glucose-6-Phosphate	244.14	$C_6H_{13}O_8P$	J Mol Biol 1995, 254(5): 900-17		
	2PRJ	NBG	Ki	6.1	2.3	Glycogen Phosphorylase	1-N-Acetyl-Beta-D-Glucosamine	221.21	$C_8H_{15}NO_6$	Protein Sci 1995, 4(12): 2469-77		
	2PYD	GLC	Ki	3.8	1.93	Glycogen Phosphorylase, Muscle Form	Alpha-D-Glucose	180.16	$C_6H_{12}O_6$	Proteins 2007, 71(3): 1307-23		
	2PYI	DL8	Ki	5.1	1.88	Glycogen Phosphorylase, Muscle Form	N-[(4-Phenyl-1H-1,2,3-Triazol-1-Yl)Acetyl]-Beta-D-Glucopyranosylamine	364.36	$C_{16}H_{20}N_4O_6$	Proteins 2007, 71(3): 1307-23		
	2QLM	F68	Ki	7.7	2.1	Glycogen Phosphorylase, Muscle Form	N-([(4-Methylphenyl)Carbonyl]Carbamoyl)-Beta-D-Glucopyranosylamine	340.33	$C_{15}H_{20}N_2O_7$	Bioorg Med Chem 2009, 17: 4773-85		
	2QMJ	ACR	Ki	5.7	1.9	Maltase-Glucoamylase, Intestinal	Alpha-Acarbose	645.61	$C_{25}H_{43}NO_{18}$	J Mol Biol 2008, 375(3): 782-92	Neutral state used (article)	
	2QN7	HBZ	Ki	7.1	1.83	Glycogen Phosphorylase, Muscle Form	N-([(4-Hydroxyphenyl)Carbonyl]Carbamoyl)-Beta-D-Glucopyranosylamine	342.3	$C_{14}H_{18}N_2O_8$	Bioorg Med Chem 2009, 17: 4773-85		
	2QN8	NBY	Ki	7.5	1.9	Glycogen Phosphorylase, Muscle Form	N-([(4-Nitrophenyl)Carbonyl]Carbamoyl)-Beta-D-Glucopyranosylamine	371.3	$C_{14}H_{17}N_3O_9$	Bioorg Med Chem 2009, 17: 4773-85	Two ligand copies, catalytic and allosteric (article) Affinity measurement was competitive, so allosteric copy deleted	
	2QN9	NBX	Ki	7.1	2	Glycogen Phosphorylase, Muscle Form	N-([(4-Aminophenyl)Carbonyl]Carbamoyl)-Beta-D-Glucopyranosylamine	341.32	$C_{14}H_{19}N_3O_7$	Bioorg Med Chem 2009, 17: 4773-85		
	2QNB	BZD	Ki	7.3	1.8	Glycogen Phosphorylase, Muscle Form	N-Benzoyl-N'-Beta-D-Glucopyranosyl Urea	326.3	$C_{14}H_{18}N_2O_7$	Bioorg Med Chem 2009, 17: 4773-85	Two ligand copies, catalytic and allosteric (article) Affinity measurement was competitive, so allosteric copy deleted	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2QRG	M07	Ki	7.1	1.85	Glycogen Phosphorylase, Muscle Form	(5R,7R,8S,9S,10R)-7-(Hydroxymethyl)-3-(4-Methoxyphenyl)-1,6-Dioxa-2-Azaspiro[4.5]Dec-2-Ene-8,9,10-Triol	325.32	$C_{15}H_{19}NO_7$	Bioorg Med Chem 2009, 17: 7368-80		
	2QRH	M08	Ki	6.4	1.83	Glycogen Phosphorylase, Muscle Form	(5R,7R,8S,9S,10R)-7-(Hydroxymethyl)-3-Phenyl-1,6-Dioxa-2-Azaspiro[4.5]Dec-2-Ene-8,9,10-Triol	295.29	$C_{14}H_{17}NO_6$	Bioorg Med Chem 2009, 17: 7368-80		
	2QRM	M09	Ki	5.5	1.9	Glycogen Phosphorylase, Muscle Form	(1R)-3'-(4-Nitrophenyl)-Spiro[1,5-Anhydro-D-Glucitol-1,5'-Isoxazoline]	342.3	$C_{14}H_{18}N_2O_8$	Bioorg Med Chem 2009, 17: 7368-80		
	2QRP	S06	Ki	8.5	1.86	Glycogen Phosphorylase, Muscle Form	1R)-3'-(2-Naphthyl)-Spiro[1,5-Anhydro-D-Glucitol-1,5'-Isoxazoline]	347.36	$C_{18}H_{21}NO_6$	Bioorg Med Chem 2009, 17: 7368-80		
	2QRQ	S13	Ki	7.0	1.8	Glycogen Phosphorylase, Muscle Form	(1R)-3'-(4-Methylphenyl)-Spiro[1,5-Anhydro-D-Glucitol-1,5'-Isoxazoline]	311.33	$C_{15}H_{21}NO_6$	Bioorg Med Chem 2009, 17: 7368-80		
	2QWB	SIA	Ki	3.7	2	Neuraminidase, R292K Mutant	O-Sialic Acid	309.27	$C_{11}H_{19}NO_9$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462	An extra SIA molecule in a secondary site probably under crystal conditions (article), deleted	
	2QWC	DAN	Ki	4.8	1.6	Neuraminidase, R292K Mutant	2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	291.26	$C_{11}H_{17}NO_8$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWD	4AM	Ki	6.6	2	Neuraminidase, R292K Mutant	4-Amino-2-Deoxy-2,3-Dehydro-N-Neuraminic Acid	290.27	$C_{11}H_{18}N_2O_7$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462	Neutral state (NH2) used (article)	
	2QWE	GNA	Ki	10.2	2	Neuraminidase, R292K Mutant	2,4-Deoxy-4-Guanidino-5-N-Acetyl-Neuraminic Acid	334.33	$C_{12}H_{22}N_4O_7$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWF	G20	Ki	7.7	1.9	Neuraminidase, R292K Mutant	4-Acetyl-4-Guanidino-6-Methyl(Propyl)Carboxamide-4,5-Dihydro-2H-Pyran-2-Carboxylic Acid	341.37	$C_{14}H_{23}N_5O_5$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWG	G28	IC50	5.0	1.8	Neuraminidase, R292K Mutant	5-N-Acetyl-4-Amino-6-Diethylcarboxamide-4,5-Dihydro-2H-Pyran-2-Carboxylic Acid	301.34	$C_{13}H_{23}N_3O_5$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWH	G39	IC50	6.7	1.8	Neuraminidase, R292K Mutant	(3R,4R,5S)-4-(Acetylamino)-5-Amino-3-(Pentan-3-Yloxy)Cyclohex-1-Ene-1-Carboxylic Acid	284.35	$C_{14}H_{24}N_2O_4$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWI	G20	Ki	11.4	2	Neuraminidase, Wild-Type	4-Acetyl-4-Guanidino-6-Methyl(Propyl)Carboxamide-4,5-Dihydro-2H-Pyran-2-Carboxylic Acid	341.37	$C_{14}H_{23}N_5O_5$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2QWJ	G28	IC50	9.1	2	Neuraminidase, Wild-Type	5-N-Acetyl-4-Amino-6-Diethylcarboxamide-4,5-Dihydro-2H-Pyran-2-Carboxylic Acid	301.34	$C_{13}H_{23}N_3O_5$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462		
	2QWK	G39	IC50	11.9	1.8	Neuraminidase, Wild-Type	(3R,4R,5S)-4-(Acetylamino)-5-Amino-3-(Pentan-3-Yloxy)Cyclohex-1-Ene-1-Carboxylic Acid	284.35	$C_{14}H_{24}N_2O_4$	Structure 1998, 6: 735-46 J Virol 1998, 72: 2456-2462	Neutral state (NH2) used (article)	
	2R0H	CTO	Ka	4.8	1.9	Cgl3 Lectin	Triacetylchitotriose	627.6	$C_{24}H_{41}N_3O_{16}$	J Mol Biol 2008, 379(1): 146-59		A
	2RFY	CBI	Kd	7.1	1.7	Cellulose 1,4-Beta-Cellobiosidase	Cellobiose	342.3	$C_{12}H_{22}O_{11}$	Protein Sci 2008, 17(8): 1383-94		A
H	2RI9	_LYM	Ki	4.4	1.95	Mannosyl-Oligosaccharide Alpha-1,2-Mannosidase	Methyl-A-D-Lyxopyranosyl-(1' ,2)-A-D-Mannopyranoside	326.3	$C_{12}H_{22}O_{10}$	Acta Crystallogr D Biol Crystallogr 2008, 64(Pt 3): 227-36	LDY+MMA Glycerol molecule (GOL) overlapping with ligand, deleted	A
	2RIA	289	IC50	3.5	1.8	Pulmonary Surfactant-Associated Protein D	D-Glycero-Alpha-D-Manno-Heptopyranose	210.18	$C_7H_{14}O_7$	Biochemistry 2008, 47(2): 710-20		A
	2RIB	GMH	IC50	3.8	1.8	Pulmonary Surfactant-Associated Protein D	L-Glycero-D-Manno-Heptopyranose	210.18	$C_7H_{14}O_7$	Biochemistry 2008, 47(2):710-20.		A
	2SIM	DAN	Ki	4.7	1.6	Sialidase	2-Deoxy-2,3-Dehydro-N-Acetyl-Neuraminic Acid	291.26	$C_{11}H_{17}NO_8$	J Biochem (Tokyo) 1991, 110(3): 462 J Biol Chem 2000, 275: 39385		
H	2UVH	_DADA	Ka	6.5	2.2	Abc Type Periplasmic Sugar-Binding Protein	Di-Galactouronic Acid	370.26	$C_{12}H_{18}O_{13}$	J Mol Biol 2007, 369(3): 759-70	ADA+ADA	
	2UVI	UNG	Ka	7.2	2.3	Abc Type Periplasmic Sugar-Binding Protein	4-O-(4-Deoxy-Beta-L-Threo-Hex-4-Enopyranuronosyl)-Alpha-D-Galactopyranuronic Acid	352.25	$C_{12}H_{16}O_{12}$	J Mol Biol 2007, 369(3): 759-70		
H	2UVJ	_TADA	Ka	5.1	1.8	Abc Type Periplasmic Sugar-Binding Protein	Tri-Galactouronic Acid	546.39	$C_{18}H_{26}O_{19}$	J Mol Biol 2007, 369(3): 759-70	ADA+ADA+ADA	
	2V4V	XYP	Ka	4.4	1.5	Gh59 Galactosidase	Beta-D-Xylopyranose	150.13	$C_5H_{10}O_5$	Biochemistry 2009, 48(43): 10395-404		
	2V72	GAL	Ka	4.3	2.25	Exo-Alpha-Sialidase	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	Biochemistry 2007, 46(40): 11352-60		
	2VEZ	G6P	Ki	3.5	1.45	Putative Glucosamine 6-Phosphate Acetyltransferase	Alpha-D-Glucose-6-Phosphate	260.14	$C_6H_{13}O_9P$	FEBS Lett 2007, 581: 5597-5600		
	2VFZ	UPF	Ki	4.9	2.4	N-Acetyllactosaminide Alpha-1,3-Galactosyl Transferase	Uridine-5'-Diphosphate-2-Deoxy-2-Fluorogalactose	568.3	$C_{15}H_{23}FN_2O_{16}P_2$	J Mol Biol 2007, 369(5): 1270-81		A
α	2VMC	A2G	Kd	4.0	1.9	Discoidin-2	N-Acetyl-2-Deoxy-2-Amino-Galactose	221.21	$C_8H_{15}NO_6$	Proteins 2008, 73(1): 43-52		
	2VMD	MBG	Kd	4.1	1.9	Discoidin-2	Methyl-Beta-Galactose	194.18	$C_7H_{14}O_6$	Proteins 2008, 73(1): 43-52		
	2VMG	MBG	Ka	3.7	1.9	Fibronectin Type Iii Domain Protein	Methyl-Beta-Galactose	194.18	$C_7H_{14}O_6$	J Biol Chem 2008, 283(18): 12604-13		
	2VNV	MMA	Kd	7.6	1.7	Bcla	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	Biochem J 2008, 411(2): 307-18	Ligand copy in chain A deleted	AB
	2VUR	YX1	IC50	6.2	2.2	O-Glcnacase Nagj	2-Deoxy-2-([(2-Hydroxy-1-Methylhydrazino)Carbonyl]Amino)-Beta-D-Glucopyranose	267.24	$C_8H_{17}N_3O_7$	Chem Biol 2008, 15(8): 799-807		A

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	2VVN	NHT	Kd	10.0	1.85	O-Glcnacase Bt_4395	(3Ar,5R,6S,7R,7Ar)-2-(Ethylamino)-5-(Hydroxymethyl)-5,6,7,7A-Tetrahydro-3Ah-Pyrano[3,2-D][1,3]Thiazole-6,7-Diol	248.3	$C_9H_{16}N_2O_4S$	Nat Chem Biol 2008, 4(8): 483-90		A
	2VVO	A6P	IC50	3.7	1.85	Ribose-5-Phosphate Isomerase B	6-O-Phosphono-Alpha-D-Allopyranose	260.14	$C_6H_{13}O_9P$	J Mol Biol 2008, 382(3): 667-79	Ligand copy in chain B deleted	AB
	2VVS	OAN	Ki	10.0	2.24	O-Glcnacase Bt_4395	O-(2-Acetamido-2-Deoxy D-Glucopyranosylidene) Amino-N-Phenylcarbamate	353.33	$C_{15}H_{19}N_3O_7$	J Biol Chem 2008, 283(50): 34687-95		
	2VZR	GCU	Ka	5.8	1.95	Exo-Beta-D-Glucosaminidase	D-Glucuronic Acid	194.14	$C_6H_{10}O_7$	Proc Natl Acad Sci USA, 2009, 106(9): 3065-70		A
	2W4X	STZ	Ki	6.6	2.42	O-Glcnacase Bt_4395	Streptozotocin	265.22	$C_8H_{15}N_3O_7$	Carbohydr Res 2009, 344(5): 627-31		
	2WCV	FUC	Kd	3.8	1.9	L-Fucose Mutarotase	Alpha-L-Fucose	164.16	$C_6H_{12}O_5$	J Mol Biol 2009, 391(1): 178-91	Ligand copies in chains E and H deleted	AE H
H	2XG3	_BNAL	Kd	6.5	1.2	Galectin-3	3′-Benzamido-N-Acetyllactosamine	486.47	$C_{21}H_{30}N_2O_{11}$	J Am Chem Soc 2010, 132(41): 14577-89	UNU+GAL+NAG	
XO	2Z65	E55	IC50	12.0	2.7	Lymphocyte Antigen 96	3-O-Decyl-2-Deoxy-6-O-(2-Deoxy-3-O-[(3R)-3-Methoxydecyl]-6-O-Methyl-2-[(11Z)-Octadec-11-Enoylamino]-4-O-Phosphono-Beta-D-Glucopyranosyl)-2-[(3-Oxotetradecanoyl)Amino]-1-O-Phosphono-Alpha-D-Glucopyranose	1313.67	$C_{66}H_{126}N_2O_{19}P_2$	Cell 2007, 130(5): 906-17 J Med Chem 2008, 51: 6621-6	Very big ligand, mostly aglycone	AC
XM	3A22	ARA	Ka	3.0	1.9	Putative Secreted Alpha-Galactosidase	Alpha-L-Arabinose	150.13	$C_5H_{10}O_5$	J Biol Chem 2009, 284(37): 25097-106 Biochem J 2000, 350(3): 933-41	Multiple ligand copies per protein chain	
XM	3A23	GAL	Ka	3.8	1.9	Putative Secreted Alpha-Galactosidase	Beta-D-Galactose	180.16	$C_6H_{12}O_6$	J Biol Chem 2009, 284(37): 25097-106 Biochem J 2000, 350(3): 933-41	Multiple ligand copies per protein chain	
	3B50	SLB	Kd	10.3	1.4	Sialic Acid-Binding Periplasmic Protein Siap	5-N-Acetyl-Beta-D-Neuraminic Acid	309.27	$C_{11}H_{19}NO_9$	J Biol Chem 2008, 283(2): 855-65		
	3BCS	CJB	Ki	7.1	2	Glycogen Phosphorylase, Muscle Form	1-Beta-D-Glucopyranosylpyrimidine-2,4(1H,3H)-Dione	274.23	$C_{10}H_{14}N_2O_7$	Bioorg Med Chem 2010, 18: 3413-25 Curr Med Chem 2008, 15: 2933-2983		
	3BD8	C3B	Ki	7.0	2.1	Glycogen Phosphorylase, Muscle Form	4-Amino-1-Beta-D-Glucopyranosylpyrimidin-2(1H)-One	273.24	$C_{10}H_{15}N_3O_6$	Bioorg Med Chem 2010, 18: 3413-25 Curr Med Chem 2008, 15: 2933-2983		
	3BXF	FBP	Kd	7.6	1.7	Central Glycolytic Gene Regulator	Beta-Fructose-1,6-Diphosphate	340.12	$C_6H_{14}O_{12}P_2$	Mol Microbiol 2008, 69(4): 895-910	chain B has a different ligand	A
	3BXG	BG6	Kd	6.8	1.8	Central Glycolytic Gene Regulator	Beta-D-Glucose-6-Phosphate	260.14	$C_6H_{13}O_9P$	Mol Microbiol 2008, 69(4): 895-910		A
	3BXH	F6P	Kd	5.5	1.85	Central Glycolytic Gene Regulator	Fructose-6-Phosphate	260.14	$C_6H_{13}O_9P$	Mol Microbiol 2008, 69(4): 895-910		A

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
WM	3DCQ	2G0	IC50	7.1	1.8	Fucose-Binding Lectin PA-IIL	(2S)-1-[(2S)-6-Amino-2-(([(2S,3S,4R,5S,6S)-3,4,5-Trihydroxy-6-Methyltetrahydro-2H-Pyran-2-Yl]Acetyl)Amino)Hexanoyl]-N-[(1S)-1-Carbamoyl-3-Methylbutyl]Pyrrolidine-2-Carboxamide	543.65	$C_{25}H_{45}N_5O_8$	Chem Biol 2008, 15(12): 1249-57	Missing a flexible tail, solvent exposed and not interacting with binding site residues Ligand copy in chain B deleted	AB
	3DJE	FSA	Ki	7.2	1.6	Fructosyl Amine: Oxygen Oxidoreductase	1-S-(Carboxymethyl)-1-Thio-Beta-D-Fructopyranose	254.25	$C_8H_{14}O_7S$	J Biol Chem 2008, 283(40): 27007–27016		A
	3DWB	RDF	IC50	8.4	2.38	Endothelin-Converting Enzyme 1	N-Alpha-L-Rhamnopyranosyloxy(Hydroxyphosphinyl)-L-Leucyl-L-Tryptophan	543.51	$C_{23}H_{34}N_3O_{10}P$	J Mol Biol 2009, 385(1): 178-87		
	3E6Y	CW1	Kd	7.6	2.5	14-3-3-Like Protein C	Cotylenin A	652.78	$C_{34}H_{52}O_{12}$	J Mol Biol. 2009, 386(4): 913-9		A
	3F8F	DM1	Kd	9.0	2.2	Transcriptional Regulator, Padr-Like Family	Daunomycin	527.53	$C_{27}H_{29}NO_{10}$	EMBO J 2009, 28(2): 156-66	Neutral state (NH2) used (article)	AB
	3G2H	KOT	Ki	5.2	2.03	Glycogen Phosphorylase, Muscle Form	1-Beta-D-Glucopyranosyl-4-Phenyl-1H-1,2,3-Triazole	307.3	$C_{14}H_{17}N_3O_5$	Bioorg Med Chem 2010, 18: 1171-80		
	3G2I	RUG	Ki	6.3	2	Glycogen Phosphorylase, Muscle Form	1-Beta-D-Glucopyranosyl-4-(Hydroxymethyl)-1H-1,2,3-Triazole	261.23	$C_9H_{15}N_3O_6$	Bioorg Med Chem 2010, 18: 1171-80		
	3G2J	9GP	Ki	6.4	2.14	Glycogen Phosphorylase, Muscle Form	N-(Hydroxyacetyl)-Beta-D-Glucopyranosylamine	237.21	$C_8H_{15}NO_7$	Bioorg Med Chem 2010, 18: 1171-80		
	3G2K	SKY	Ki	6.1	2	Glycogen Phosphorylase, Muscle Form	1-Beta-D-Glucopyranosyl-4-Naphthalen-2-Yl-1H-1,2,3-Triazole	357.36	$C_{18}H_{19}N_3O_5$	Bioorg Med Chem 2010, 18: 1171-80		
	3G2L	LEW	Ki	5.3	2.3	Glycogen Phosphorylase, Muscle Form	1-Beta-D-Glucopyranosyl-4-Naphthalen-1-Yl-1H-1,2,3-Triazole	357.36	$C_{18}H_{19}N_3O_5$	Bioorg Med Chem 2010, 18: 1171-80		
	3G2N	OAK	Ki	5.2	2.1	Glycogen Phosphorylase, Muscle Form	N-(Phenylcarbonyl)-Beta-D-Glucopyranosylamine	283.28	$C_{13}H_{17}NO_6$	Bioorg Med Chem 2010, 18: 1171-80		
	3GA5	RGG	Kd	7.5	1.87	D-Galactose-Binding Periplasmic Protein	(2R)-2,3-Dihydroxypropyl Beta-D-Galactopyranoside	254.24	$C_9H_{18}O_8$	FEBS J 2009, 276(7): 2116-24 Eur J Biochem 1969, 10: 66–73		A
	3GF4	UPG	Kd	4.3	2.45	Udp-Galactopyranose Mutase	Uridine-5'-Diphosphate-Glucose	566.3	$C_{15}H_{24}N_2O_{17}P_2$	J Mol Biol 2009, 391: 327-40		A
	3GPB	G1P	Kd	2.8	2.3	Glycogen Phosphorylase B	Alpha-D-Glucose-1-Phosphate	260.14	$C_6H_{13}O_9P$	Biochemistry 1990, 29(48): 10745-57 Mol Cell Biochem 1976, 11: 35-50	Two ligand copies, catalytic and allosteric (article) Affinity measurement was competitive, so allosteric copy deleted	
XM	3H2K	BOG	Kd	5.6	2.1	Esterase	B-Octylglucoside	292.37	$C_{14}H_{28}O_6$	Plant Cell 2009, 21(6): 1860-73	Two very close ligand molecules, one in the catalytic site and the other very close interacting with the former	
	3HDQ	GDU	Kd	5.0	2.36	UDP-Galactopyranose Mutase	Galactose-Uridine-5'-Diphosphate	566.3	$C_{15}H_{24}N_2O_{17}P_2$	J Mol Biol 2009, 394: 864-77	Protonation state from Epik (oxidized form of the protein)	A
	3HDY	GDU	Kd	5.7	2.4	UDP-Galactopyranose Mutase	Galactose-Uridine-5'-Diphosphate	566.3	$C_{15}H_{24}N_2O_{17}P_2$	J Mol Biol 2009, 394: 864-77	Protonation state from Epik (reduced form of the protein, cf. 3HDQ)	

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	3HKN	MFS	Ki	7.3	1.8	Carbonic Anhydrase 2	(1S)-2,3,6-Tri-O-Acetyl-1,5-Anhydro-1-Sulfamoyl-4-O-(2,3,4,6-Tetra-O-Acetyl-Beta-D-Galactopyranosyl)-D-Glucitol	699.63	$C_{26}H_{37}NO_{19}S$	J Med Chem 2009, 52: 6421-32		
	3HKQ	1SD	Ki	7.3	1.7	Carbonic Anhydrase 2	(2S,3R,4S,5R,6R)-3,4,5-Trihydroxy-6-(Hydroxymethyl)Oxane-2-Sulfonamide	243.23	$C_6H_{13}NO_7S$	J Med Chem 2009, 52: 6421-32		
	3HKT	2SD	Ki	7.3	2.36	Carbonic Anhydrase 2	(1S)-1,5-Anhydro-4-O-Alpha-D-Galactopyranosyl-1-Sulfamoyl-D-Galactitol	405.38	$C_{12}H_{23}NO_{12}S$	J Med Chem 2009, 52: 6421-32		
	3HKU	TOR	Ki	11.3	1.8	Carbonic Anhydrase 2	Topiramate	339.36	$C_{12}H_{21}NO_8S$	J Med Chem 2009, 52: 6421-32		
XM	3HP8	SUC	Kd	2.9	2	Cyanovirin-N-Like Protein	Sucrose	342.3	$C_{12}H_{22}O_{11}$	Proteins 2009, 77(4): 904-15	Two ligand molecules per protein chain, in on-identical binding sites	
XM	3HP8	SUC	Kd	2.5	2	Cyanovirin-N-Like Protein	Sucrose	342.3	$C_{12}H_{22}O_{11}$	Proteins 2009, 77(4): 904-15	Two ligand molecules per protein chain, in on-identical binding sites	
	3IJH	KO2	Kd	5.5	2.1	Immunoglobulin Heavy Chain (Igg3)	Prop-2-En-1-Yl D-Glycero-Alpha-D-Talo-Oct-2-Ulopyranosidonic Acid	294.26	$C_{11}H_{18}O_9$	Glycobiology 2010, 20(2): 138-47		AB
H	3IJY	_KDAO	Kd	8.2	2.85	Immunoglobulin Heavy Chain (Igg3)	Kdo(2→8)Kdo	498.44	$C_{19}H_{30}O_{15}$	Glycobiology 2010, 20(2): 138-47	KDA+KDO	AB
H	3IKC	_KDKM	Kd	10.2	2.6	Immunoglobulin Heavy Chain (Igg3)	Kdo(2→8)7-O-Me-Kdo	512.46	$C_{20}H_{32}O_{15}$	Glycobiology 2010, 20(2): 138-47	KDO+KME	AB
	3L79	DKX	Ki	3.4	1.86	Glycogen Phosphorylase, Muscle Form	1-(3-Deoxy-3-Fluoro-Beta-D-Glucopyranosyl)Pyrimidine-2,4(1H,3H)-Dione	276.22	$C_{10}H_{13}FN_2O_6$	Bioorg Med Chem 2010, 18: 3413-3425		
	3L7A	DKY	Ki	5.9	1.9	Glycogen Phosphorylase, Muscle Form	1-(3-Deoxy-3-Fluoro-Beta-D-Glucopyranosyl)-4-[(Phenylcarbonyl)Amino]Pyrimidin-2(1H)-One	379.34	$C_{17}H_{18}FN_3O_6$	Bioorg Med Chem 2010, 18: 3413-3425		
	3L7B	DKZ	Ki	3.3	2	Glycogen Phosphorylase, Muscle Form	4-Amino-1-(3-Deoxy-3-Fluoro-Beta-D-Glucopyranosyl)Pyrimidin-2(1H)-One	275.23	$C_{10}H_{14}FN_3O_5$	Bioorg Med Chem 2010, 18: 3413-3425		
	3L7C	DK4	Ki	3.3	1.93	Glycogen Phosphorylase, Muscle Form	1-(3-Deoxy-3-Fluoro-Beta-D-Glucopyranosyl)-5-Fluoropyrimidine-2,4(1H,3H)-Dione	294.21	$C_{10}H_{12}F_2N_2O_6$	Bioorg Med Chem 2010, 18: 3413-3425		
	3L7D	DK5	Ki	3.0	2	Glycogen Phosphorylase, Muscle Form	1-(2,3-Dideoxy-3-Fluoro-Beta-D-Arabino-Hexopyranosyl)-4-[(Phenylcarbonyl)Amino]Pyrimidin-2(1H)-One	363.34	$C_{17}H_{18}FN_3O_5$	Bioorg Med Chem 2010, 18: 3413-3425		
	3LXE	TOR	Ki	9.0	1.9	Carbonic Anhydrase 1	Topiramate	339.36	$C_{12}H_{21}NO_8S$	Org Biomol Chem 2010, 8: 3528–33		A
H	3MBP	MLR	Kd	9.3	1.7	Maltodextrin-Binding Protein	Maltotriose	504.44	$C_{18}H_{32}O_{16}$	Structure 1997, 5(8): 997-1015	GLC+GLC+GLC	
H	3OY8	_LBA	Kd	5.0	2.19	Galectin-1	Lactobionic Acid	358.3	$C_{12}H_{22}O_{12}$	Cancer Letters 2010, 299(2): 95–110	GAL+GCO	A

Flags	PDB ID	HET ID	Affinity	$-\Delta G_{exp}$	Res. (Å)	Protein	Ligand Name	Mol. Wt.	Formula	Reference	Comments	Chain(s)
	3OYW	TDG	Kd	5.6	2.5	Galectin-1	Thiodigalactoside	358.36	$C_{12}H_{22}O_{10}S$	Cancer Letters 2010, 299(2): 95–110		A
H	4MBP	MTT	Kd	7.7	1.7	Maltodextrin Binding Protein	Maltotetraose	666.58	$C_{24}H_{42}O_{21}$	Structure 1997, 5(8): 997-1015	GLC+GLC+GLC+GLC	
A	5ABP	GLA	Kd	9.1	1.8	L-Arabinose-Binding Protein	Alpha D-Galactose	180.16	$C_6H_{12}O_6$	Nature 1989, 340(6232): 404-7		
	5CNA	MMA	IC50	4.7	2	Concanavalin A	O1-Methyl-Mannose	194.18	$C_7H_{14}O_6$	Acta Crystallogr Sect D 1994, 50: 847-858 Bioorg Med Chem Lett 2008, 18: 6573-5		A
α	6ABP	ARA	Kd	8.7	1.67	L-Arabinose-Binding Protein	Alpha-L-Arabinose	150.13	$C_5H_{10}O_5$	Biochemistry 1991, 30(28): 6861-6		
α	7ABP	FCA	Kd	8.8	1.67	L-Arabinose-Binding Protein	Alpha-D-Fucose	164.16	$C_6H_{12}O_5$	Biochemistry 1991, 30(28): 6861-6		
	8A3H	IDC	Ki	5.5	0.97	Protein (Endoglucanase)	5-Hydroxymethyl-5,6,7,8-Tetrahydro-Imidazo[1,2-A]Pyridin-6Yl-7,8-Diol-Glucopyranoside	362.34	$C_{14}H_{22}N_2O_9$	J Am Chem Soc 1999, 121: 2621-22		
α	8ABP	GLA	Kd	10.9	1.49	L-Arabinose-Binding Protein	Alpha D-Galactose	180.16	$C_6H_{12}O_6$	Biochemistry 1991, 30(28): 6861-6		
XR	9ABP	GLA	Kd	9.1	1.97	L-Arabinose-Binding Protein	Alpha D-Galactose	180.16	$C_6H_{12}O_6$	Biochemistry 1991, 30(28): 6861-6	redundant, 5ABP	

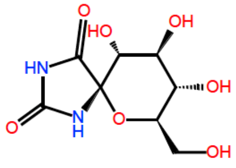
Appendix 2: Structures of carbohydrate ligands in the studied complexes

Ligands are grouped according to molecular weight:

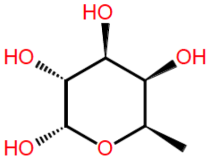
- **Group 1**: Mol. Wt. ≤ 370
- **Group 2**: 370 < Mol. Wt. ≤ 700
- **Group 3**: Mol. Wt. > 700

Within each group, ligands are arranged alphabetically according to the corresponding PDB code.

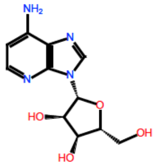
Group 1



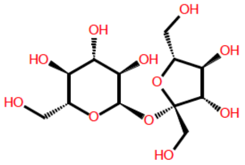
PDB: 1A8I
HET: GLS
Mol.Wt.: 248.194



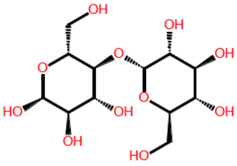
PDB: 1ABF
HET: FCA
Mol.Wt.: 164.16



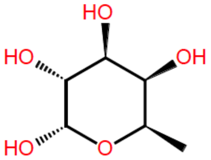
PDB: 1ADD
HET: 1DA
Mol.Wt.: 266.259



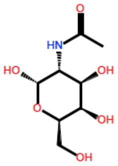
PDB: 1AF6
HET: SUC
Mol.Wt.: 342.303



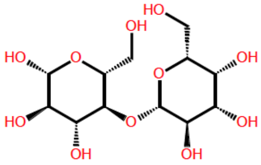
PDB: 1ANF
HET: MAL
Mol.Wt.: 342.303



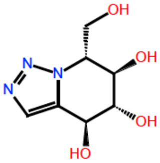
PDB: 1APB
HET: FCA
Mol.Wt.: 164.16



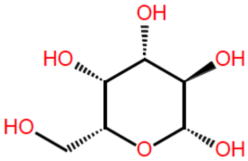
PDB: 1AX0
HET: A2G
Mol.Wt.: 221.212



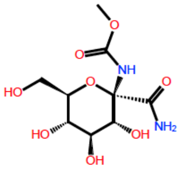
PDB: 1AX1
HET: LAT
Mol.Wt.: 342.303



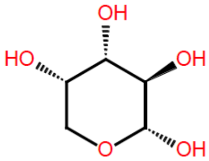
PDB: 1AXR
HET: HTP
Mol.Wt.: 201.183



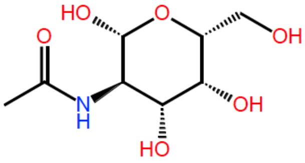
PDB: 1AXZ
HET: GAL
Mol.Wt.: 180.159



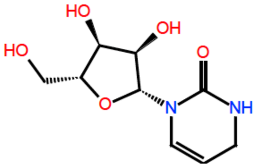
PDB: 1B4D
HET: CRA
Mol.Wt.: 280.236



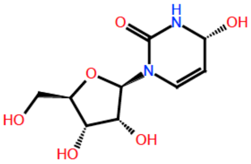
PDB: 1BAP
HET: ARA
Mol.Wt.: 150.132



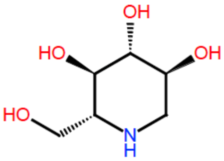
PDB: 1BCH
HET: NGA
Mol.Wt.: 221.212



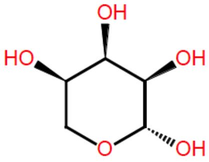
PDB: 1CTT
HET: DHZ
Mol.Wt.: 230.222



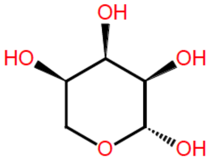
PDB: 1CTU
HET: ZEB
Mol.Wt.: 246.222



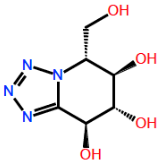
PDB: 1DOG
HET: NOJ
Mol.Wt.: 163.175



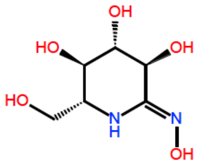
PDB: 1DRJ
HET: RIP
Mol.Wt.: 150.132



PDB: 1DRK
HET: RIP
Mol.Wt.: 150.132

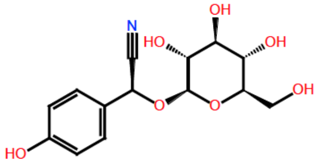


PDB: 1E6Q
HET: NTZ
Mol.Wt.: 202.171

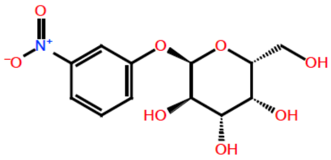


PDB: 1E6S
HET: GOX
Mol.Wt.: 192.173

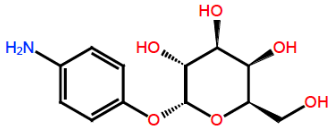
Group 1 (continued)



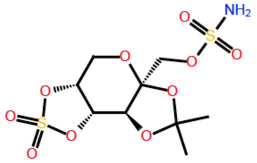
PDB: 1E55
HET: _DHR
Mol.Wt.: 311.294



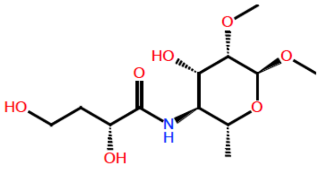
PDB: 1EEI
HET: GAA
Mol.Wt.: 301.255



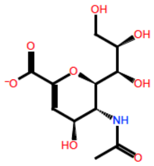
PDB: 1EFI
HET: GAT
Mol.Wt.: 271.272



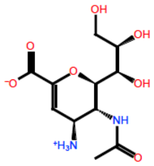
PDB: 1EOU
HET: SMS
Mol.Wt.: 361.349



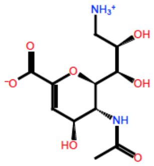
PDB: 1F4X
HET: MGS
Mol.Wt.: 293.32



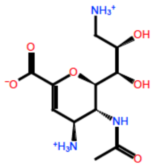
PDB: 1F8B
HET: DAN
Mol.Wt.: 290.252



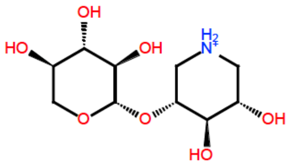
PDB: 1F8C
HET: 4AM
Mol.Wt.: 290.275



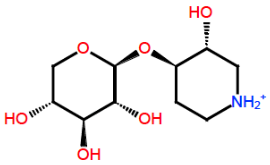
PDB: 1F8D
HET: 9AM
Mol.Wt.: 290.275



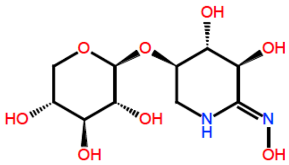
PDB: 1F8E
HET: 49A
Mol.Wt.: 290.299



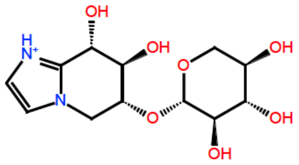
PDB: 1FH7
HET: _XDNJ
Mol.Wt.: 266.273



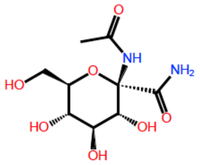
PDB: 1FH8
HET: _XIFG
Mol.Wt.: 250.274



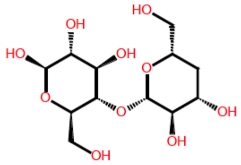
PDB: 1FH9
HET: _XLOX
Mol.Wt.: 294.264



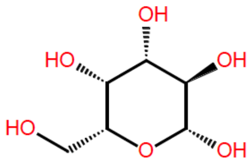
PDB: 1FHD
HET: _XXIM
Mol.Wt.: 303.294



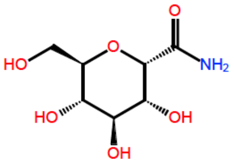
PDB: 1FU8
HET: CR6
Mol.Wt.: 264.237



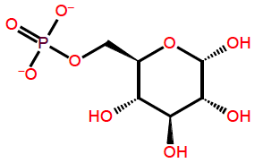
PDB: 1GA8
HET: DEL
Mol.Wt.: 326.303



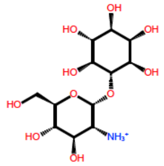
PDB: 1GCA
HET: GAL
Mol.Wt.: 180.159



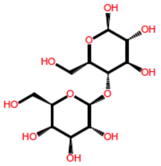
PDB: 1GG8
HET: GLG
Mol.Wt.: 207.185



PDB: 1GPY
HET: G6P
Mol.Wt.: 258.123

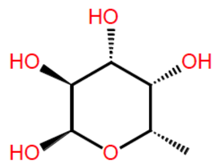


PDB: 1GYM
HET: MYG
Mol.Wt.: 342.326

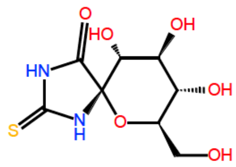


PDB: 1GZC
HET: LAT
Mol.Wt.: 342.303

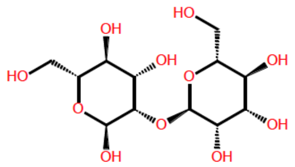
Group 1 (continued)



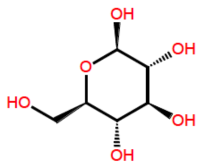
PDB: 1GZT
HET: FUC
Mol.Wt.: 164.16



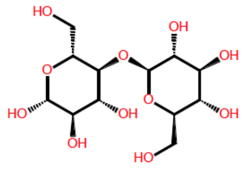
PDB: 1HLF
HET: GL4
Mol.Wt.: 265.267



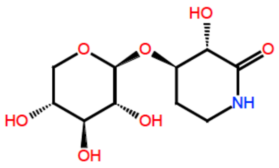
PDB: 1I3H
HET: _2MAN
Mol.Wt.: 342.303



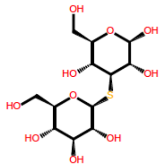
PDB: 1I8A
HET: BGC
Mol.Wt.: 180.159



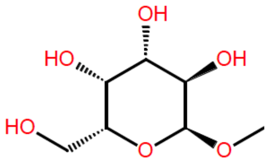
PDB: 1I82
HET: CBI
Mol.Wt.: 342.303



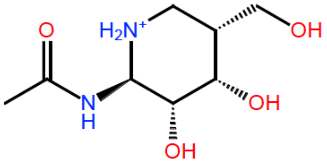
PDB: 1J01
HET: XIL
Mol.Wt.: 263.249



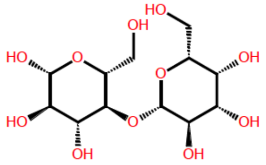
PDB: 1J8V
HET: LAM
Mol.Wt.: 358.367



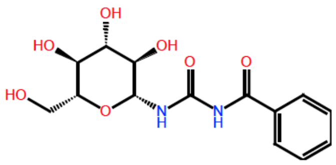
PDB: 1JAC
HET: AMG
Mol.Wt.: 194.186



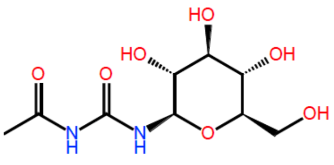
PDB: 1JAK
HET: IFG
Mol.Wt.: 205.236



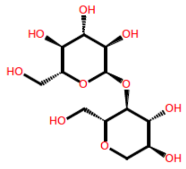
PDB: 1JZN
HET: LAT
Mol.Wt.: 342.303



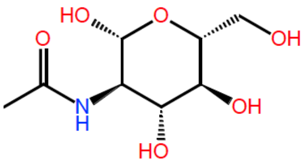
PDB: 1K06
HET: BZD
Mol.Wt.: 326.309



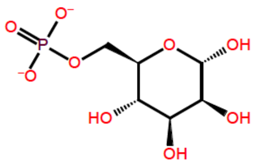
PDB: 1KTI
HET: AZC
Mol.Wt.: 264.237



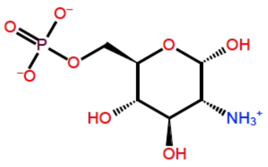
PDB: 1LAX
HET: MAL
Mol.Wt.: 326.303



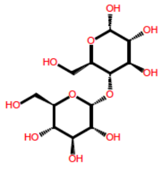
PDB: 1M01
HET: NAG
Mol.Wt.: 221.212



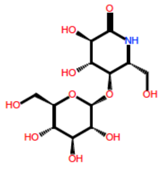
PDB: 1M6P
HET: M6P
Mol.Wt.: 258.123



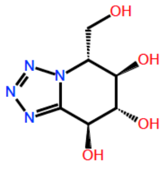
PDB: 1MOQ
HET: GLP
Mol.Wt.: 258.146



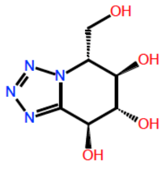
PDB: 1N3W
HET: MAL
Mol.Wt.: 342.303



PDB: 1NAA
HET: ABL
Mol.Wt.: 339.302

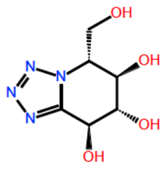


PDB: 1NOI
HET: NTZ
Mol.Wt.: 202.171

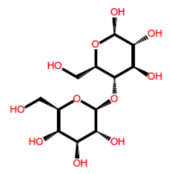


PDB: 1NOJ
HET: NTZ
Mol.Wt.: 202.171

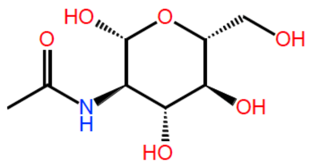
Group 1 (continued)



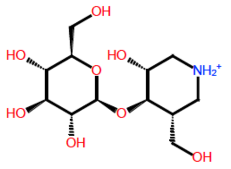
PDB: 1NOK
HET: NTZ
Mol.Wt.: 202.171



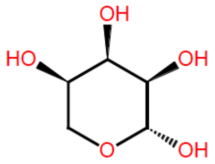
PDB: 1O7O
HET: LAT
Mol.Wt.: 342.303



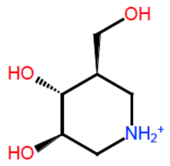
PDB: 1O9W
HET: NAG
Mol.Wt.: 221.212



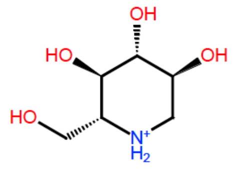
PDB: 1OCQ
HET: _GIFG
Mol.Wt.: 310.327



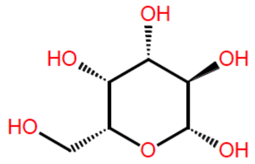
PDB: 1OGD
HET: RIP
Mol.Wt.: 150.132



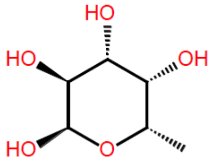
PDB: 1OIF
HET: IFM
Mol.Wt.: 148.183



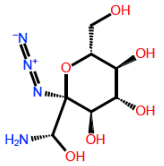
PDB: 1OIM
HET: NOJ
Mol.Wt.: 164.183



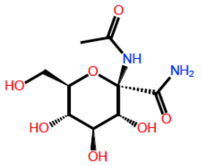
PDB: 1OKO
HET: GAL
Mol.Wt.: 180.159



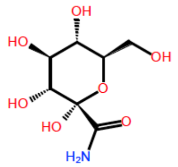
PDB: 1OXC
HET: FUC
Mol.Wt.: 164.16



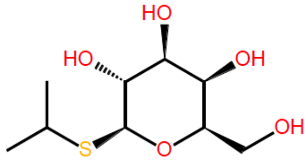
PDB: 1P4G
HET: CGF
Mol.Wt.: 250.213



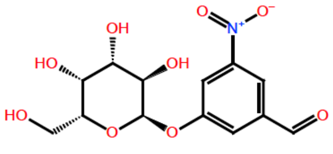
PDB: 1P4H
HET: CR6
Mol.Wt.: 264.237



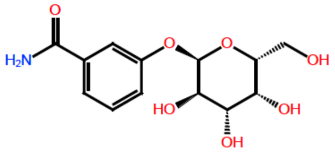
PDB: 1P4J
HET: CBF
Mol.Wt.: 223.184



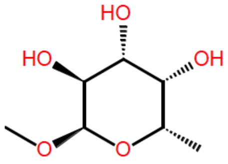
PDB: 1PX4
HET: IPT
Mol.Wt.: 238.305



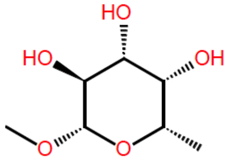
PDB: 1PZI
HET: 1DM
Mol.Wt.: 329.266



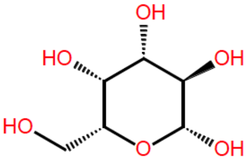
PDB: 1PZK
HET: J12
Mol.Wt.: 299.283



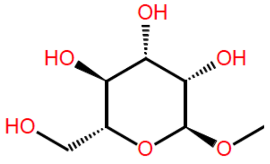
PDB: 1RDI
HET: MFU
Mol.Wt.: 178.187



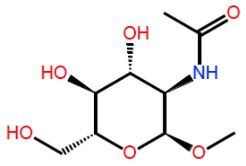
PDB: 1RDJ
HET: MFB
Mol.Wt.: 178.187



PDB: 1RDK
HET: GAL
Mol.Wt.: 180.159

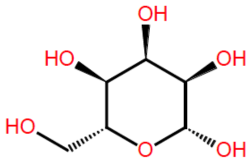


PDB: 1RDL
HET: MMA
Mol.Wt.: 194.186

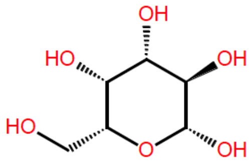


PDB: 1RDN
HET: _MNDG
Mol.Wt.: 235.239

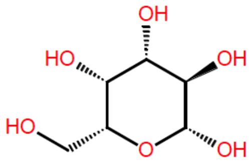
Group 1 (continued)



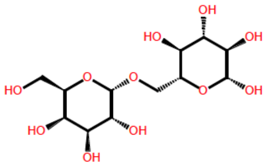
PDB: 1RPJ
HET: ALL
Mol.Wt.: 180.159



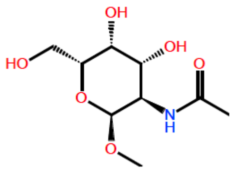
PDB: 1TLG
HET: GAL
Mol.Wt.: 180.159



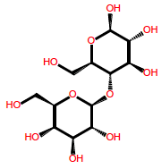
PDB: 1UGW
HET: GAL
Mol.Wt.: 180.159



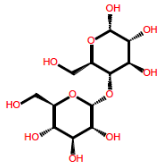
PDB: 1UGY
HET: _GAGC
Mol.Wt.: 342.303



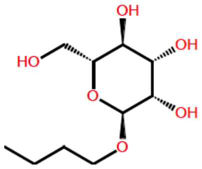
PDB: 1UH0
HET: MGC
Mol.Wt.: 235.239



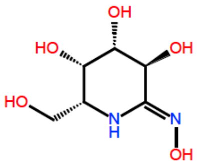
PDB: 1ULC
HET: LAT
Mol.Wt.: 342.303



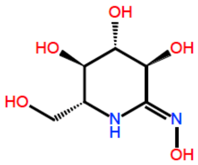
PDB: 1URG
HET: MAL
Mol.Wt.: 342.303



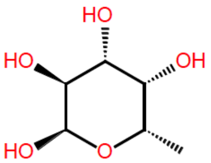
PDB: 1UWF
HET: DEG
Mol.Wt.: 236.267



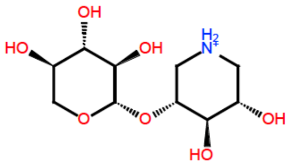
PDB: 1UWT
HET: GTL
Mol.Wt.: 192.173



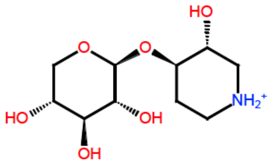
PDB: 1UWU
HET: GOX
Mol.Wt.: 192.173



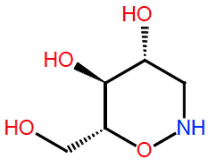
PDB: 1UZV
HET: FUC
Mol.Wt.: 164.16



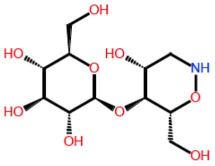
PDB: 1V0K
HET: _XDNJ
Mol.Wt.: 266.273



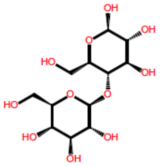
PDB: 1V0L
HET: _XIFG
Mol.Wt.: 250.274



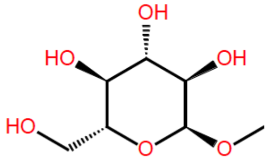
PDB: 1W3J
HET: OXZ
Mol.Wt.: 149.148



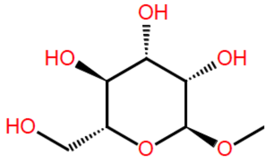
PDB: 1W3K
HET: _CELB
Mol.Wt.: 311.291



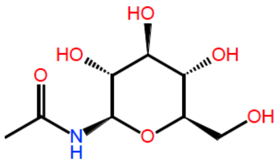
PDB: 1W6O
HET: LAT
Mol.Wt.: 342.303



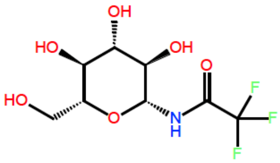
PDB: 1WS4
HET: GYP
Mol.Wt.: 194.186



PDB: 1WS5
HET: MMA
Mol.Wt.: 194.186

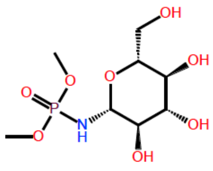


PDB: 1WW2
HET: NBG
Mol.Wt.: 221.212

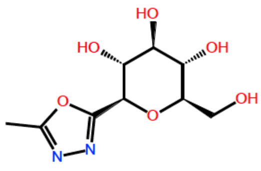


PDB: 1WW3
HET: NTF
Mol.Wt.: 275.183

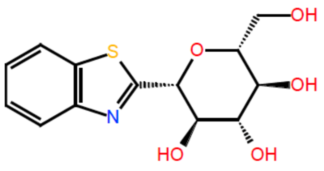
Group 1 (continued)



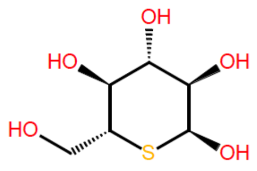
PDB: 1XC7
HET: GL6
Mol.Wt.: 287.208



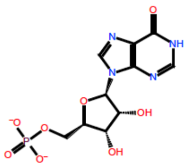
PDB: 1XL0
HET: OX2
Mol.Wt.: 246.222



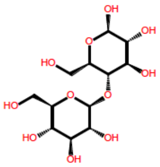
PDB: 1XL1
HET: TH1
Mol.Wt.: 297.332



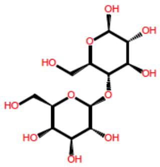
PDB: 1XLI
HET: GLT
Mol.Wt.: 196.224



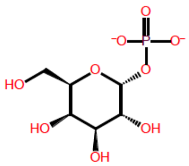
PDB: 1YFZ
HET: IMP
Mol.Wt.: 346.195



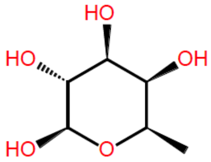
PDB: 1Z3T
HET: CBI
Mol.Wt.: 342.303



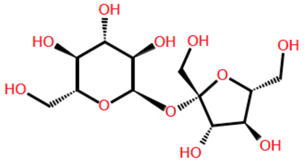
PDB: 1Z3V
HET: LAT
Mol.Wt.: 342.303



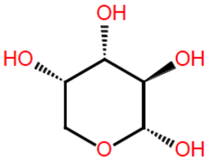
PDB: 1Z4O
HET: GL1
Mol.Wt.: 258.123



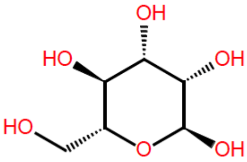
PDB: 2AAC
HET: FCB
Mol.Wt.: 164.16



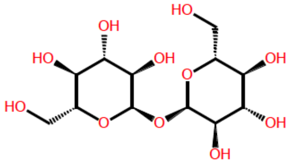
PDB: 2ADD
HET: SUC
Mol.Wt.: 342.303



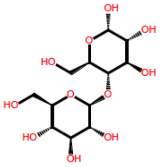
PDB: 2ARC
HET: ARA
Mol.Wt.: 150.132



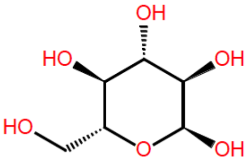
PDB: 2ARE
HET: MAN
Mol.Wt.: 180.159



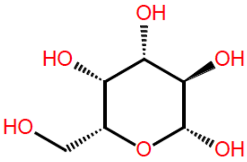
PDB: 2B1Q
HET: TRE
Mol.Wt.: 342.303



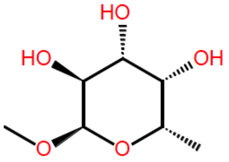
PDB: 2B1R
HET: CBI
Mol.Wt.: 342.303



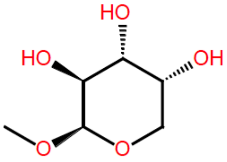
PDB: 2B3B
HET: GLC
Mol.Wt.: 180.159



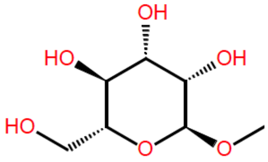
PDB: 2B3F
HET: GAL
Mol.Wt.: 180.159



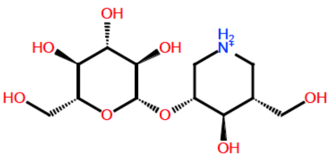
PDB: 2BOI
HET: MFU
Mol.Wt.: 178.187



PDB: 2BOJ
HET: ARW
Mol.Wt.: 164.16

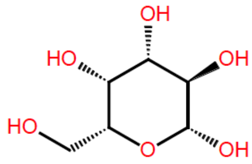


PDB: 2BV4
HET: MMA
Mol.Wt.: 194.186

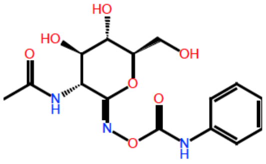


PDB: 2BVD
HET: ISX
Mol.Wt.: 310.327

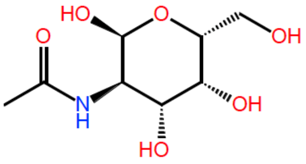
Group 1 (continued)



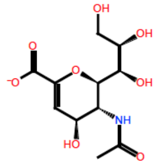
PDB: 2BZD
HET: GAL
Mol.Wt.: 180.159



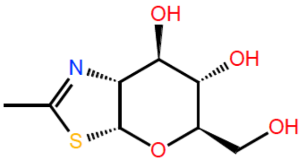
PDB: 2CBJ
HET: OAN
Mol.Wt.: 353.335



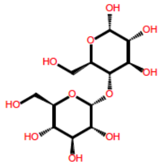
PDB: 2CCV
HET: A2G
Mol.Wt.: 221.212



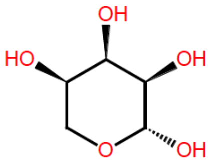
PDB: 2CEX
HET: DAN
Mol.Wt.: 290.252



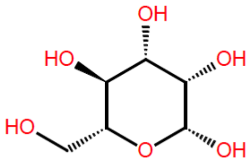
PDB: 2CHN
HET: NGT
Mol.Wt.: 219.261



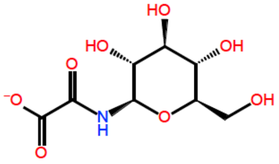
PDB: 2D2V
HET: MAL
Mol.Wt.: 342.303



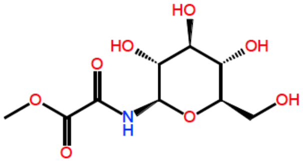
PDB: 2DRI
HET: RIP
Mol.Wt.: 150.132



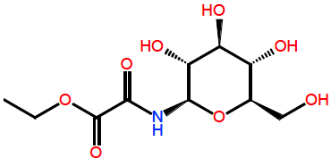
PDB: 2E22
HET: MAN
Mol.Wt.: 180.159



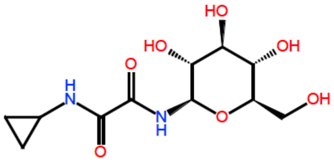
PDB: 2F3P
HET: 4GP
Mol.Wt.: 250.187



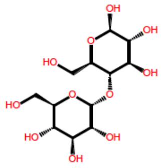
PDB: 2F3Q
HET: 6GP
Mol.Wt.: 265.222



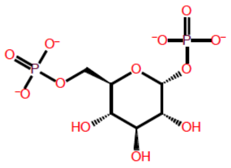
PDB: 2F3S
HET: 7GP
Mol.Wt.: 279.249



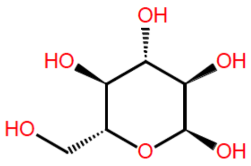
PDB: 2F3U
HET: 8GP
Mol.Wt.: 290.275



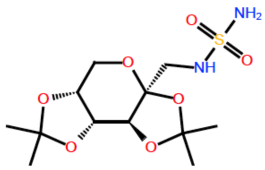
PDB: 2F5T
HET: MAL
Mol.Wt.: 342.303



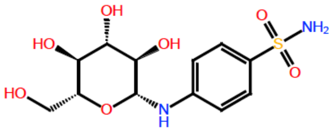
PDB: 2FKF
HET: G16
Mol.Wt.: 336.087



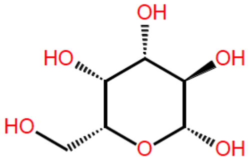
PDB: 2GPB
HET: GLC
Mol.Wt.: 180.159



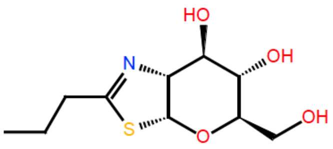
PDB: 2H15
HET: B19
Mol.Wt.: 338.382



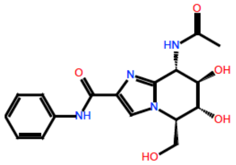
PDB: 2HL4
HET: BO1
Mol.Wt.: 334.35



PDB: 2J1A
HET: GAL
Mol.Wt.: 180.159

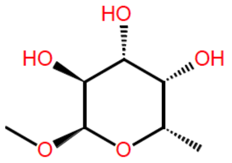


PDB: 2J4G
HET: NB1
Mol.Wt.: 247.315

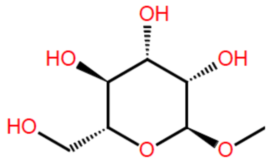


PDB: 2J47
HET: GDV
Mol.Wt.: 360.373

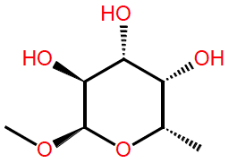
Group 1 (continued)



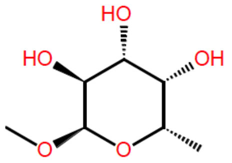
PDB: 2JDM
HET: MFU
Mol.Wt.: 178.187



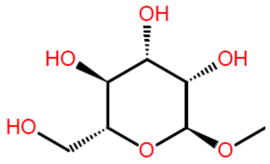
PDB: 2JDN
HET: MMA
Mol.Wt.: 194.186



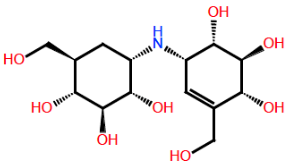
PDB: 2JDP
HET: MFU
Mol.Wt.: 178.187



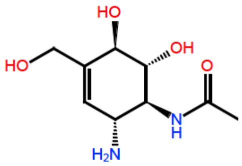
PDB: 2JDU
HET: MFU
Mol.Wt.: 178.187



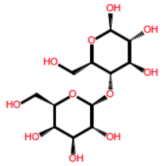
PDB: 2JDY
HET: MMA
Mol.Wt.: 194.186



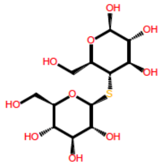
PDB: 2JF4
HET: VDM
Mol.Wt.: 335.357



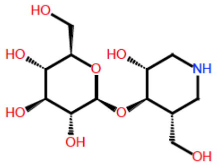
PDB: 2JIW
HET: BEU
Mol.Wt.: 216.239



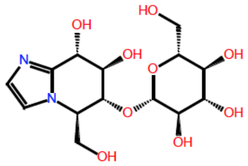
PDB: 2NMO
HET: LAT
Mol.Wt.: 342.303



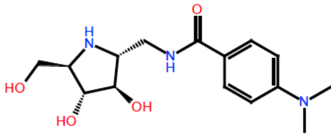
PDB: 2O9R
HET: TCB
Mol.Wt.: 358.367



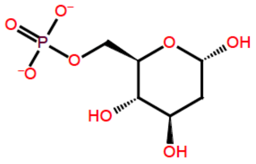
PDB: 2OYK
HET: 9MR
Mol.Wt.: 309.319



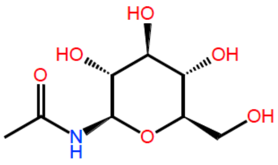
PDB: 2OYL
HET: IDC
Mol.Wt.: 362.339



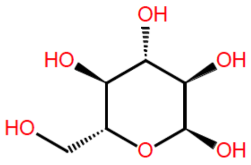
PDB: 2OYM
HET: MNI
Mol.Wt.: 309.368



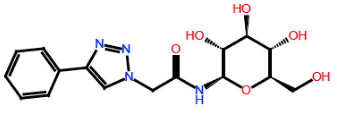
PDB: 2PRI
HET: D6G
Mol.Wt.: 242.124



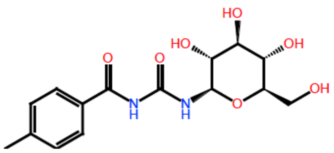
PDB: 2PRJ
HET: NBG
Mol.Wt.: 221.212



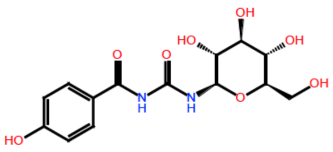
PDB: 2PYD
HET: GLC
Mol.Wt.: 180.159



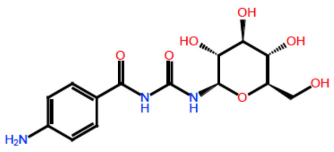
PDB: 2PYI
HET: DL8
Mol.Wt.: 364.361



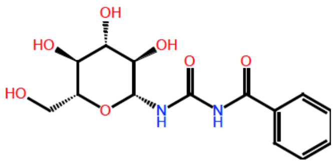
PDB: 2QLM
HET: F68
Mol.Wt.: 340.336



PDB: 2QN7
HET: HBZ
Mol.Wt.: 342.308

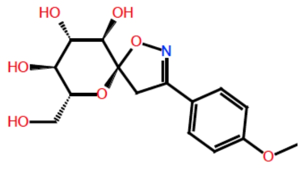


PDB: 2QN9
HET: NBX
Mol.Wt.: 341.323

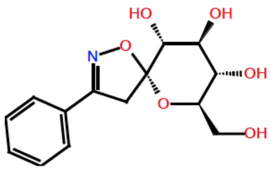


PDB: 2QNB
HET: BZD
Mol.Wt.: 326.309

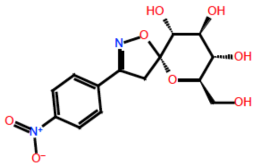
Group 1 (continued)



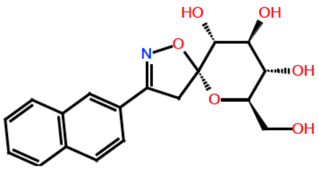
PDB: 2QRG
HET: M07
Mol.Wt.: 325.321



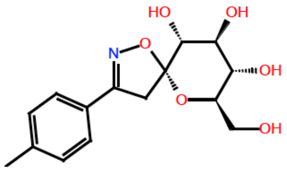
PDB: 2QRH
HET: M08
Mol.Wt.: 295.295



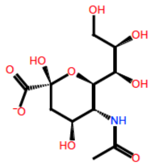
PDB: 2QRM
HET: M09
Mol.Wt.: 340.292



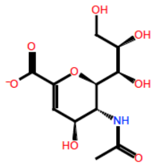
PDB: 2QRP
HET: S06
Mol.Wt.: 345.355



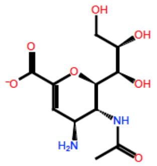
PDB: 2QRQ
HET: S13
Mol.Wt.: 309.322



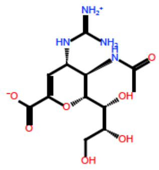
PDB: 2QWB
HET: SIA
Mol.Wt.: 308.267



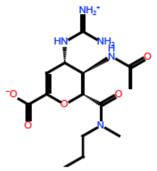
PDB: 2QWC
HET: DAN
Mol.Wt.: 290.252



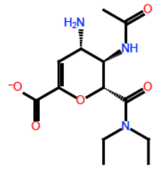
PDB: 2QWD
HET: 4AM
Mol.Wt.: 289.267



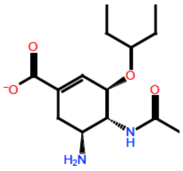
PDB: 2QWE
HET: GNA
Mol.Wt.: 332.316



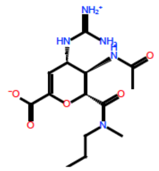
PDB: 2QWF
HET: G20
Mol.Wt.: 341.37



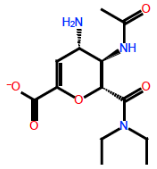
PDB: 2QWG
HET: G28
Mol.Wt.: 298.321



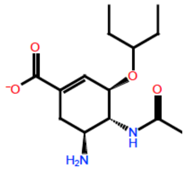
PDB: 2QWH
HET: G39
Mol.Wt.: 283.35



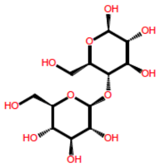
PDB: 2QWI
HET: G20
Mol.Wt.: 341.37



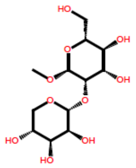
PDB: 2QWJ
HET: G28
Mol.Wt.: 298.321



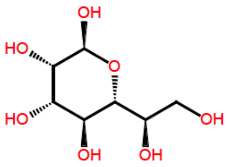
PDB: 2QWK
HET: G39
Mol.Wt.: 283.35



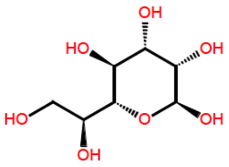
PDB: 2RFY
HET: CBI
Mol.Wt.: 342.303



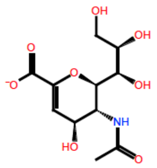
PDB: 2RI9
HET: _LYM
Mol.Wt.: 326.303



PDB: 2RIA
HET: 289
Mol.Wt.: 210.185

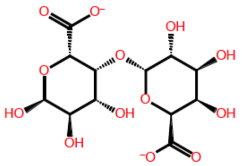


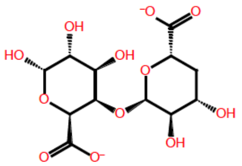
PDB: 2RIB
HET: GMH
Mol.Wt.: 210.185

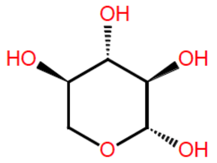


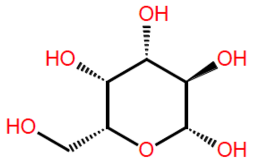
PDB: 2SIM
HET: DAN
Mol.Wt.: 290.252

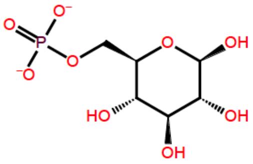
Group 1 (continued)

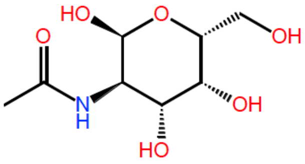

PDB: 2UVH
HET: _DADA
Mol.Wt.: 368.254

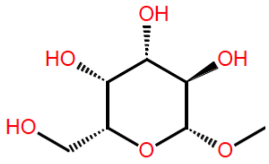

PDB: 2UVI
HET: UNG
Mol.Wt.: 352.254

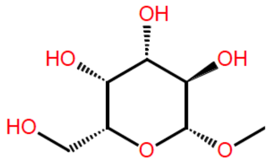

PDB: 2V4V
HET: XYP
Mol.Wt.: 150.132

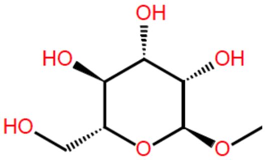

PDB: 2V72
HET: GAL
Mol.Wt.: 180.159

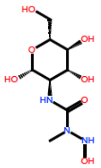

PDB: 2VEZ
HET: G6P
Mol.Wt.: 258.123

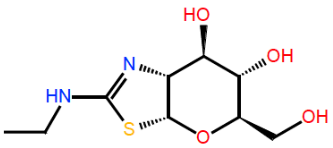

PDB: 2VMC
HET: A2G
Mol.Wt.: 221.212

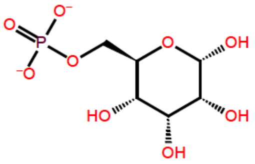

PDB: 2VMD
HET: MBG
Mol.Wt.: 194.186

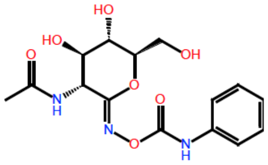

PDB: 2VMG
HET: MBG
Mol.Wt.: 194.186

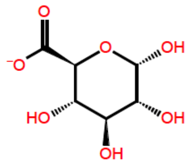

PDB: 2VNV
HET: MMA
Mol.Wt.: 194.186

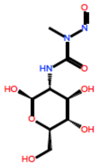

PDB: 2VUR
HET: YX1
Mol.Wt.: 267.241

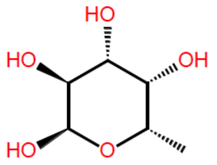

PDB: 2VVN
HET: NHT
Mol.Wt.: 248.303

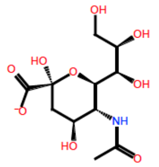

PDB: 2VVO
HET: A6P
Mol.Wt.: 258.123

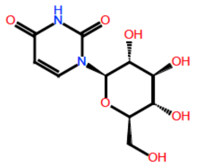

PDB: 2VVS
HET: OAN
Mol.Wt.: 353.335

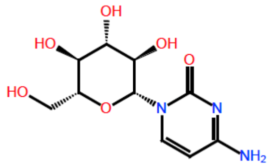

PDB: 2VZR
HET: GCU
Mol.Wt.: 193.134

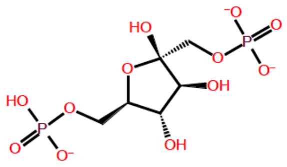

PDB: 2W4X
HET: STZ
Mol.Wt.: 265.225


PDB: 2WCV
HET: FUC
Mol.Wt.: 164.16

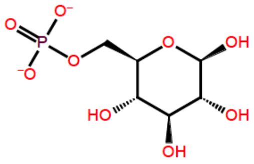

PDB: 3B50
HET: SLB
Mol.Wt.: 308.267


PDB: 3BCS
HET: CJB
Mol.Wt.: 274.232

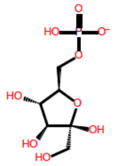

PDB: 3BD8
HET: C3B
Mol.Wt.: 273.248


PDB: 3BXF
HET: FBP
Mol.Wt.: 337.095

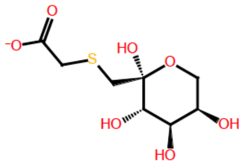
Group 1 (continued)



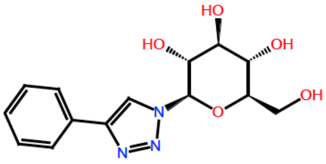
PDB: 3BXG
HET: BG6
Mol.Wt.: 258.123



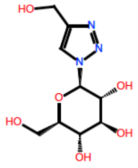
PDB: 3BXH
HET: F6P
Mol.Wt.: 259.131



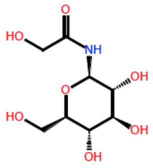
PDB: 3DJE
HET: FSA
Mol.Wt.: 253.253



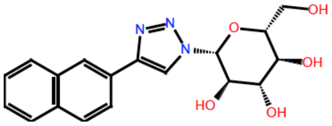
PDB: 3G2H
HET: KOT
Mol.Wt.: 307.309



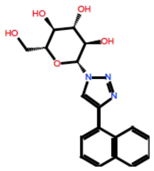
PDB: 3G2I
HET: RUG
Mol.Wt.: 261.236



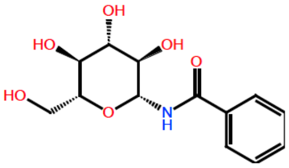
PDB: 3G2J
HET: 9GP
Mol.Wt.: 237.211



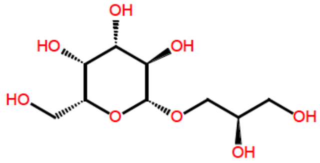
PDB: 3G2K
HET: SKY
Mol.Wt.: 357.369



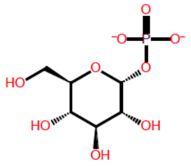
PDB: 3G2L
HET: LEW
Mol.Wt.: 357.369



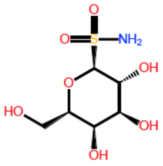
PDB: 3G2N
HET: OAK
Mol.Wt.: 283.284



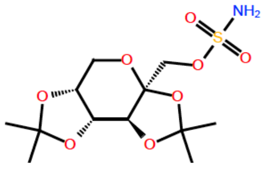
PDB: 3GA5
HET: RGG
Mol.Wt.: 254.239



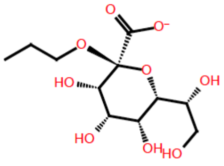
PDB: 3GPB
HET: G1P
Mol.Wt.: 258.123



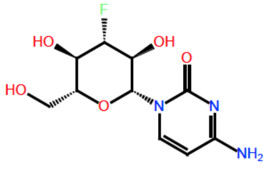
PDB: 3HKQ
HET: 1SD
Mol.Wt.: 243.237



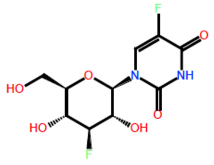
PDB: 3HKU
HET: TOR
Mol.Wt.: 339.367



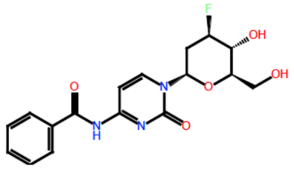
PDB: 3IJH
HET: KO2
Mol.Wt.: 295.269



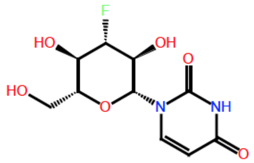
PDB: 3L7B
HET: DKZ
Mol.Wt.: 275.239



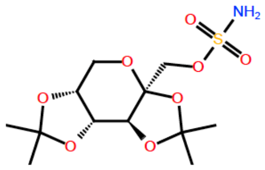
PDB: 3L7C
HET: DK4
Mol.Wt.: 294.214



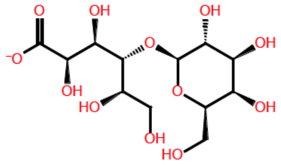
PDB: 3L7D
HET: DK5
Mol.Wt.: 363.349



PDB: 3L79
HET: DKX
Mol.Wt.: 276.223

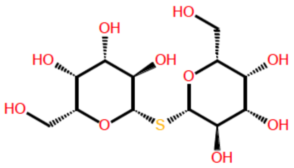


PDB: 3LXE
HET: TOR
Mol.Wt.: 339.367

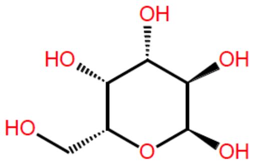


PDB: 3OY8
HET: _LBA
Mol.Wt.: 357.294

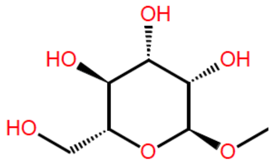
Group 1 (continued)



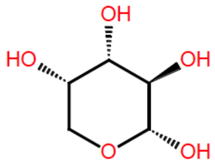
PDB: 3OYW
HET: TDG
Mol.Wt.: 358.367



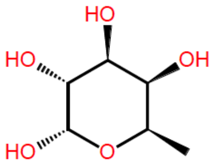
PDB: 5ABP
HET: GLA
Mol.Wt.: 180.159



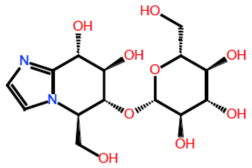
PDB: 5CNA
HET: MMA
Mol.Wt.: 194.186



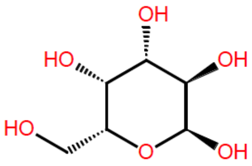
PDB: 6ABP
HET: ARA
Mol.Wt.: 150.132



PDB: 7ABP
HET: FCA
Mol.Wt.: 164.16

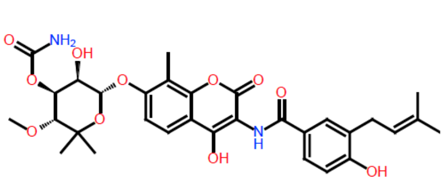


PDB: 8A3H
HET: IDC
Mol.Wt.: 362.339

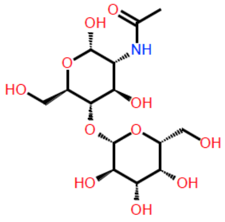


PDB: 8ABP
HET: GLA
Mol.Wt.: 180.159

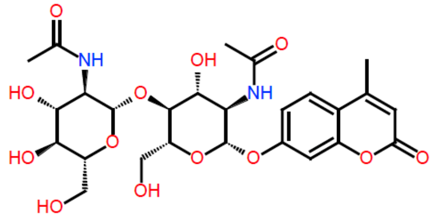
Group 2



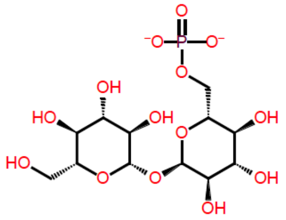
PDB: 1AJ6
HET: NOV
Mol.Wt.: 612.639



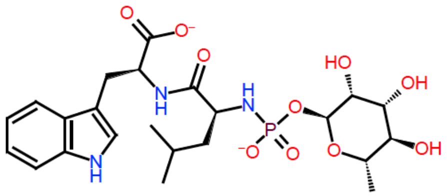
PDB: 1AX2
HET: _NAL
Mol.Wt.: 383.355



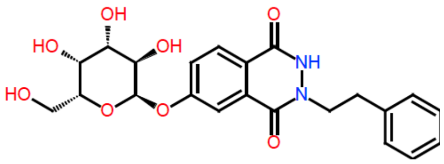
PDB: 1BB7
HET: GUM
Mol.Wt.: 582.566



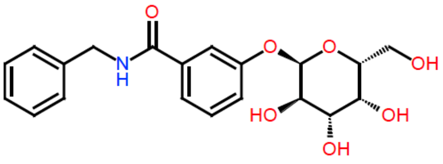
PDB: 1BYK
HET: T6P
Mol.Wt.: 420.267



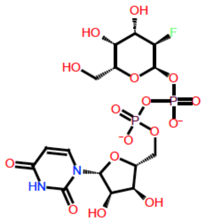
PDB: 1DMT
HET: RDF
Mol.Wt.: 541.499



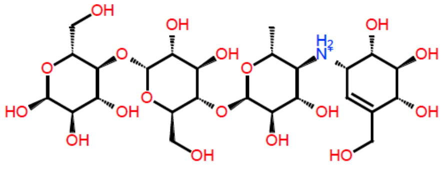
PDB: 1EEF
HET: _PEPG
Mol.Wt.: 444.445



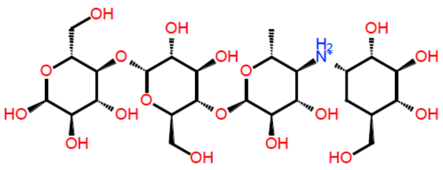
PDB: 1FD7
HET: AI1
Mol.Wt.: 389.409



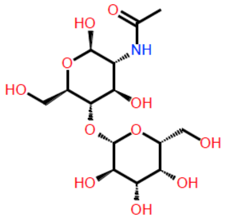
PDB: 1GA8
HET: UPF
Mol.Wt.: 566.284



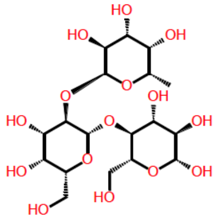
PDB: 1GAH
HET: ACR
Mol.Wt.: 646.625



PDB: 1GAI
HET: GAC
Mol.Wt.: 648.641

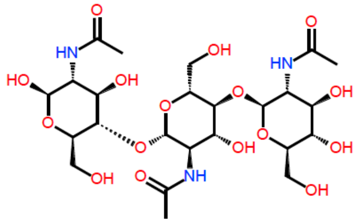


PDB: 1GX4
HET: _NAL
Mol.Wt.: 383.355

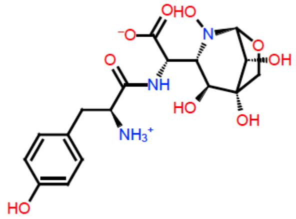


PDB: 1GZ9
HET: _FLAC
Mol.Wt.: 488.447

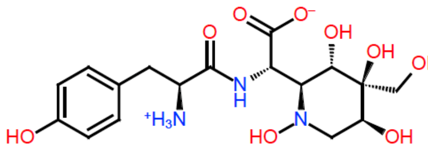
Group 2 (continued)



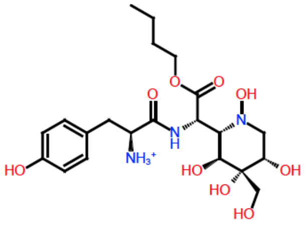
PDB: 1HEW
HET: _NAG3
Mol.Wt.: 627.605



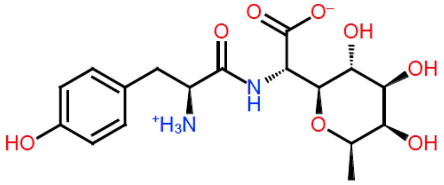
PDB: 1JII
HET: 383
Mol.Wt.: 413.388



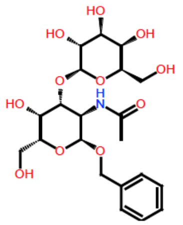
PDB: 1JIJ
HET: 629
Mol.Wt.: 415.404



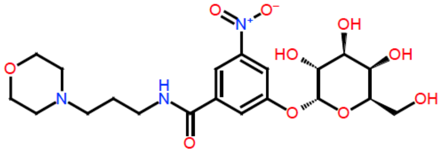
PDB: 1JIK
HET: 545
Mol.Wt.: 472.52



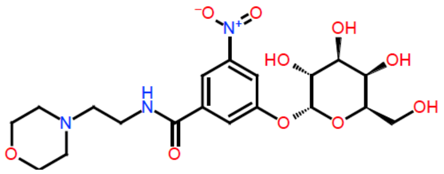
PDB: 1JIL
HET: 485
Mol.Wt.: 384.389



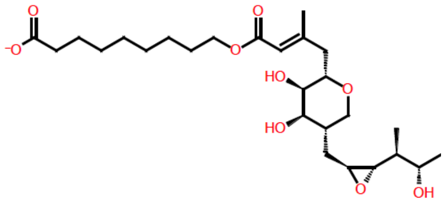
PDB: 1JLX
HET: _TDSC
Mol.Wt.: 473.481



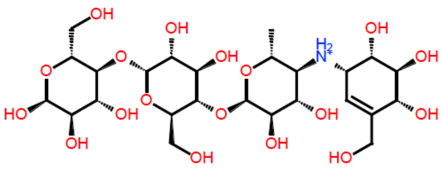
PDB: 1JQY
HET: A32
Mol.Wt.: 471.468



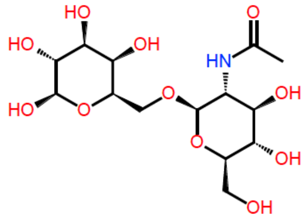
PDB: 1JR0
HET: A24
Mol.Wt.: 457.441



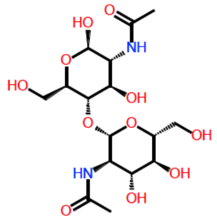
PDB: 1JZS
HET: MRC
Mol.Wt.: 499.627



PDB: 1K1Y
HET: ACR
Mol.Wt.: 646.625

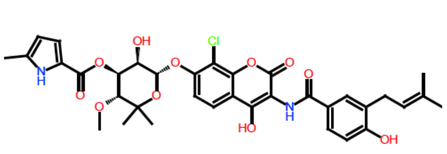


PDB: 1K7T
HET: _NGGA
Mol.Wt.: 383.355

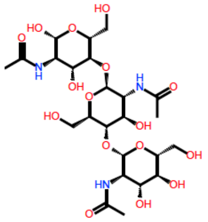


PDB: 1K7U
HET: _2NAG
Mol.Wt.: 424.408

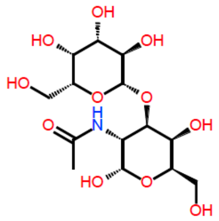
Group 2 (continued)



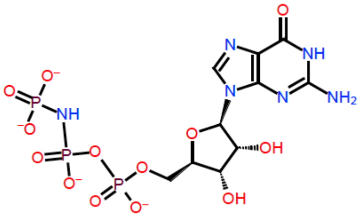
PDB: 1KZN
HET: CBN
Mol.Wt.: 697.145



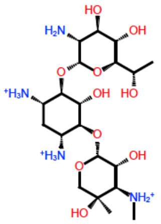
PDB: 1LZB
HET: _NAGT
Mol.Wt.: 627.605



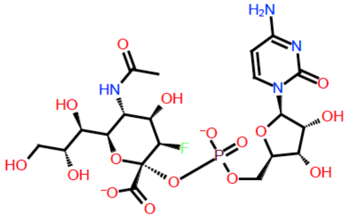
PDB: 1M26
HET: _TANT
Mol.Wt.: 383.355



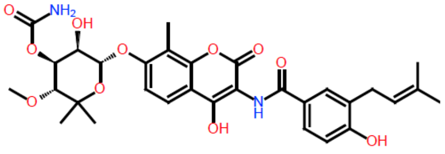
PDB: 1NF3
HET: GNP
Mol.Wt.: 518.169



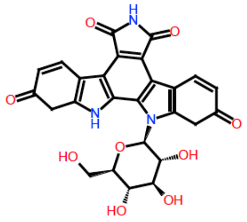
PDB: 1NJJ
HET: GET
Mol.Wt.: 499.587



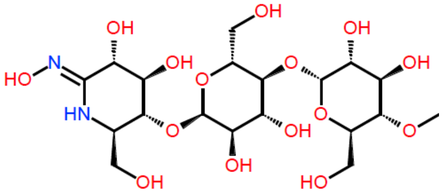
PDB: 1RO7
HET: CSF
Mol.Wt.: 630.436



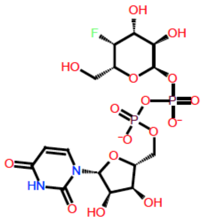
PDB: 1S14
HET: NOV
Mol.Wt.: 612.639



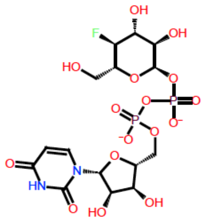
PDB: 1SEU
HET: SA3
Mol.Wt.: 519.472



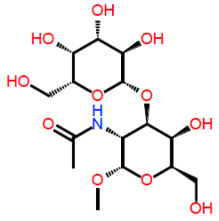
PDB: 1U33
HET: LM2
Mol.Wt.: 530.487



PDB: 1UDA
HET: UFG
Mol.Wt.: 566.284

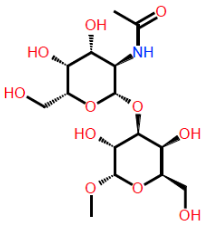


PDB: 1UDB
HET: _UFGC
Mol.Wt.: 566.284

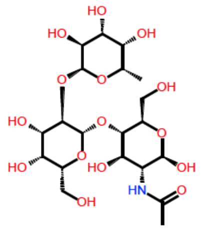


PDB: 1UGX
HET: _MTNT
Mol.Wt.: 397.383

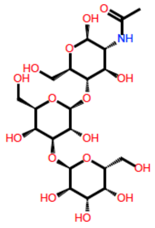
Group 2 (continued)



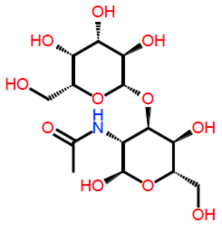
PDB: 1UH1
HET: _NGMG
Mol.Wt.: 397.383



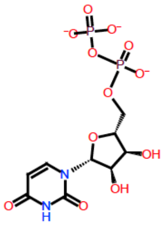
PDB: 1ULD
HET: _BLDH
Mol.Wt.: 529.5



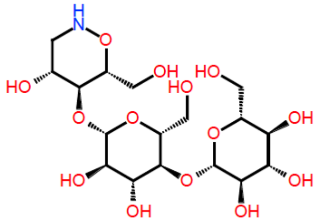
PDB: 1ULE
HET: _LNB2
Mol.Wt.: 545.499



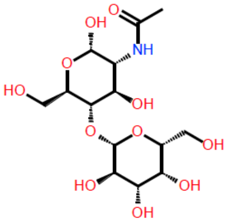
PDB: 1ULG
HET: _TFAN
Mol.Wt.: 383.355



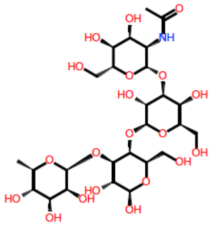
PDB: 1VZT
HET: UDP
Mol.Wt.: 401.142



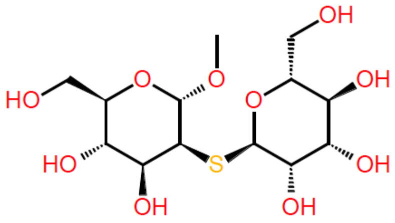
PDB: 1W3L
HET: _CELT
Mol.Wt.: 473.435



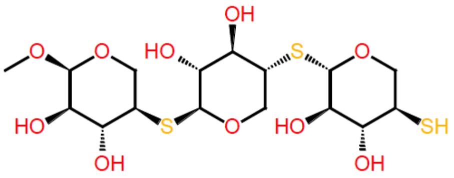
PDB: 1W6P
HET: _GAND
Mol.Wt.: 383.355



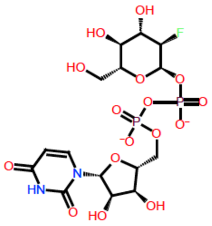
PDB: 1W8F
HET: _LNPV
Mol.Wt.: 691.643



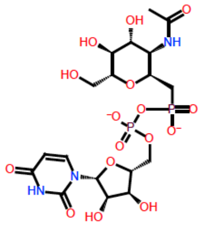
PDB: 1X9D
HET: SMD
Mol.Wt.: 372.394



PDB: 1XNK
HET: XS2
Mol.Wt.: 476.588

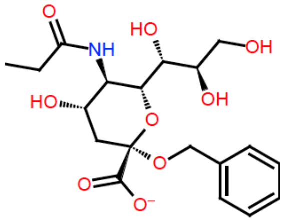


PDB: 2AM4
HET: U2F
Mol.Wt.: 566.284

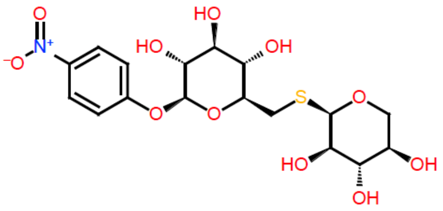


PDB: 2APC
HET: UDM
Mol.Wt.: 603.374

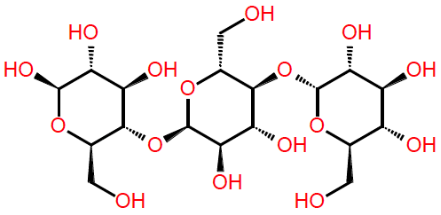
Group 2 (continued)



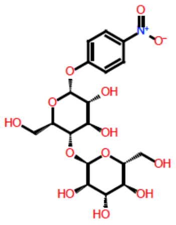
PDB: 2BVE
HET: PH5
Mol.Wt.: 412.42



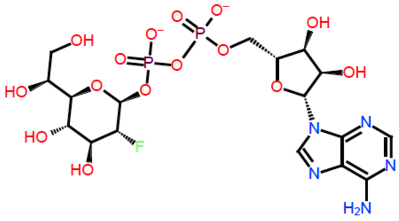
PDB: 2F2H
HET: XTG
Mol.Wt.: 449.437



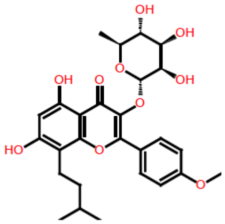
PDB: 2GGU
HET: MLR
Mol.Wt.: 504.446



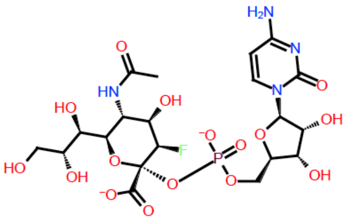
PDB: 2GGX
HET: NPJ
Mol.Wt.: 463.399



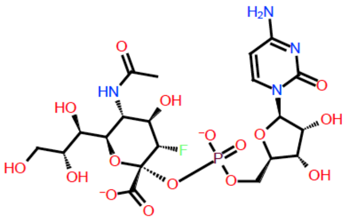
PDB: 2H1H
HET: AFH
Mol.Wt.: 619.351



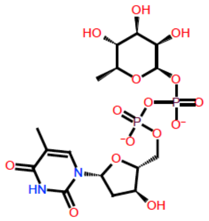
PDB: 2H44
HET: 7CA
Mol.Wt.: 516.55



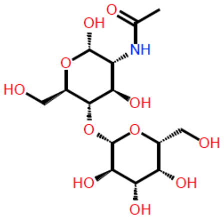
PDB: 2IHJ
HET: CSF
Mol.Wt.: 630.436



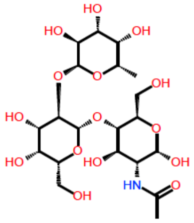
PDB: 2IHK
HET: _CSFE
Mol.Wt.: 630.436



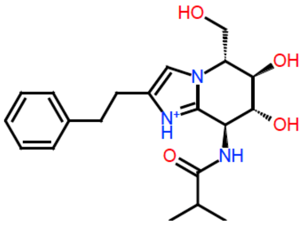
PDB: 2IXH
HET: TRH
Mol.Wt.: 546.322



PDB: 2J1E
HET: _GAND
Mol.Wt.: 383.355

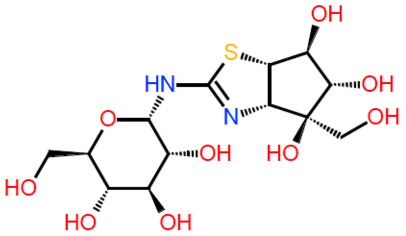


PDB: 2J7M
HET: _BLDH
Mol.Wt.: 529.5

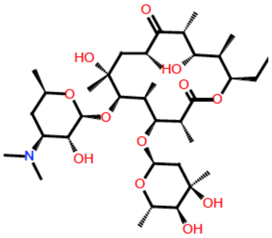


PDB: 2J62
HET: GSZ
Mol.Wt.: 374.464

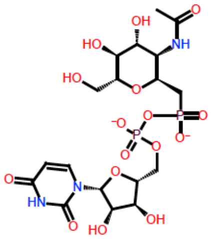
Group 2 (continued)



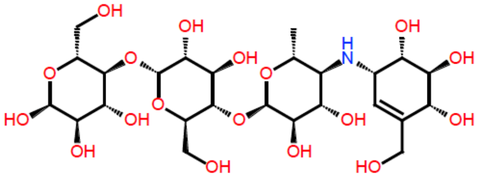
PDB: 2JG0
HET: TTZ
Mol.Wt.: 382.392



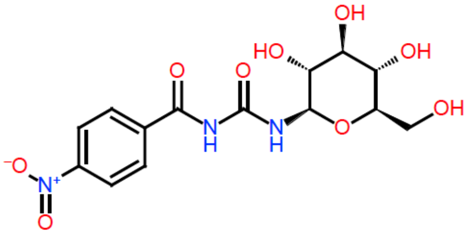
PDB: 2JJO
HET: EY5
Mol.Wt.: 703.919



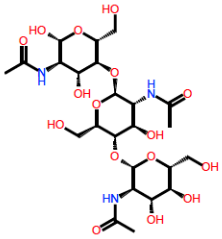
PDB: 2JLB
HET: UDM
Mol.Wt.: 603.374



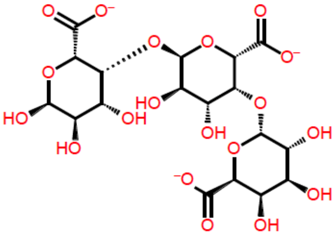
PDB: 2QMJ
HET: ACR
Mol.Wt.: 645.617



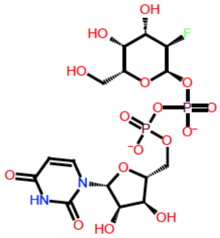
PDB: 2QN8
HET: NBY
Mol.Wt.: 371.306



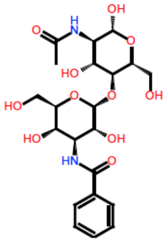
PDB: 2R0H
HET: CTO
Mol.Wt.: 627.605



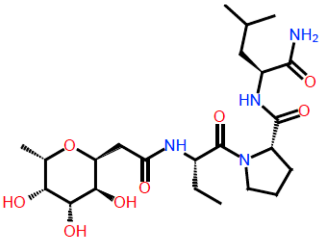
PDB: 2UVJ
HET: _TADA
Mol.Wt.: 543.373



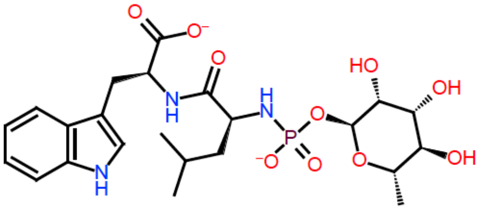
PDB: 2VFZ
HET: UPF
Mol.Wt.: 566.284



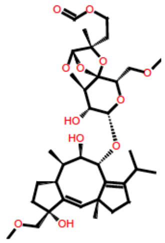
PDB: 2XG3
HET: _BNAL
Mol.Wt.: 486.48



PDB: 3DCQ
HET: 2G0
Mol.Wt.: 500.597

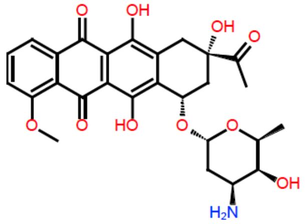


PDB: 3DWB
HET: RDF
Mol.Wt.: 541.499

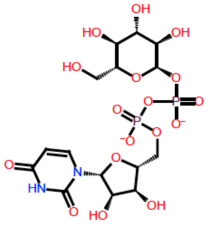


PDB: 3E6Y
HET: CW1
Mol.Wt.: 652.786

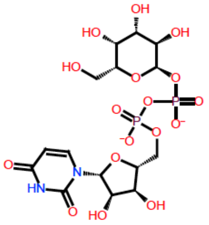
Group 2 (continued)



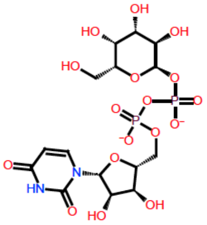
PDB: 3F8F
HET: DM1
Mol.Wt.: 527.533



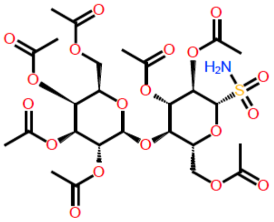
PDB: 3GF4
HET: UPG
Mol.Wt.: 564.293



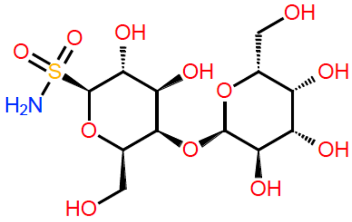
PDB: 3HDQ
HET: GDU
Mol.Wt.: 564.293



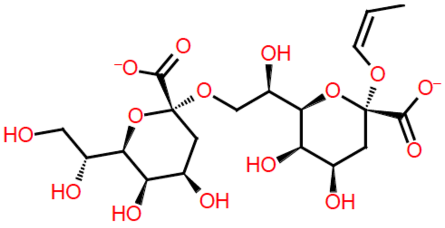
PDB: 3HDY
HET: GDU
Mol.Wt.: 564.293



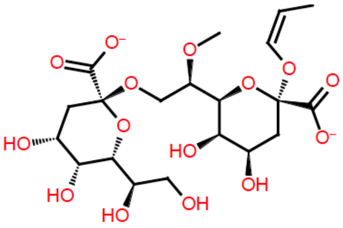
PDB: 3HKN
HET: MFS
Mol.Wt.: 699.644



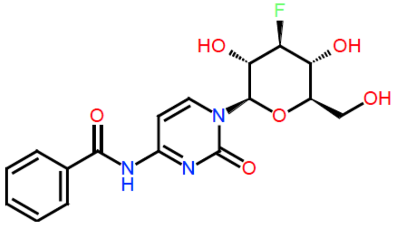
PDB: 3HKT
HET: 2SD
Mol.Wt.: 405.381



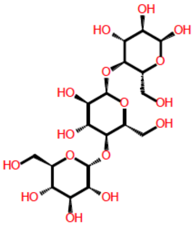
PDB: 3IJY
HET: _KDAO
Mol.Wt.: 496.426



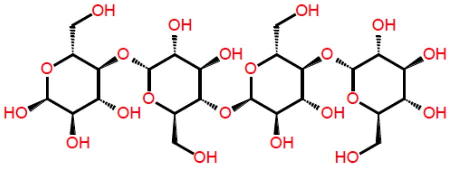
PDB: 3IKC
HET: _KDKM
Mol.Wt.: 510.453



PDB: 3L7A
HET: DKY
Mol.Wt.: 379.348

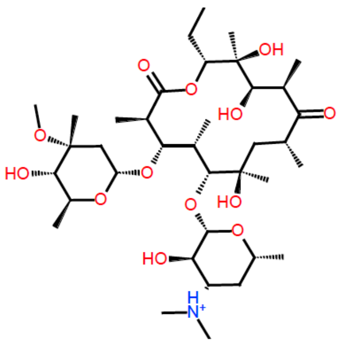


PDB: 3MBP
HET: MLR
Mol.Wt.: 504.446

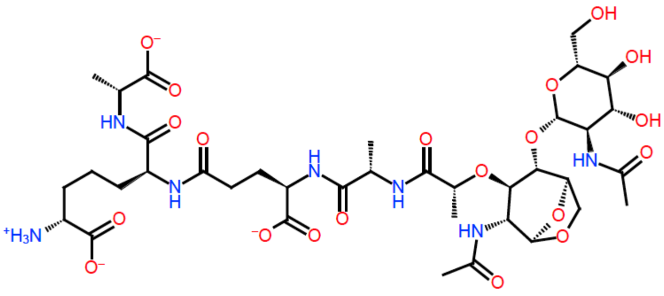


PDB: 4MBP
HET: MTT
Mol.Wt.: 666.59

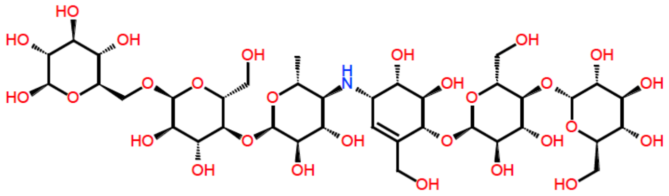
Group 3



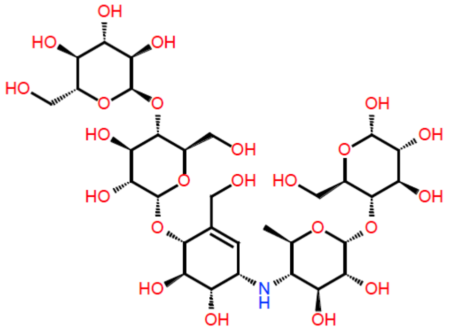
PDB: 2J0D
HET: ERY
Mol.Wt.: 734.953



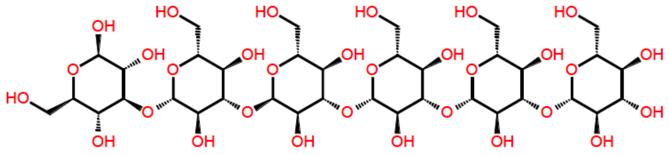
PDB: 2CB3
HET: MLD
Mol.Wt.: 919.902



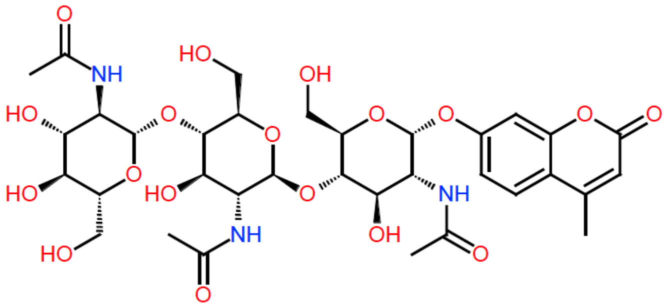
PDB: 1XD1
HET: 6SA
Mol.Wt.: 969.905



PDB: 1XD0
HET: ARE
Mol.Wt.: 807.761



PDB: 1W9W
HET: _LMHX
Mol.Wt.: 990.877



PDB: 1BB6
HET: UMG
Mol.Wt.: 785.763

Appendix 3: *Int. J. Mol. Sci.* 2013, *14*, 684–700

Int. J. Mol. Sci. **2013**, *14*, 684-700; doi:10.3390/ijms14010684

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Article

A Molecular-Modeling Toolbox Aimed at Bridging the Gap between Medicinal Chemistry and Computational Sciences

Sameh Eid [†], Adam Zalewski [†], Martin Smieško, Beat Ernst and Angelo Vedani *

Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50, 4056 Basel, Switzerland; E-Mails: sameh.eid@unibas.ch (S.E.); adam.zalewski@unibas.ch (A.Z.); martin.smiesko@unibas.ch (M.S.); beat.ernst@unibas.ch (B.E.)

[†] These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: angelo.vedani@unibas.ch; Tel.: +41-61-267-1659; Fax: +41-61-267-1552.

Received: 12 December 2012; in revised form: 21 December 2012 / Accepted: 26 December 2012 / Published: 4 January 2013

Abstract: In the current era of high-throughput drug discovery and development, molecular modeling has become an indispensable tool for identifying, optimizing and prioritizing small-molecule drug candidates. The required background in computational chemistry and the knowledge of how to handle the complex underlying protocols, however, might keep medicinal chemists from routinely using *in silico* technologies. Our objective is to encourage those researchers to exploit existing modeling technologies more frequently through easy-to-use graphical user interfaces. In this account, we present two innovative tools (which we are prepared to share with academic institutions) facilitating computational tasks commonly utilized in drug discovery and development: (1) the *VirtualDesignLab* estimates the binding affinity of small molecules by simulating and quantifying their binding to the three-dimensional structure of a target protein; and (2) the *MD Client* launches molecular dynamics simulations aimed at exploring the time-dependent stability of ligand–protein complexes and provides residue-based interaction energies. This allows medicinal chemists to identify sites of potential improvement in their candidate molecule. As a case study, we present the application of our tools towards the design of novel antagonists for the FimH adhesin.

Keywords: computer-aided drug discovery; structure-based design; multi-dimensional QSAR; molecular dynamics; single-click molecular modeling

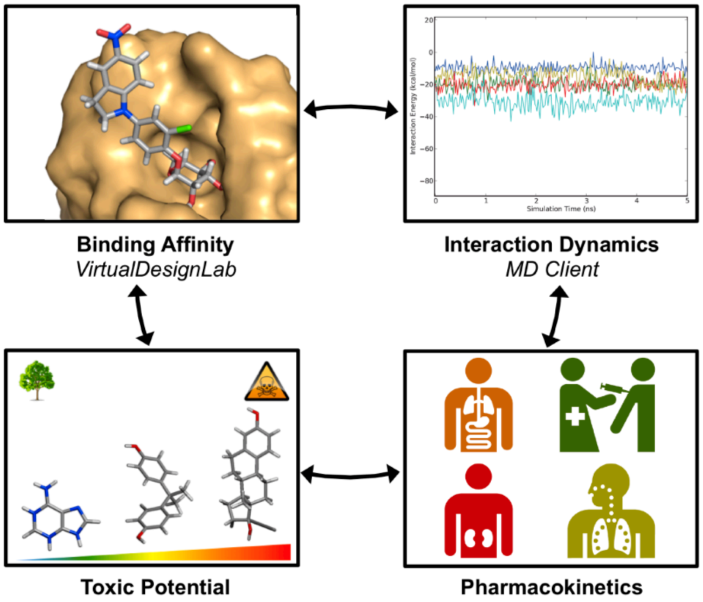
1. Introduction

Molecular modeling has become an integral part of drug discovery and development, with numerous documented examples of successful employment of computational approaches to answer key questions in the field of molecular design. For instance, structure-based design techniques, including small-molecule docking and scoring, can provide structural and energetic information on ligand–protein binding and, hence, guide the design of more potent candidate molecules [1,2]. Additionally, quantitative structure-activity relationships (QSAR) models can provide reliable estimates of binding affinities, particularly of hypothetical ligands—prior to their laborious and costly synthesis and biological testing [3,4]. Molecular dynamics (MD) simulations address more challenging questions regarding the dynamic nature of ligand–receptor interactions [5–11]. Overall, virtual screening can increase the efficiency and reduce time and cost of lead identification [12,13]. A number of commercially available software packages handle one or more of these tasks, e.g., the Schrödinger Suite [14], the Accelrys Discovery Studio [15], the SYBYL-X Suite [16] or the Molecular Operating Environment [17]. Furthermore, a wealth of modeling tools are available free-of-charge, including AutoDock for automated docking [18], Quasar^X for multi-dimensional QSAR [19], Desmond for molecular dynamics simulations [20] and DOCK Blaster for virtual screening [21].

Medicinal chemists involved in design of new ligands for some macromolecular target are nowadays knowledgeable of the binding site's topology at the molecular level. This degree of familiarity with the target provides valuable guidance for modeling techniques, such as docking proposed ligands to that target or developing a QSAR for predicting binding affinities. Such optimized modeling methodologies could, in turn, guide the medicinal chemists' decision making. However, making the best use of these and other modeling techniques requires a tedious and repetitive process of setting and calibrating parameters, as well as collecting and organizing the results. Designing an intuitive interface that encapsulates and hides the complexity of the underlying technologies from the end-user would, thus, motivate medicinal chemists to use modeling tools more frequently.

In this article, we present two novel platforms addressing commonly required tasks in modern drug design workflow: the *VirtualDesignLab* for predicting binding mode and affinity and the *MD Client* for investigating interaction dynamics of ligand–protein complexes (Figure 1). We discuss the development of the underlying models and technologies used in both tools and demonstrate their recent employment in our lab for the design and optimization of novel antagonists for FimH [22–24], a bacterial lectin playing a crucial role in the initial stages of urinary tract infections. Since the goal of the present work is to develop versatile tools that can be easily tuned for any structure-based drug design project, we will conclude with reviewing the steps required to apply/extend our tools for use with other protein targets.

Figure 1. Tools presented in this article handle two common tasks in modern computer-aided drug design workflow. The *VirtualDesignLab* predicts binding mode and estimates the associated binding affinity of prospective ligands. The *MD Client* facilitates simulation and analysis of the dynamics in ligand–protein complexes. In concert with other software predicting pharmacokinetic (e.g., QikProp [25]) and toxicological profiles (e.g., the VirtualToxLab [26]), our tools equip medicinal chemists with a multi-purpose molecular-modeling kit.



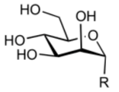
2. Methods

2.1. VirtualDesignLab

The *VirtualDesignLab* is an *in silico* tool developed at our institute (based on the *VirtualToxLab* framework [26] shared by the Biographics Laboratory 3R) simulating and quantifying the binding of small molecules to a macromolecular target. The technology employs automated, flexible docking combined with multi-dimensional quantitative structure-activity relationships (mQSAR). Controlled by an easy-to-use interface, the *VirtualDesignLab* allows medicinal chemists to perform quick and straightforward design, screening and structural inspection of any compound of interest [27].

In order to provide a reliable *in silico* affinity estimate for a given system, it is necessary to account for protein-ligand interactions, solvation and entropic phenomena. In our example system, FimH adhesin, we utilized a set of 108 compounds, along with their experimental affinity data, to develop and validate a corresponding mQSAR model (Table 1). When generating the model, the initial compound structures were constructed using the integrated model-building tool and then optimized with MacroModel [28]. Atomic partial charges were computed using the AMSOL package [29]. All structures were subjected to the conformational-searching algorithm ConfGen [30], resulting in sets of low-energy conformations for each molecule in aqueous solution. Energetically feasible binding conformations (within 10 kcal/mol from the lowest-energy structure) were identified by means of automated docking to two three-dimensional structures (“in” and “out” state, cf. below) of the FimH carbohydrate-binding domain. The employed alignment (Alignator) [31] and docking (Cheetah) [32] protocols allowed for flexibility of both ligand and the protein (induced fit), as well as dynamic solvation. Several templates (based on experimental structures) were used for the pre-alignment in order to account for distinct modes of binding to FimH (referred to as “in” and “out”) reported previously [23,33]. The underlying protein structures were retrieved from the Protein Data Bank (PDB codes 1UWF and 3MCY available at 1.69 Å and 2.90 Å resolution, respectively) and pre-processed (calculation of hydrogen-atom positions, hydrogen-bond network optimization, energy minimization) with the Protein Preparation Wizard in Maestro [34]. A total of 282 docking poses (allowing for multiple poses per ligand) comprising a 4D data set were then used as input (84 training and 24 test substances) for the mQSAR software Quasar [35] to generate a series of quasi-atomistic binding-site models. The underlying model families (comprising 200 members) were evaluated in consensus-scoring mode—along with a direct force-field scoring in Cheetah [32] and the comparison of a molecule’s interaction energy in a box of pre-equilibrated water and in the binding site. For validation, we additionally employed an alternative receptor-modeling concept, Raptor [28], featuring a substantially different scoring function.

Table 1. Structures and binding affinities (pIC50: negative logarithm of IC_{50} [M]) for 52 compounds employed to develop the QSAR model. The remaining data cannot be disclosed at this time, due to pending patent applications.



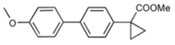
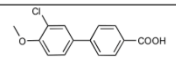
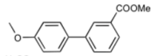
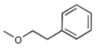
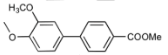
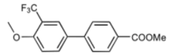
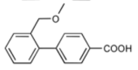
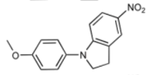
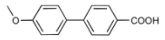
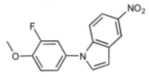
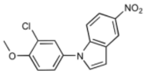
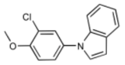
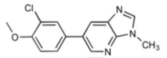
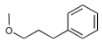
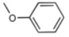
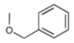
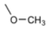
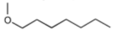
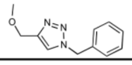
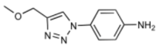
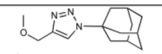
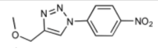
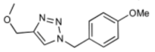
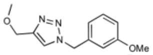
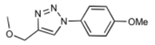
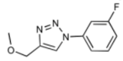
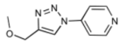
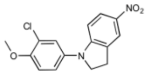
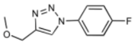
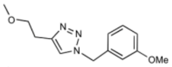
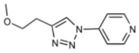
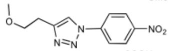
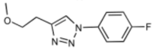
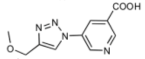
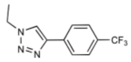
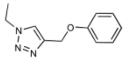
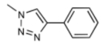
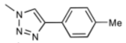
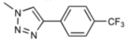
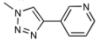
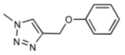
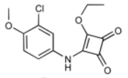
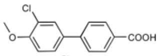
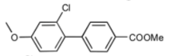
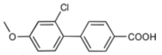
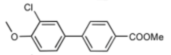
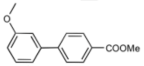
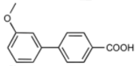
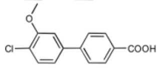
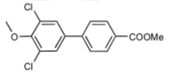
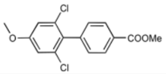
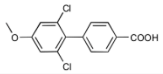
	R	Exp. affinity	Pred. affinity	Residual		R	Exp. affinity	Pred. affinity	Residual
1		7.3	7.6	0.3	**2**		7.0	7.4	0.4
3		7.5	7.5	0	**4**		6.7	6.5	−0.2
5		8.5	7.7	−0.8	**6**		8.6	7.8	−0.8
7		6.2	6.7	0.5	**8**		7.8	7.0	−0.8
9		8.0	7.4	−0.6	**10**		7.5	7.2	−0.3
11		7.8	7.3	−0.5	**12**		8.1	7.0	−1.1
13		7.8	7.5	−0.3	**14**		6.6	7.0	0.4
15		6.8	6.6	−0.2	**16**		6.4	6.2	−0.2
17		5.5	6.0	0.5	**18**		7.2	6.9	−0.3
19		6.2	7.1	0.9	**20**		6.4	6.7	0.3

Table 1. *Cont.*

	R	Exp. affinity	Pred. affinity	Residual		R	Exp. affinity	Pred. affinity	Residual
21		6.9	7.1	0.2	**22**		6.4	7.1	0.7
23		6.3	7.1	0.8	**24**		6.3	7.0	0.7
25		6.8	7.4	0.6	**26**		6.5	6.7	0.2
27		7.2	6.8	−0.4	**28**		8.6	8.0	−0.6
29		6.1	6.8	0.7	**30**		6.6	7.3	0.7
31		6.8	6.8	0	**32**		7.0	7.2	0.2
33		6.7	6.9	0.2	**34**		6.8	6.8	0
35		6.7	6.7	0	**36**		6.5	6.3	−0.2
37		6.6	6.8	0.2	**38**		6.6	6.6	0
39		6.8	7.3	0.5	**40**		6.7	6.5	−0.2
41		6.3	6.5	0.2	**42**		8.0	7.7	−0.3
43		8.2	7.5	−0.7	**44**		7.7	7.6	−0.1
45		7.6	7.5	−0.1	**46**		8.3	7.8	−0.5
47		7.7	7.4	−0.3	**48**		7.4	7.3	−0.1
49		7.5	7.3	−0.2	**50**		7.8	7.4	−0.4
51		8.3	7.9	−0.4	**52**		8.0	7.4	−0.6

Every compound submitted to the *VirtualDesignLab* server (by means of imported PDB files or the integrated model builder) is subjected to identical protocols as those employed to train and validate the underlying mQSAR model(s) (Figure 2). The affinity is calculated based on multiple components of the binding energy (Figure 3). Protein–ligand interaction and internal strain energies (Cheetah and Quasar) are obtained using a directional force field with polarization terms [36]. The desolvation costs are calculated for the global minimum obtained from the conformational search, using a continuum solvation model. Loss of entropy is approximated from the number of rotatable bonds constrained upon binding to the protein. Induced-fit energy calculation is an inherent function of the Quasar algorithm. The affinity predictions are based on (up to) eight docking poses as obtained from Alignator/Cheetah (4D) and take into account (up to) six induced-fit mechanisms (5D) and two solvation (6D) scenarios in order to account for the unique properties of certain binding sites (e.g., the surface-exposed FimH binding pocket). Protein–ligand structures may be viewed (binding pocket) and/or downloaded (in PDB format) upon job completion. The latter files also serve as input for other software, including the *MD Client*.

Figure 2. *VirtualDesignLab* flowchart (left) and an example of a typical workflow based on the FimH receptor (right). The compound of interest is designed using the built-in 3D model builder or imported from an external file. The main step involves the conformational sampling of the ligand in the protein's binding pocket, where all feasible poses are retained and used as input for the subsequent mQSAR. Structure management options and job controls are all accessible from a central interface window. References to the individual pieces of software are given in text.

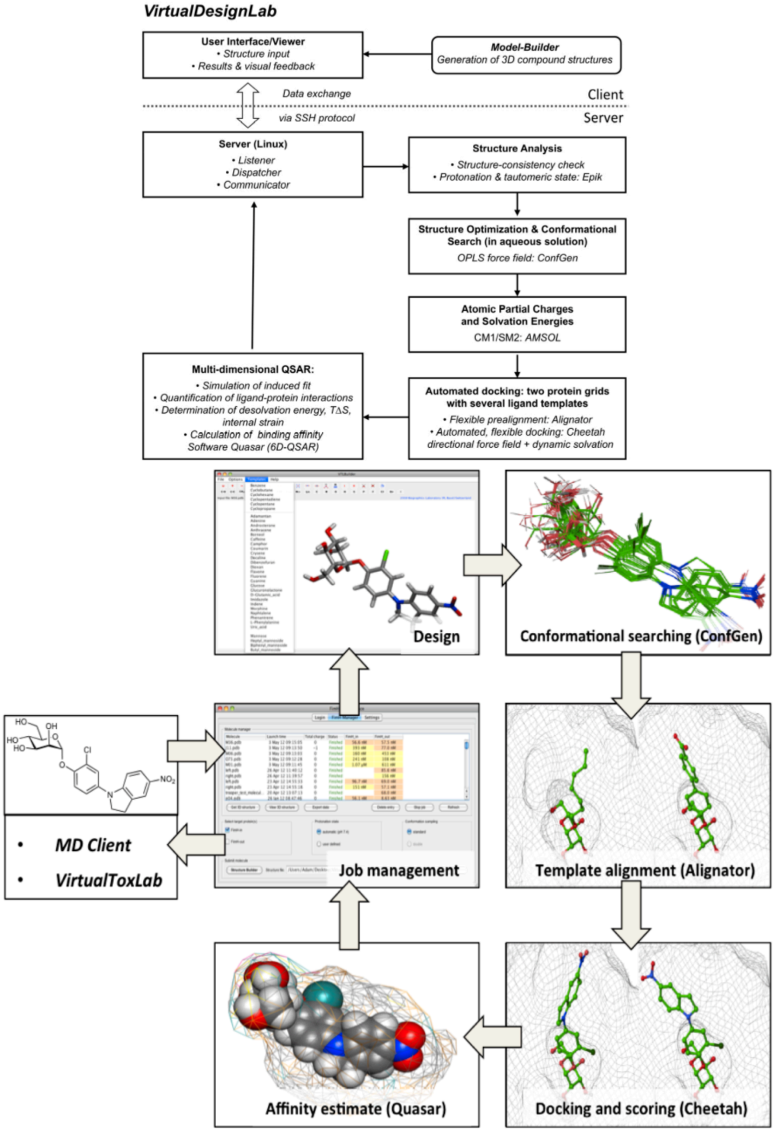

VirtualDesignLab
User Interface/Viewer
• Structure input
• Results & visual feedback
Model-Builder
Generation of 3D compound structures
Data exchange
via SSH protocol
Client
Server
Server (Linux)
• Listener
• Dispatcher
• Communicator
Structure Analysis
• Structure-consistency check
• Protonation & tautomeric state: Epik
Structure Optimization & Conformational Search (in aqueous solution)
OPLS force field: ConfGen
Atomic Partial Charges and Solvation Energies
CM1/SM2: AMSOL
Multi-dimensional QSAR:
• Simulation of induced fit
• Quantification of ligand-protein interactions
• Determination of desolvation energy, TΔS, internal strain
• Calculation of binding affinity
Software Quasar (6D-QSAR)
Automated docking: two protein grids with several ligand templates
• Flexible prealignment: Alignator
• Automated, flexible docking: Cheetah
directional force field + dynamic solvation
Design
Conformational searching (ConfGen)
HO
OH
HO
HO
O
Cl
O
N
NO2
• MD Client
• VirtualToxLab
Job management
Template alignment (Alignator)
Affinity estimate (Quasar)
Docking and scoring (Cheetah)

Figure 3. The equation for calculating the binding energy used in the *VirtualDesignLab/VirtualToxLab* and the directional force field employed in Cheetah and Quasar [32]. The individual terms—quantifying experimentally accessible quantities, such as bond lengths, bond angles, torsion angles, van der Waals contacts, geometries of hydrogen bonds, electrostatic and metal-ligand interactions, as well as ligand→protein polarization—are described in greater detail in the software documentation found at http://www.biograf.ch/index.php?id=software.

$$E_{\text{binding}} = E_{\text{ligand-receptor}} - E_{\text{ligand desolv.}} - T\Delta S - E_{\text{ligand strain}} - E_{\text{induced fit}}$$

$$E_{ligand-receptor} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{nb\ pairs} \frac{q_i \cdot q_j}{4\pi\varepsilon_0 D(r) r_{ij}} + \sum_{nb\ pairs} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}}\right) +$$

$$\sum_{H\ bonds} \left(\frac{C}{r_{ij}^{12}} - \frac{D}{r_{ij}^{10}}\right) \cdot \cos^2(\theta_{Don-H\cdots Acc}) \cdot \cos^n(\omega_{H\cdots Acc-LP}) + \sum_{metal\ pairs} \frac{q_i^{CT} \cdot q_j^{CT}}{4\pi\varepsilon_0 D(r) r_{ij}} + \sum_{metal\ pairs} \left(\frac{E}{r_{ij}^{12}} - \frac{F}{r_{ij}^{10}}\right) + \sum_{atoms} -\frac{1}{2}\alpha_i\left[\vec{E}_i^{\circ} \cdot \vec{E}_i\right]$$

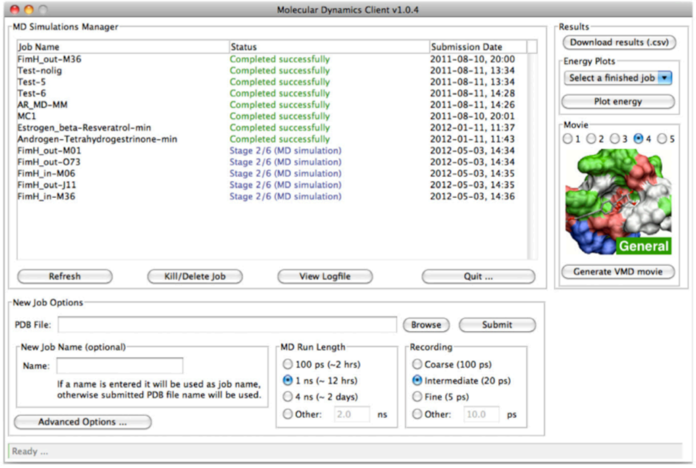
2.2. MD Client

In addition to binding affinities estimated from mQSAR based on the docking simulations, medicinal chemists might wish to analyze the kinetic stability of ligand–protein interaction used by means of molecular-dynamics (MD) simulations. MD has been successful in studying structural fluctuations in proteins [37–39], lipids [40–42] and nucleic acids [43,44], as well as in the refinement of structures solved by X-ray crystallography and NMR [5]. Despite the availability of a wealth of software packages for performing MD simulations, (e.g., Desmond [20], Amber [45], CHARMM [46], GROMOS [47] and GROMACS [48]), the lengthy setup and laborious post-processing act as a barrier, preventing users from routinely utilizing these simulations. We therefore developed the *MD Client* to overcome this limitation by requiring as few settings as possible to quickly and reliably highlight basic features of the dynamics of the studied protein–ligand complex. Our *MD Client* is designed specifically for use by bench medicinal chemists interested in exploring ligand–protein interaction dynamics.

2.2.1. The *MD Client* Interface

The *MD Client* utilizes a simple and intuitive GUI front-end and a more sophisticated back-end that handles all “under-the-hood” tasks, from cleaning the input structure to post-processing MD trajectory and gathering energy results. Both front- and back-end programs were developed in python 2.6 (http://www.python.org) using standard extensions, such as the *TkInter* GUI package and the *matplotlib* library for rendering interactive 2D plots. The front-end has been compiled for Mac OS X, Linux and Windows operating systems. The communication between the front-end and the back-end on the remote server is carried out via a Secure Shell (SSH) protocol. A molecular dynamics simulation of a ligand–protein complex (as obtained, for instance, from the *VirtualDesignLab*) is launched by a single-click in the *MD Client* interface. The *MD Client* provides control over the basic parameters of submitted MD simulation: namely its length and frequency of taking snapshots for subsequent energy analysis and movie production (Figure 4). The “Advanced Options”, button enables the user to control more details of the MD simulation; yet the default options are adequate in most cases. A list box keeps track of jobs currently on the server and their current status. The user can monitor the progress of running jobs or, when needed, terminate them at any stage.

Figure 4. Appearance of the user interface of *MD Client*; top-left: list of jobs currently on remote server; bottom: basic simulation settings; and right: results download and analysis.



Molecular Dynamics Client v1.0.4

MD Simulations Manager

Job Name	Status	Submission Date
FimH_out-M36	Completed successfully	2011-08-10, 20:00
Test-nolig	Completed successfully	2011-08-11, 13:34
Test-5	Completed successfully	2011-08-11, 13:34
Test-6	Completed successfully	2011-08-11, 14:28
AR_MD-MM	Completed successfully	2011-08-11, 14:26
MC1	Completed successfully	2011-08-10, 20:01
Estrogen_beta-Resveratrol-min	Completed successfully	2012-01-11, 11:37
Androgen-Tetrahydrogestrinone-min	Completed successfully	2012-01-11, 11:43
FimH_out-M01	Stage 2/6 (MD simulation)	2012-05-03, 14:34
FimH_out-O73	Stage 2/6 (MD simulation)	2012-05-03, 14:34
FimH_in-M06	Stage 2/6 (MD simulation)	2012-05-03, 14:35
FimH_out-J11	Stage 2/6 (MD simulation)	2012-05-03, 14:35
FimH_in-M36	Stage 2/6 (MD simulation)	2012-05-03, 14:36

New Job Options

PDB File:

New Job Name (optional)

Name:

If a name is entered it will be used as job name, otherwise submitted PDB file name will be used.

MD Run Length

○ 100 ps (~2 hrs)
● 1 ns (~ 12 hrs)
○ 4 ns (~ 2 days)
○ Other: 2.0 ns

Recording

○ Coarse (100 ps)
● Intermediate (20 ps)
○ Fine (5 ps)
○ Other: 10.0 ps

Movie ○ 1 ○ 2 ○ 3 ● 4 ○ 5

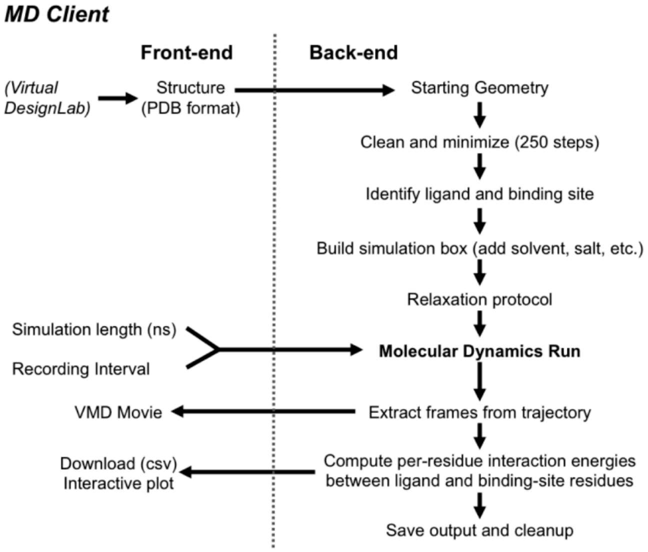
Ready ...

MD simulations require different types of data input. The most important is the structure file containing input geometries of the ligand–protein complex. Currently, the *MD Client* accepts files with Protein Data Bank (PDB) format [49]. The *VirtualDesignLab* output structures (*i.e.*, ligand-protein complexes) can be directly used as input for the *MD Client*. When an MD simulation is completed, the user can download extracted frames as standard PDB files for viewing. For the 3D visualization of structures, several free 3D-rendering tools, such as Visual Molecular Dynamics (VMD) [50], are available. To facilitate the importing of MD trajectories into VMD, we added a functionality that automatically generates a VMD visualization-state file linked to the downloaded frames. The user can choose between different pre-defined visualization styles and simply click on "Generate VMD Movie" in the *MD Client* interface (Figure 4) to produce a file that can be loaded directly into VMD. This spares the user the time and effort needed to load individual PDB files and set up the view options in VMD. Most importantly, the user can use *MD Client's* built-in plotting tool to analyze ligand–protein interaction dynamics (cf. Figure 7) and compare them amongst multiple systems, which we are going to demonstrate in the results section on selected antagonists binding to the FimH receptor.

2.2.2. The *MD Client* Back-end

The *MD Client* back-end resides on the remote server, where all computational jobs are to take place. It utilizes Schrödinger's Python API (http://www.schrodinger.com/pythonapi) for reading structures, launching MD simulations and computing per-residue interaction energies. It receives input structure and primary simulation settings from the front-end. Figure 5 shows how the *MD Client* back-end processes input structures into useful quantities. It starts by constructing atom connectivity and bond orders for the submitted structure and doing a short energy minimization to relieve structural inconsistencies in bond lengths, angles, steric clashes, *etc.* The *MD Client* back-end automatically identifies the ligand-like molecule and defines binding site residues as all residues within 8 Å (default) from ligand atoms. This definition is employed for subsequent use in energy computations and movie production.

Figure 5. Workflow of the *MD Client* and communication between front-end and back-end.



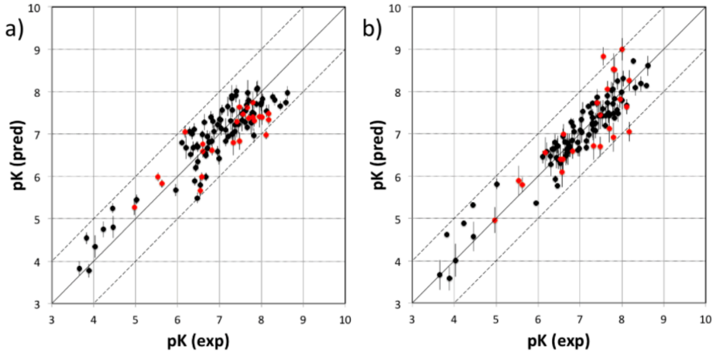
The *MD Client* back-end employs the Desmond package from the D. E. Shaw Research laboratory to perform the MD simulations [51,52]. Desmond and its source code are distributed under free license to non-commercial and academic users. It uses novel parallel algorithms and numerical techniques to achieve high performance and accuracy on platforms containing a large number of processors, but may also be executed on a single-processor computer [20]. Desmond's System Builder soaks the submitted ligand–protein complex into a TIP3P water box extending 10 Å beyond any of the complex's atoms. It adds counter ions to neutralize the simulation box and 0.15 M sodium and chloride ions to approximate physiological conditions. The complex is first minimized to a convergence gradient threshold of 1.0 kcal/(mol·Å). The molecular-dynamics protocol utilizes the OPLS2005 force field and the NPT ensemble (constant number of particles, pressure and temperature) at 300 K, with periodic boundary conditions. The production run of the user-defined length is preceded by 24 ps of the Desmond default relaxation protocol. After completion of the MD simulation, the *MD Client* back-end extracts frames at the user-defined intervals and saves them as standard PDB files. The user can download these frames for later viewing and analysis. Finally, the extracted frames are analyzed using the component-interactions script in Maestro [34] to compute interaction energies between the ligand and individual amino acids defining the binding site along the MD simulations. Ligand-residue interaction energies are calculated as the sum of the (OPLS2005) van der Waals and electrostatic terms. This dynamic-interaction profile is saved as a time series in a comma-separated-values (csv) file for subsequent download by the user or for interactive plotting and analysis using the *MD Client* front-end interface.

3. Results and Discussion

3.1. VirtualDesignLab

The current FimH Quasar model for the *VirtualDesignLab* was established based on structural and biological data of 108 mannose-based inhibitors (with IC_{50} values ranging from 220 µM to 2.4 nM) displaying diverse PK/PD profiles. Compound synthesis, biological assays [53] and model development were performed in-house, ensuring consistency of all results. Table 1 shows the structures, experimental and predicted affinities of compounds used for developing the model. The QSAR model based on a genetic algorithm converged at a cross-validated r^2 of 0.805 and yielded a predictive r^2 of 0.596 (Figure 6a). The only modest value of the predictive r^2 is a consequence of the relatively narrow range of test compound affinities (as some substances were necessary for the training set due to their structural uniqueness). The performance of the model is therefore better reflected by the individual predictions (23 out of 24 test substances within a factor of 10 from their experimental affinity). We further challenged the FimH model by using Y-scrambling and consensus scoring with the software Raptor (dual-shell 5D-QSAR; Figures 6b and S1) [54]. All tests, including the processing of additional external compounds, confirmed the predictive power of the mQSAR model-based framework.

Figure 6. Comparison between experimental (horizontal axis) and predicted pK values (vertical axis) for (**a**) the Quasar model and (**b**) the Raptor model. Black and red points represent compounds of the training and test set, respectively. Vertical bars indicate the estimated standard deviations of the prediction. Dashed lines are drawn at factors of 10 from the experimental value.



The *VirtualDesignLab* is aimed at predicting the binding affinity for a given compound within a factor of 10 from the experimental value. Currently, the affinity prediction for a single compound requires approximately one hour of CPU time—a good balance between accuracy and processing time. Special treatment may, however, be required for compounds retaining flexibility upon binding. In such cases, improved entropy estimation (a method is currently in development at our institute) or a non-static approach, such as the one offered by the *MD Client* (cf. below), may be necessary. We would like to emphasize that the framework is independent of the FimH mQSAR model (presented here), as it only requires the generation and validation of a new QSAR model for any target protein of interest. This can be developed using by freely available software, e.g., Quasar [35].

3.2. MD Client

In the *MD Client*, the outcome of an MD simulation includes a set of frames extracted from the MD trajectory and a dynamic-interaction profile comprising per-residue interaction energies between ligand and protein for all time points. An interactive plot of computed interaction profiles is readily accessible from the *MD Client* interface (Figure 7). The plot created by the *matplotlib* python extension is cumulative, *i.e.*, it can incorporate dynamic profiles from several simulations in the same plot with automated coloring and legend generation. Comparing dynamic profiles of different simulations may provide valuable clues, for instance, about interaction modes of different ligands and/or key residues in ligand recognition and binding. We chose five structurally distinct FimH ligands (9, 17, 18, 28 and 37) to demonstrate the usefulness of dynamic-interaction profiles. Examination of their dynamic interaction energies with two key FimH residues (Gln133 and Phe1) indicates that these interactions are maintained throughout the entire simulation and that they don't significantly differ among different ligand classes (Figure 7a,b). These residues are typically involved in an extended hydrogen-bond network with the mannose moiety common to FimH binders. The profiles also show that the interaction with the N-terminal NH_3^+ moiety of Phe1 results in a considerably larger contribution to the binding enthalpy compared to Gln133. Automated docking of FimH ligands typically predicts a hydrogen bond from the 3-OH of the mannose moiety to the Asp140 residue to be thermodynamically favorable. Interestingly, this hydrogen bond does not seem to be kinetically stable, since it is broken within the first 0.5 ns and is never re-established throughout the entire simulation as can be observed in the profiles of all studied compounds (Figure 7c).

Figure 7. Dynamic per-residue interaction plots for five FimH ligands generated by the interactive plotting feature of *MD Client*; (**a**) Gln133, (**b**) Phe1, (**c**) Asp140 and (**d**) Tyr48. Vertical axis: protein–ligand interaction energies (kcal/mol); horizontal axis: molecular dynamics simulation time (ns). The colors mark the individual compounds shown above.

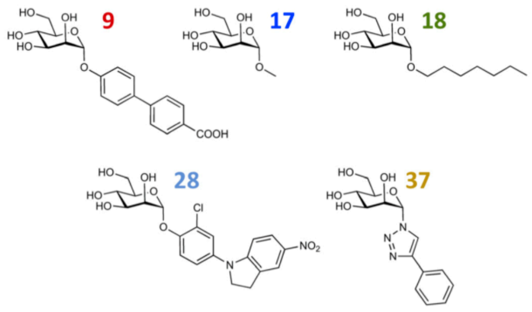
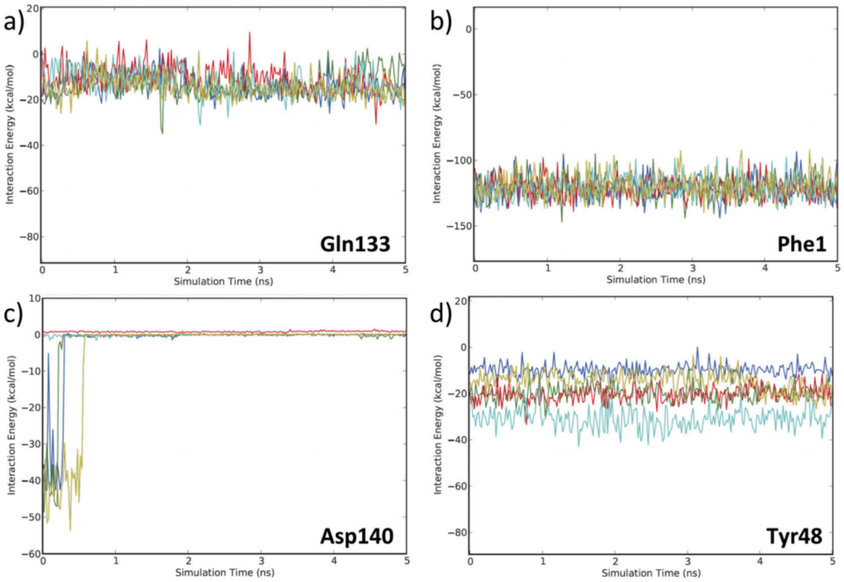
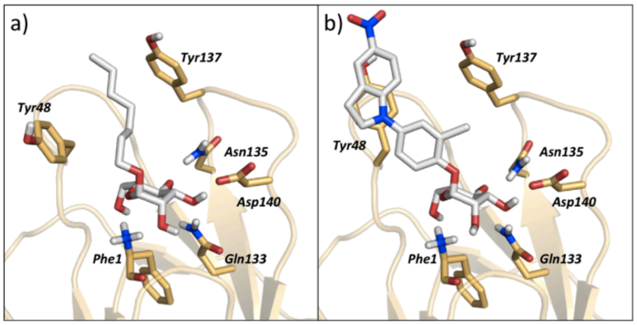


Figure 7. *Cont.*



The various classes of FimH ligands differ in their interaction patterns with the so-called *tyrosine gate* lined by Tyr48 and Tyr137 located at the entrance to the mannose-binding site (Figure 8) [23]. The interaction dynamics show that compound 28 exhibits the strongest interaction with Tyr48, which could be explained by its unique scaffold that allows a preferred interaction with the tyrosine side chain (Figure 7d). Compounds 9 and 18 display a favorable interaction with Tyr48, yet of lower magnitude than 28, which explains the superior affinity of the latter (see Table 1). Finally, two ligands seem to lack this favorable interaction with Tyr48; namely 17 and 37, which also coincides with their relatively lower affinities. This could be rationalized by the lack of a side chain capable of interacting with the Tyr48 in 17 and the inherent rigidity of 37 due to a lack of glycosidic oxygen in the linker between mannose and the aromatic aglycone.

Figure 8. Illustration of (**a**) *in* binding mode of 18 and (**b**) *out* binding mode of 28, within the tyrosine gate of FimH binding site. References to different binding modes of FimH ligands are given in the methods section.



4. Software Extension/Repurposing

The philosophy behind all our software is to allow for extendibility, as well as redirection, towards different targets of interest. In the following, we provide a brief overview of the corresponding requirements.

4.1. VirtualDesignLab

Assuming that a decent number of ligands with experimental affinity data are available for the given target, a three-dimensional protein-ligand structure that will serve for automated, flexible docking is required. These are usually obtained by means of crystallography or homology modeling and must typically be further refined (addition of hydrogen atoms, completion of missing amino-acid residues or their side chains, completion/generation of the solvent shell, energy minimization). These tasks can be accomplished through numerous, freely accessible computational tools. With the structure at hand, potential binding poses of all tested compounds need to be obtained. For this step, any flexible-docking software may be employed, including Alignator/Cheetah discussed in this article. The ensemble of potential binding modes can be compiled into a 4D data set to serve as input for the generation of the binding site surrogate. Though this task is best handled using the Quasar software, a QSAR model of different origin could also potentially be utilized. It should be noted, however, that even though structure preparation and pose generation are relatively simple tasks requiring no more than a few days of work, developing and validating a robust and reliable QSAR model is a complex and lengthy procedure. Also, given the vast diversity of the computational methods, personal communication with the authors of this article would likely be necessary in order to integrate a QSAR model with the *VirtualDesignLab* framework.

4.2. MD Client

The *MD Client* relies on the Desmond package at its back-end terminal to perform MD simulations. Desmond can be obtained free of charge for academics and non-commercial users. Once a working Desmond installation is available, the user needs to point the *MD Client* to where the back-end is located by providing the necessary SSH credentials. The *MD Client* can basically take it from there, since it has internal routines for identifying protein and ligand, job submission and monitoring, as well as calculating, organizing and plotting the interaction energy results.

5. Conclusions

Over the past three decades, much progress has been made in developing and validating innovative computational algorithms for common drug design-related tasks. In their perspective on the future of medicinal chemistry, Satyanarayanajois and Hill [55] stated that emerging medicinal chemists should additionally acquire "computational and cheminformatics acumen considerably greater than in years past". In a related analysis, Ritchie and McLay [56] concluded that the goal of encouraging medicinal chemists to rely more on computational chemistry tools could be best achieved via specially designed tools that are "well-thought-out, suitable for their needs, able to generate useful, timely and valid results". Similarly, we trust that adopting this strategy will ultimately maximize the benefit of state-of-the-art modeling technologies in the field of drug design and development.

To this end, we designed versatile single-click tools to assist medicinal chemists in performing two routine modeling tasks: the *VirtualDesignLab* for predicting binding mode and affinity of potential drug candidates and the *MD Client* for investigating dynamic behavior and energetics of ligand-protein complexes. Thanks to their modular design based mainly on self-developed algorithms, our tools allow easy modification, extension, as well as reorientation towards other targets and platforms of interest. Our group previously introduced the *OpenVirtualToxLab* for prediction of the toxic potential of drug candidates and made it freely available to academic organizations [57]. The two new tools introduced in this article, *VirtualDesignLab* and *MD Client*, can also be made available on request. Our future plans for *MD Client* include adding support for more molecular dynamics packages, as well as more analysis functionalities (for instance, surface area and entropy computations) to give more insight into ligand–protein interaction processes.

In closing, we wish to emphasize that our tools are not intended to, neither can they, replace the expert molecular modeler. In fact, their main purpose is to facilitate handling routine drug-design related tasks, thus leaving the more time-consuming detailed investigation only for interesting cases. Our vision is to place our interfacing technologies right on the medicinal chemists' workbench and keep all the complicated 'machinery' on a transparently maintained server. However, the simplicity of such tools, although tempting, should never cloud the medicinal chemists' judgment. On the contrary, medicinal chemists should always employ their expertise to question the results obtained from such tools.

Acknowledgments

Financial support from the Swiss National Foundation (grant No. 315230_122478) and the Department of Pharmaceutical Sciences, University of Basel, is gratefully acknowledged.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Bottegoni, G. Protein-ligand docking. *Front. Biosci.* **2012**, *17*, 2289–2306.
2. Kitchen, D.B.; Decornez, H.; Furr, J.R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* **2004**, *3*, 935–949.
3. Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Naenna, T.; Prachayasittikul, V. A practical overview of Quantitative Structure-Activity Relationship. *EXCLI Journal* **2009**, *8*, 74–88.
4. Verma, J.; Khedkar, V.M.; Coutinho, E.C. 3D-QSAR in drug design—A review. *Curr. Top. Med. Chem.* **2010**, *10*, 95–115.
5. Karplus, M.; McCammon, J.A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.

6. Cheatham, T.E.; Kollman, P.A. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435–471.
7. Alonso, H.; Bliznyuk, A.A.; Gready, J.E. Combining docking and molecular dynamic simulations in drug design. *Med. Res. Rev.* **2006**, *26*, 531–568.
8. Rapaport, D.C. *The Art of Molecular Dynamics Simulation*. Cambridge University Press: Cambridge, UK, 2004; pp. 1–10.
9. Törnroth-Horsefield, S.; Wang, Y.; Hedfalk, K.; Johanson, U.; Karlsson, M.; Tajkhorshid, E.; Neutze, R.; Kjellbom, P. Structural mechanism of plant aquaporin gating. *Nature* **2006**, *439*, 688–694.
10. Beckstein, O.; Tai, K.; Sansom, M.S. Not ions alone: Barriers to ion permeation in nanopores and channels. *J. Am. Chem. Soc.* **2007**, *126*, 14694–14695.
11. Berendsen, H.J.C. Simulating the Physical World: Hierarchical Modeling from Quantum Mechanics to Fluid Dynamics; Cambridge University Press: Cambridge, UK; 2007; p. 624.
12. Muegge, I. Synergies of virtual screening approaches. *Mini Rev. Med. Chem.* **2008**, *8*, 927–933.
13. Kirchmair, J.; Distinto, S.; Schuster, D.; Spitzer, G.; Langer, T.; Wolber, G. Enhancing drug discovery through in silico screening: Strategies to increase true positives retrieval rates. *Curr. Med. Chem.* **2008**, *15*, 2040–2053.
14. Schrödinger Suite. Available online: http://www.schrodinger.com (accessed on 12 December 2012).
15. Accelrys Discovery Studio. Available online: http://accelrys.com/products/discovery-studio (accessed on 12 December 2012).
16. SYBYL-X Suite. Available online: http://tripos.com (accessed on 12 December 2012).
17. Molecular Operating Environment. Available online: http://www.chemcomp.com (accessed on 12 December 2012).
18. Morris, G.M.; Goodsell, D.S.; Halliday, R.S.; Huey, R.; Hart, W.E.; Belew, R.K.; Olson, A.J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.
19. Vedani, A.; Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* **2002**, *45*, 2139–2149.
20. Bowers, K.J.; Chow, E.; Xu, H.; Dror, R.O.; Eastwood, M.P.; Gregersen, B.A.; Klepeis, J.L.; Kolossvary, I.; Moraes, M.A.; Sacerdoti, F.D.; *et al.* Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In *Proceedings of the ACM/IEEE Conference on Supercomputing (SC06)*, Tampa, FL, USA, 11–17 November 2006.
21. Irwin, J.J.; Shoichet, B.K.; Mysinger, M.M.; Huang, N.; Colizzi, F.; Wassam, P.; Cao, Y. Automated docking screens: A feasibility study. *J. Med. Chem.* **2009**, *52*, 5712–5720.
22. Klein, T.; Abgottspon, D.; Wittwer, M.; Rabbani, S.; Herold, J.; Jiang, X.; Kleeb, S.; Luthi, C.; Scharenberg M.; Bezencon, J.; *et al*. FimH Antagonists for the oral treatment of urinary tract infections: From design and synthesis to *in vitro* and *in vivo* evaluation. *J. Med. Chem.* **2010**, *53*, 8627–8641.
23. Schwardt, O.; Rabbani, S.; Hartmann, M.; Abgottspon, D.; Wittwer, M.; Kleeb, S.; Zalewski, A.; Smieško, M.; Cutting, B.; Ernst, B. Design, synthesis and biological evaluation of mannosyl triazoles as FimH antagonists. *Bioorg. Med. Chem.* **2011**, *19*, 6454–6473.

24. Jiang, X.; Abgottspon, D.; Kleeb, S.; Rabbani, S.; Scharenberg, M.; Wittwer, M.; Haug, M.; Schwardt, O.; Ernst, B. Antiadhesion therapy for urinary tract infections—A balanced PK/PD profile proved to be key for success. *J. Med. Chem.* **2012**, *55*, 4700–4713.
25. QikProp, version 3.4; Schrödinger, LLC: New York, NY, USA, 2011.
26. Vedani, A.; Dobler, M.; Smieško, M. VirtualToxLab—A platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicol. Appl. Pharmacol.* **2012**, *261*, 142–153.
27. VirtualToxLab Online Documentation. Available online: http://www.biograf.ch/downloads/VirtualToxLab.pdf (accessed on 12 December 2012).
28. MacroModel, version 9.9; Schrödinger, LLC: New York, NY, USA, 2011.
29. Storer, J.W.; Giesen, D.J.; Cramer, C.J.; Truhlar, D.G. Class IV charge models: A new semiempirical approach in quantum chemistry. *J. Comput. Aided Mol. Des.* **1995**, *9*, 87–110.
30. ConfGen, version 2.1; Schrödinger, LLC: New York, NY, USA, 2009.
31. Smieško, M. University of Basel, Basel, Switzerland. Personal communication, 2012.
32. Rossato, G.; Ernst, B.; Smiesko, M.; Spreafico, M.; Vedani, A. Probing small-molecule binding to cytochrome P450 2D6 and 2C9: An *in silico* protocol for generating toxicity alerts. *ChemMedChem* **2010**, *5*, 2088–2101.
33. Han, Z.; Pinkner, J.S.; Ford, B.; Obermann, R.; Nolan, W.; Wildman, S.A.; Hobbs, D.; Ellenberger, T.; Cusumano, C.K.; Hultgren, S.J.; *et al*. Structure-based drug design and optimization of mannoside bacterial FimH antagonists. *J. Med. Chem.* **2010**, *53*, 4779–4792.
34. Maestro, version 9.2; Schrödinger, LLC: New York, NY, USA, 2011.
35. Vedani, A.; Dobler, M.; Lill, M.A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005**, *48*, 3700–3703.
36. Vedani, A.; Huhta, D.W. A new force field for modeling metalloproteins. *J. Amer. Chem. Soc.* **1990**, *112*, 4759–4767.
37. Rojewska, D.; Elber, R. Molecular dynamics study of secondary structure motions in proteins: application to myohemerythrin. *Proteins* **1990**, *7*, 265–279.
38. Wan, S.; Flower, D.R.; Coveney, P.V. Toward an atomistic understanding of the immune synapse: Large-scale molecular dynamics simulation of a membrane-embedded TCR-pMHC-CD4 complex. *Mol. Immunol.* **2008**, *45*, 1221–1230.
39. Knapp, B.; Omasits, U.; Bohle, B.; Maillere, B.; Ebner, C.; Schreiner, W.; Jahn-Schmid, B. 3-Layer-based analysis of peptide-MHC interaction: *In silico* prediction, peptide binding affinity and T cell activation in a relevant allergen-specific model. *Mol. Immunol*. **2009**, *46*, 1839–1844.
40. Coll, E.P.; Kandt, C.; Bird, D.A.; Samuels, A.L.; Tieleman, D.P. The distribution and conformation of very long-chain plant wax components in a lipid bilayer. *J. Phys. Chem. B* **2007**, *111*, 8702–8704.
41. Kandt, C.; Ash, W.L.; Tieleman, D.P. Setting up and running molecular dynamics simulations of membrane proteins. *Methods* **2007**, *41*, 475–488.
42. Heller, H.; Schaefer, M.; Schulten, K. Molecular dynamics simulation of a bilayer of 200 lipids in the gel and in the liquid crystal phase. *J. Phys. Chem.* **1993**, *97*, 8343–8360.

43. Miller, J.L.; Kollman, P.A. Theoretical studies of an exceptionally stable RNA tetraloop: Observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.* **1997**, *270*, 436–450.
44. Luo, J.; Bruice, T.C. Nanosecond molecular dynamics of hybrid triplex and duplex of polycation deoxyribonucleic guanidine strands with a complimentary DNA strand. *J. Amer. Chem. Soc.* **1998**, *120*, 1115–1123.
45. Case, D.A.; Cheatham, T.E., III.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
46. Brooks, B.R.; Bruccoleri, R.E.; Olafson, B.D.; States, D.J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.
47. Christen, M.; Hünenberger, P.H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D.P.; Heinz, T.N.; Kastenholz, M.A.; Kräutler, V.; Oostenbrink, C.; *et al*. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comp. Chem.* **2005**, *26*, 1719–1751.
48. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.
49. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
50. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
51. Desmond Molecular Dynamics System, version 3.1; D. E. Shaw Research: New York, NY, USA, 2012.
52. Maestro-Desmond Interoperability Tools, version 3.1; Schrödinger: New York, NY, USA, 2012.
53. Rabbani, S.; Jiang, X; Schwardt, O.; Ernst, B. Expression of the carbohydrate recognition domain of FimH and development of a competitive binding assay. *Anal. Biochem.* **2010**, *407*, 188–195.
54. Lill, M.A.; Vedani, A.; Dobler, M. Raptor: Combining dual-shell representation, induced-fit simulation and hydrophobicity scoring in receptor modeling: Application toward the simulation of structurally diverse ligand sets. *J. Med. Chem.* **2004**, *47*, 6174–6186.
55. Satyanarayanajois, S.D.; Hill, R.A. Medicinal chemistry for 2020. *Future Med. Chem.* **2011**, *3*, 1765–1786.
56. Ritchie, T.J.; McLay, I.M. Should medicinal chemists do molecular modelling? *Drug Discov. Today* **2012**, *17*, 534–537.
57. Open VirtualToxLab. Available online: http://www.virtualtoxlab.org (accessed on 12 December 2012).

Appendix 4: The MD Client Quick Start Guide

Molecular Dynamics Client

1.0.4b

Quick Start Guide

Starting MD simulations

Job Management

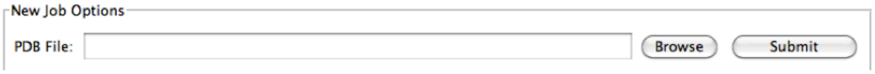
Analyzing dynamic interaction energies

Producing a movie

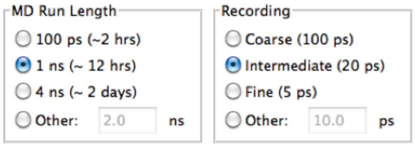
Advanced Options

Starting MD simulations

1. Click **Browse** to choose PDB file containing your ligand–protein complex.



2. Select required MD length (ns)* and frequency of taking snapshots.



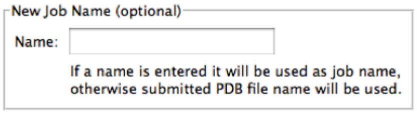
MD Run Length

○ 100 ps (~2 hrs)
● 1 ns (~ 12 hrs)
○ 4 ns (~ 2 days)
○ Other: 2.0 ns

Recording

○ Coarse (100 ps)
● Intermediate (20 ps)
○ Fine (5 ps)
○ Other: 10.0 ps

3. If necessary, provide an alternative job name (spaces are not allowed).

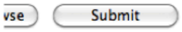


New Job Name (optional)

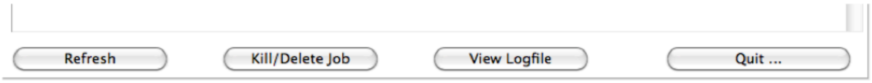
Name:

If a name is entered it will be used as job name, otherwise submitted PDB file name will be used.

4. Click **Submit**.



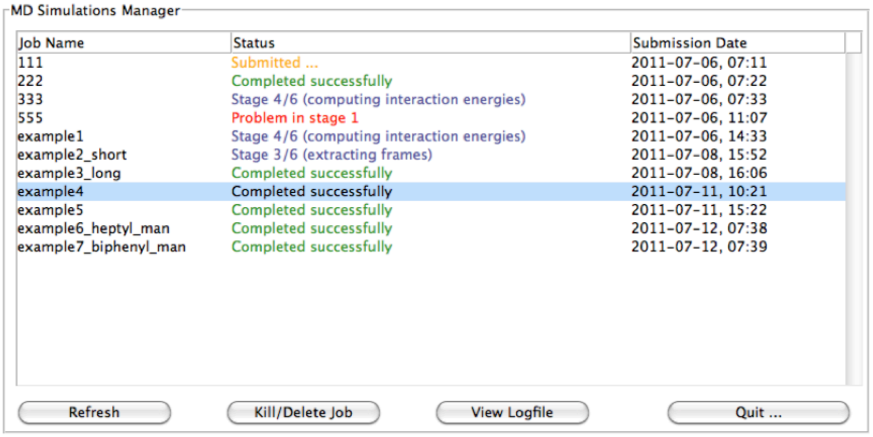
5. Click **Refresh** to update job list.



* Time requirements were estimated based on FimH (PDB: 1UWF). MD simulations for other proteins might take longer (or shorter) if they differ in size from FimH.

Job Management

MD Simulations Manager lists jobs currently available on the server. Four functions can be performed using buttons at its bottom:

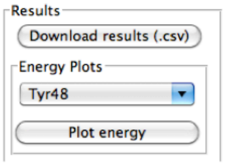


Job Name	Status	Submission Date
111	Submitted ...	2011-07-06, 07:11
222	Completed successfully	2011-07-06, 07:22
333	Stage 4/6 (computing interaction energies)	2011-07-06, 07:33
555	Problem in stage 1	2011-07-06, 11:07
example1	Stage 4/6 (computing interaction energies)	2011-07-06, 14:33
example2_short	Stage 3/6 (extracting frames)	2011-07-08, 15:52
example3_long	Completed successfully	2011-07-08, 16:06
example4	Completed successfully	2011-07-11, 10:21
example5	Completed successfully	2011-07-11, 15:22
example6_heptyl_man	Completed successfully	2011-07-12, 07:38
example7_biphenyl_man	Completed successfully	2011-07-12, 07:39

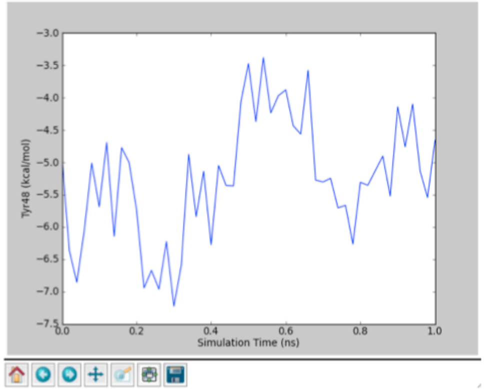
1. **Refresh**
   updates list of jobs from server and their current status
2. **Kill/Delete Job**
   stops running simulation and/or deletes all files associated with it
   keep in mind: this action can NOT be reversed
3. **View Logfile**
   reads and displays the job log file from server
4. **Quit ...**
   closes MD Client

Analyzing dynamic interaction energies

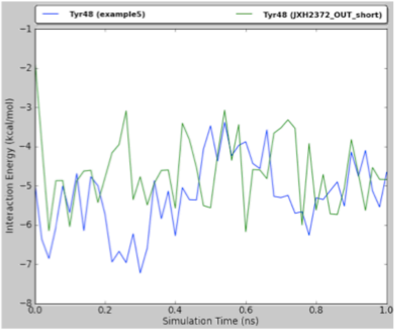
1. After MD simulation is finished and frames are harvested, interaction energies between ligand and surrounding protein residues are calculated and saved in a comma-separated values (.csv) file on the server.
2. Click **Download results (.csv)** to save a copy of this file locally.

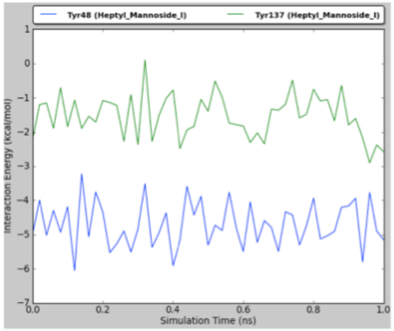


3. Select a residue and click **Plot Energy** to preview how its interaction with the ligand evolves along the MD simulation.



4. Any subsequent clicks on **Plot Energy** will add to the already existing plot, as long as its window is still open.
5. You can use this feature to compare interaction with certain residue across two (or more) simulations, …
   or to compare strengths of ligand interaction with two (or more) residues in the same simulation.

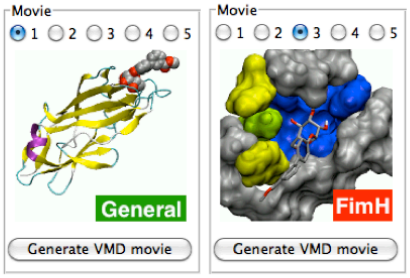




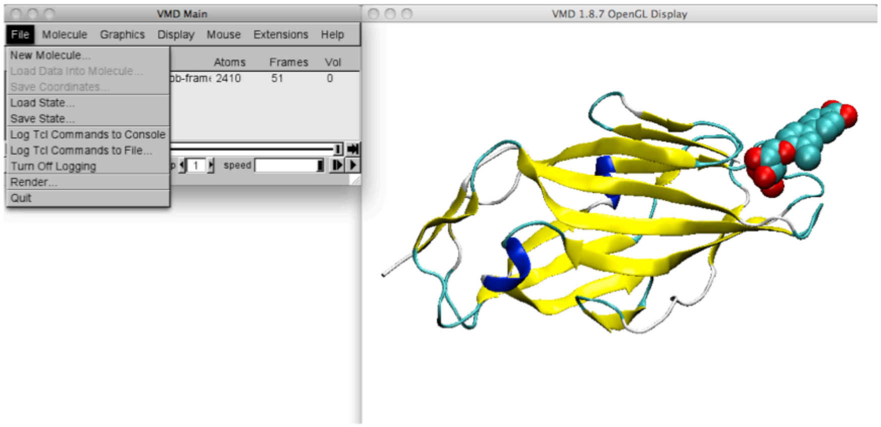
6. You can save your plots to a variety of formats depending on the *file extension* you provide. The default output is PNG, but you can also save in SVG, PDF, EPS and *high-quality* PS formats.

Producing a movie

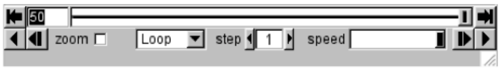
1. Select a completed simulation, choose the output style[†] and click **Generate VMD movie**, then provide a path to save the movie.



2. In the path you provided you will find a folder named *'frames'* and a *.vmd* file. Open this file in VMD using **File | Load State ...**



3. Use VMD movie control options to play your MD movie.



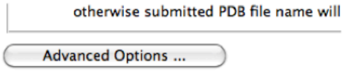
4. Keep in mind that the *location* where you saved the movie *matters*! If you move the movie folder to another location it will not work.

5. In cases where you want to place the movie folder somewhere else, simply re-issue the **Generate VMD movie** command, and provide the new location; MD Client will take care of the rest.

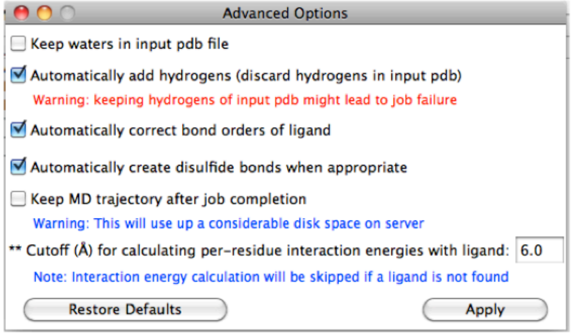
[†] Note that some styles work only with FimH.

Advanced Options

1. In addition to standard job settings, some advanced options are available to fit some purposes.


otherwise submitted PDB file name will

2. Default values are optimized to work fine in most cases. You can restore them at any time by clicking **Restore Defaults**.



☐ Keep waters in input pdb file

☑ Automatically add hydrogens (discard hydrogens in input pdb)

Warning: keeping hydrogens of input pdb might lead to job failure

☑ Automatically correct bond orders of ligand

☑ Automatically create disulfide bonds when appropriate

☐ Keep MD trajectory after job completion

Warning: This will use up a considerable disk space on server

** Cutoff (Å) for calculating per-residue interaction energies with ligand: 6.0

Note: Interaction energy calculation will be skipped if a ligand is not found